STATISTICS AND PROBABILITY



Las Positas College

LibreTexts™

Statistics and Probability

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 04/19/2026

TABLE OF CONTENTS

Licensing

1: The Nature of Statistics

- 1.1: Introduction
- 1.2: Descriptive and Inferential Statistics
- 1.3: Variables and Types of Data
  - 1.3.1: Levels of Measurement
- 1.4: Data Collection and Sampling Techniques
- 1.5: Computers and Calculators
  - 1.5.1: Using Spreadsheets for Statistics
- 1.6: Experimental Design and Ethics
  - 1.6.1: More on Experiments
  - 1.6.2: Observational Studies and Sampling Strategies
- 1.E: Sampling and Data (Exercises)

2: Frequency Distributions and Graphs

- 2.1: Prelude to Graphs
- 2.2: Histograms, Ogives, and Frequency Polygons
  - 2.2.1: Frequency Polygons and Time Series Graphs
- 2.3: Other Types of Graphs
  - 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
  - 2.3.2: Dot Plots
  - 2.3.3: Guide to Fairly Good Graphs
  - 2.3.4: Presenting Data in Tables
- 2.4: Organizing Data - Frequency Distributions
- 2.E: Graphs (Optional Exercises)

3: Data Description

- 3.1: Prelude to Descriptive Statistics
- 3.2: Measures of the Center of the Data
  - 3.2.1: Skewness and the Mean, Median, and Mode
- 3.3: Measures of Variation
  - 3.3.1: Coefficient of Variation
  - 3.3.2: The Empirical Rule and Chebyshev's Theorem
- 3.4: Measures of Position
  - 3.4.1: Measures of Location- Deciles
  - 3.4.2: Z-scores
- 3.5: Exploratory Data Analysis
- 3.E: Descriptive Statistics (Optional Exercises)
  - 1.E: Measures of Position (Optional Exercises)

4: Probability and Counting

- 4.1: Sample Spaces and Probability
  - 4.1.1: Introduction to Probability
  - 4.1.2: Terminology
- 4.2: Independent and Mutually Exclusive Events
- 4.3: The Addition and Multiplication Rules of Probability
  - 4.3.1: Contingency Tables
  - 4.3.2: Tree and Venn Diagrams
- 4.4: Counting Rules
  - 4.4.1: Permutations
  - 4.4.2: Permutations with Similar Elements
  - 4.4.3: Combinations
- 4.5: Probability And Counting Rules
- 4.E: Probability Topics (Optional Exercises)
  - 1.E: Combinations (Optional Exercises)
  - 1.E: Permutations (Optional Exercises)
  - 1.E: Permutations with Similar Elements (Optional Exercises)
  - 1.E: Probability Using Tree Diagrams and Combinations (Optional Exercises)
  - 1.E: Tree Diagrams and the Multiplication Axiom (Optional Exercises)

5: Discrete Probability Distributions

- 5.1: Prelude to Discrete Random Variables
- 5.2: Probability Distribution Function (PDF) for a Discrete Random Variable
- 5.3: Mean or Expected Value and Standard Deviation
- 5.4: Binomial Distribution
  - 5.4.1: Binomial Distribution Formula
- 5.E: Discrete Random Variables (Optional Exercises)

6: Continuous Random Variables and the Normal Distribution

- 6.1: Introduction
  - 6.1.1: Continuous Probability Functions
  - 6.1.2: The Uniform Distribution
- 6.2: The Normal Distribution
  - 6.2.1: The Standard Normal Distribution
- 6.3: Applications of the Normal Distribution
- 6.4: The Central Limit Theorem
- 6.5: Normal Approximation to the Binomial Distribution
- 6.E: The Normal Distribution (Optional Exercises)
  - 1.E: The Central Limit Theorem for Sample Means (Optional Exercises)
  - 1.E: The Standard Normal Distribution (Optional Exercises)

7: Confidence Intervals and Sample Size

- 7.1: Confidence Intervals
- 7.2: Confidence Intervals for the Mean with Known Standard Deviation
- 7.3: Confidence Intervals for the Mean with Unknown Standard Deviation

- 7.4: Confidence Intervals and Sample Size for Proportions
- 7.5: Confidence Intervals (Summary)
- 7.E: Confidence Intervals (Optional Exercises)
  - 1.E: Confidence Intervals for the Mean with Known Standard Deviation (Optional Exercises)

8: Hypothesis Testing with One Sample

- 8.1: Steps in Hypothesis Testing
  - 8.1.1: Null and Alternative Hypotheses
  - 8.1.2: Outcomes and the Type I and Type II Errors
  - 8.1.3: Distribution Needed for Hypothesis Testing
  - 8.1.4: Rare Events, the Sample, Decision and Conclusion
  - 8.1.5: Additional Information on Hypothesis Tests
- 8.2: Hypothesis Test Examples for Means
- 8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation
- 8.4: Hypothesis Test Examples for Proportions
- 8.E: Hypothesis Testing (Optional Exercises)
  - 1.E: Distribution Needed for Hypothesis Testing (Optional Exercises)
  - 1.E: Hypothesis Testing with One Sample (Optional Exercises)
  - 1.E: Null and Alternative Hypotheses (Optional Exercises)
  - 1.E: Outcomes and the Type I and Type II Errors (Optional Exercises)
  - 1.E: Rare Events, the Sample, Decision and Conclusion (Optional Exercises)

9: Inferences with Two Samples

- 9.1: Prelude to Hypothesis Testing with Two Samples
- 9.2: Inferences for Two Population Means- Large, Independent Samples
- 9.3: Inferences for Two Population Means - Unknown Standard Deviations
- 9.4: Inferences for Two Population Means - Paired Samples
- 9.5: Inferences for Two Population Proportions
- 9.6: Which Analysis Should You Conduct?
- 9.E: Hypothesis Testing with Two Samples (Optional Exercises)

10: Correlation and Regression

- 10.1: Prelude to Linear Regression and Correlation
  - 10.1.1: Review- Linear Equations
  - 10.1.2: Scatter Plots
- 10.2: Testing the Significance of the Correlation Coefficient
- 10.3: The Regression Equation
  - 10.3.1: Prediction
- 10.4: Outliers
- 10.E: Linear Regression and Correlation (Optional Exercises)
  - 1.E: Linear Equations (Optional Exercises)
  - 1.E: Outliers (Optional Exercises)
  - 1.E: Prediction (Optional Exercises)
  - 1.E: Scatter Plots (Optional Exercises)
  - 1.E: Testing the Significance of the Correlation Coefficient (Optional Exercises)
  - 1.E: The Regression Equation (Optional Exercise)

11: Chi-Square and Analysis of Variance (ANOVA)

- 11.1: Prelude to The Chi-Square Distribution
  - 11.1.1: Facts About the Chi-Square Distribution
- 11.2: Goodness-of-Fit Test
- 11.3: Tests Using Contingency tables
  - 11.3.1: Test of Independence
  - 11.3.2: Test for Homogeneity
  - 11.3.3: Comparison of the Chi-Square Tests
- 11.4: Prelude to F Distribution and One-Way ANOVA
  - 11.4.1: One-Way ANOVA
  - 11.4.2: The F Distribution and the F-Ratio
  - 11.4.3: Facts About the F Distribution
  - 11.4.4: How to Use Microsoft Excel® for Regression Analysis
- 11.E: F Distribution and One-Way ANOVA (Optional Exercises)
- 11.E: The Chi-Square Distribution (Optional Exercises)

12: Nonparametric Statistics

- 12.1: Benefits of Distribution Free Tests
- 12.2: Randomization Tests - Two Conditions
- 12.3: Randomization Tests - Two or More Conditions
- 12.4: Randomization Association
- 12.5: Fisher's Exact Test
- 12.6: Rank Randomization Two Conditions
- 12.7: Rank Randomization Two or More Conditions
- 12.8: Rank Randomization for Association
- 12.9: Statistical Literacy Standard
- 12.10: Wilcoxon Signed-Rank Test
- 12.11: Kruskal–Wallis Test
- 12.12: Spearman Rank Correlation
- 12.13: Choosing the Right Test
- 12.E: Distribution Free Tests (Exercises)

13: Appendices

- 13.1: A | Statistical Table- Standard Normal (Z)
- 13.2: A | Statistical Table- Student t Distribution
- 13.3: A | Statistical Table- Chi-Square Distribution
- 13.4: A | Statistical Table- F Distribution
- 13.5: B | Mathematical Phrases, Symbols, and Formulas

Index

Glossary

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

CHAPTER OVERVIEW

1: The Nature of Statistics

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

1.1: Introduction
1.2: Descriptive and Inferential Statistics
1.3: Variables and Types of Data
1.3.1: Levels of Measurement
1.4: Data Collection and Sampling Techniques
1.5: Computers and Calculators
1.5.1: Using Spreadsheets for Statistics
1.6: Experimental Design and Ethics
1.6.1: More on Experiments
1.6.2: Observational Studies and Sampling Strategies
1.E: Sampling and Data (Exercises)

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 1: The Nature of Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.1: Introduction

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."



Figure **1.1.1**: We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

This page titled 1.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.2: Descriptive and Inferential Statistics

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives.

Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5; 5.5; 6; 6; 6; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:

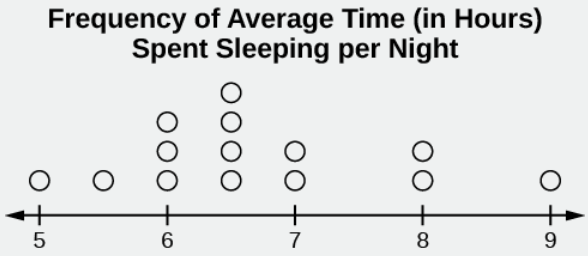


Figure **1.2.1**

- Does your dot plot look the same as or different from the example? Why?
- If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?
- Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses

probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

> The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example 1.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

- The **population** is all first year students attending ABC College this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).
- The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.
- The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.
- The **variable** could be the amount of money spent (excluding books) by one first year student. Let $\boldsymbol{X}$ = the amount of money spent (excluding books) by one first year student attending ABC College.
- The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

? Exercise 1.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Answer

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let $\boldsymbol{X}$ = the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

✓ Example 1.2.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1._____ Population 2._____ Statistic 3._____ Parameter 4._____ Sample 5._____ Variable 6._____ Data

a. all students who attended the college last year
b. the cumulative GPA of one student who graduated from the college last year
c. 3.65, 2.80, 1.50, 3.90
d. a group of students who graduated from the college last year, randomly selected
e. the average cumulative GPA of students who graduated from the college last year
f. all students who graduated from the college last year
g. the average cumulative GPA of students in the study who graduated from the college last year

Answer

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

✓ Example 1.2.3

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Answer

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.
- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.
- The **data** are either: yes, had head injury, or no, did not.

Example 1.2.4

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Answer

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.
- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

1. The Data and Story Library, https://dasl.datadescription.com/ (accessed May 1, 2013).

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once

they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B:

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

? Exercise 1.2.2

population

Answer

AIDS patients.

? Exercise 1.2.3

sample

? Exercise 1.2.4

parameter

Answer

The average length of time (in months) AIDS patients live after treatment.

? Exercise 1.2.5

statistic

? Exercise 1.2.6

variable

Answer

$X =$ the length of time (in months) AIDS patients live after treatment

Glossary

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative**(an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

---

This page titled 1.2: Descriptive and Inferential Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Variables and Types of Data

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of *counting* or *measuring* attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either *discrete* or *continuous*.

All data that are the result of counting are called *quantitative discrete data*. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the **quantitative discrete data**.

Exercise 1.3.1

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are **quantitative continuous data** because weights are measured.

Exercise 1.3.2

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Answer

quantitative continuous data

Exercise 1.3.3

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

Sample of qualitative data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are **qualitative data**.

Exercise 1.3.4

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Answer

qualitative data

Collaborative Exercise 1.3.1

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples' attitudes toward the government
n. IQ scores (This may cause some discussion.)

Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Exercise 1.3.5

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Answer

quantitative discrete

Exercise 1.3.6

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure 1.3.1. What type of data does this graph show?

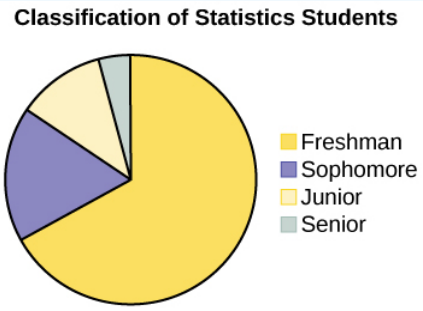


Figure 1.3.1

Answer

This pie chart shows the students in each year, which is **qualitative data**.

Exercise 1.3.7

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

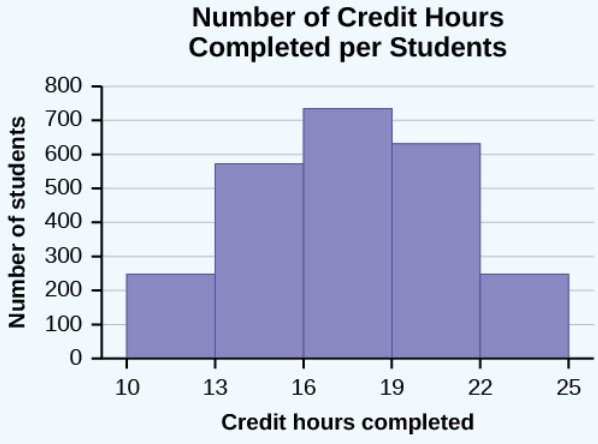


Figure 1.3.2

What type of data does this graph show?

Answer

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Table **1.3.1**: Fall Term 2007 (Census day)

De Anza College				**Foothill College**		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures **1.3.3** and **1.3.4** and determine which graph (pie or bar) you think displays the comparisons better.

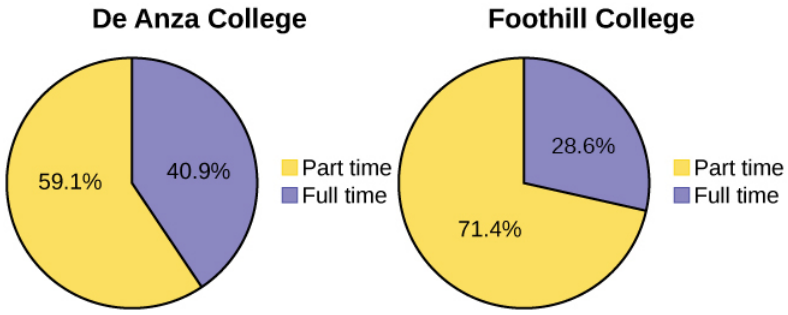


Figure **1.3.3**: Pie Charts

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

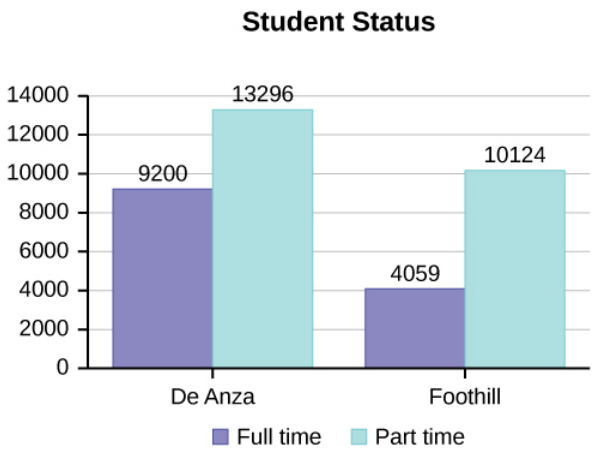


Figure **1.3.4**: Bar chart

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Table **1.3.2**: De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

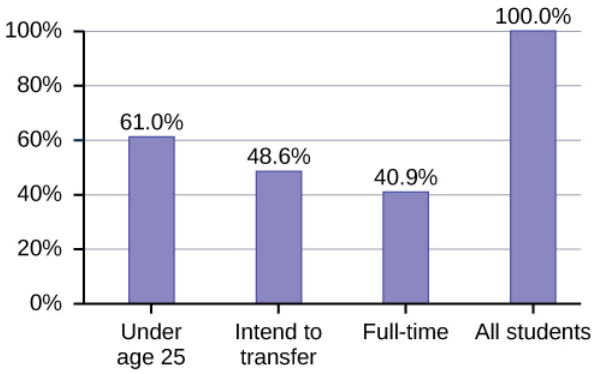


Figure **1.3.2**: Bar chart of data in Table **1.3.2**.

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Table **1.3.2**: Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

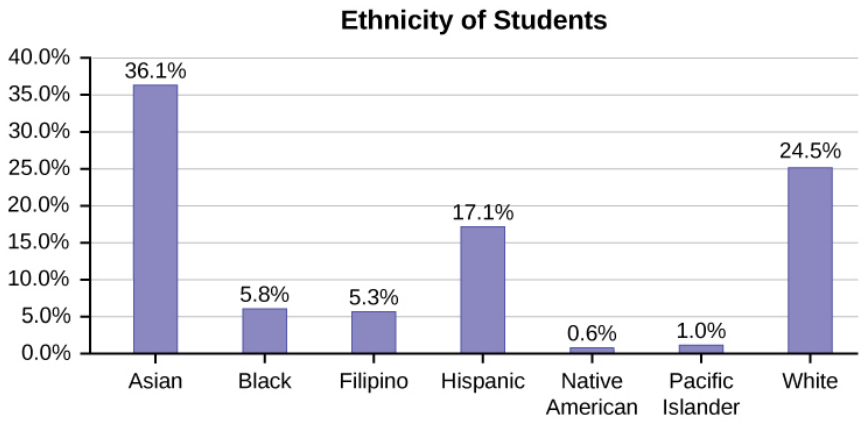


Figure **1.3.3**: Enrollment of De Anza College (Spring 2010)

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure **1.3.4** can be difficult to understand visually. The graph in Figure **1.3.5** is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

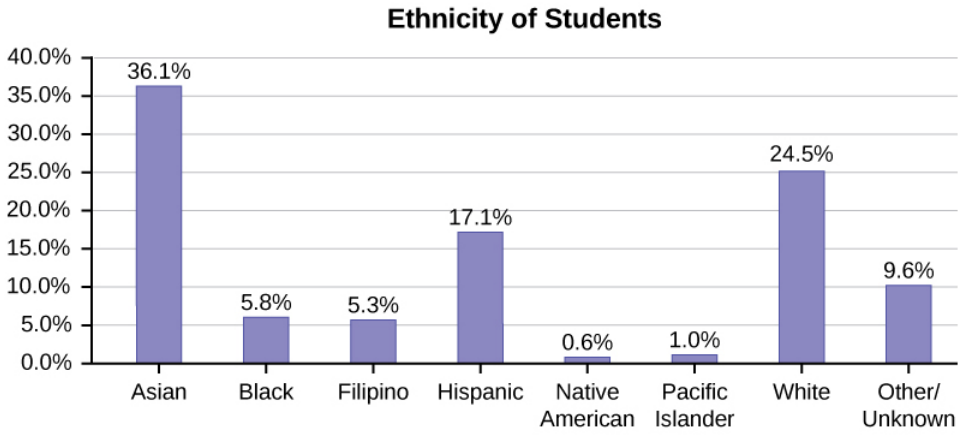


Figure **1.3.4**: Bar Graph with Other/Unknown Category

https://stats.libretexts.org/@go/page/10906

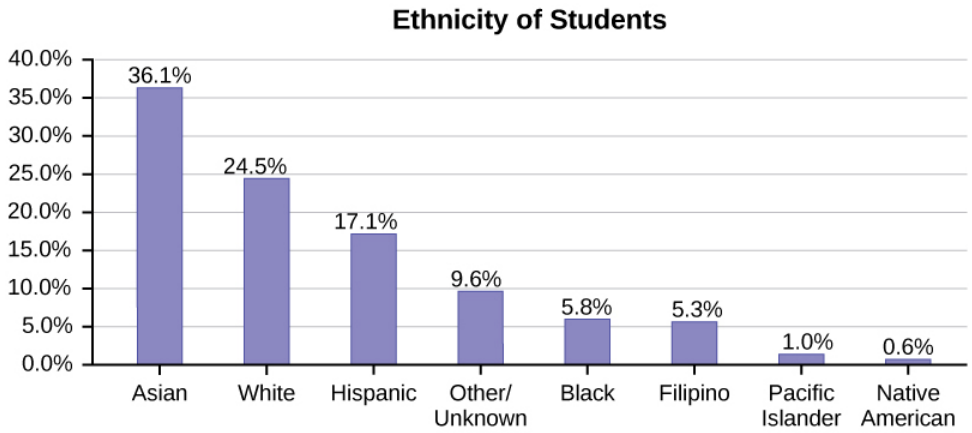


Figure 1.3.5: Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure 1.3.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 1.3.6.

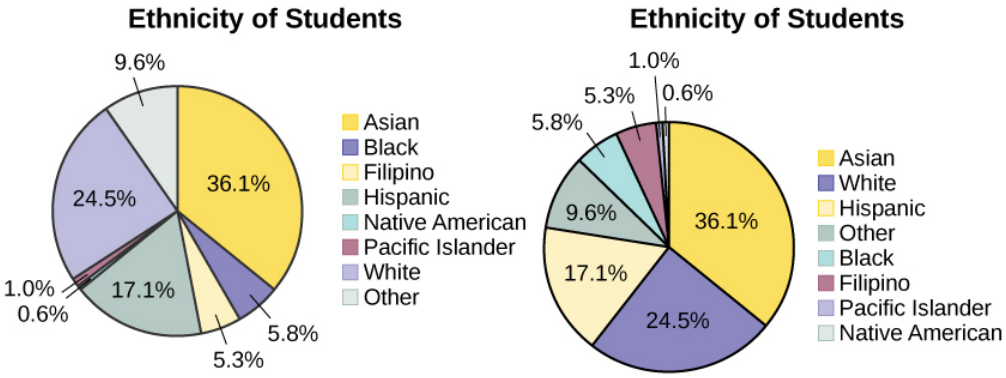


Figure 1.3.6.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.3.2:

Table 1.3.3: Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

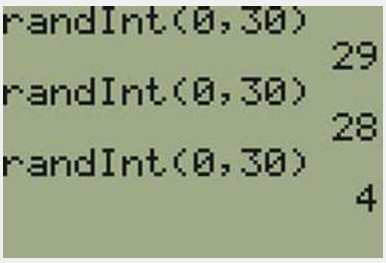


Figure 1.3.7

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Exercise 1.3.8

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Answer

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Example 1.3.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.
- Number each row one through ten.
- On your calculator, press Math and arrow over to PRB.
- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
- Press MATH and arrow over to PRB.
- Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
- The two numbers are for two of the columns.
- The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.
- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 1.3.10

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

? Exercise 1.3.11

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

✓ Example 1.3.12: Sampling

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task. Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Answer

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Answer

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

? Exercise 1.3.12

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Collaborative Exercise 1.3.8

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This *variability in samples* cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Collaborative Exercise **1.3.8**

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following table ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)		Second Experiment (20 rolls)	
Face on Die	**Frequency**	**Face on Die**	**Frequency**
1			
2			
3			
4			
5			
6			

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from www.bookofodds.com/Relationsh...-the-President
5. Dominic Lusinchi, "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).
6. "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).
9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples.

There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

This page titled 1.3: Variables and Types of Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.3.1: Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these. These will be explored in the next chapter.

Exercise 1.3.1.11

What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: Superior, Average, Above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box
d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store
j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65 year-old women
l. Common letter grades: A, B, C, D, and F

Answer

a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

References

1. “State & County QuickFacts,” U.S. Census Bureau. quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).
2. “State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).
3. “Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).
4. “Levels of Measurement,” infinity.cos.edu/faculty/wood...ata_Levels.htm (accessed May 1, 2013).
5. Courtney Taylor, “Levels of Measurement,” about.com, http://statistics.about.com/od/Helpa...easurement.htm (accessed May 1, 2013).
6. David Lane. “Levels of Measurement,” Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 1.3.1: Levels of Measurement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.4: Data Collection and Sampling Techniques

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider how data are collected so that they are reliable and help achieve the research goals.

Populations and samples

Consider the following three research questions:

1. What is the average mercury content in sword sh in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target population. In the rst question, the target **population** is all sword sh in the Atlantic ocean, and each sh represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 sword sh (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

Exercise

Exercise 1.7 For the second and third questions above, identify the target population and what represents an individual case.[10]

Anecdotal Evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating sword sh, so the average mercury concentration in sword sh must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each of the conclusions are based on some data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

[10] *(2) Notice that the rst question is only relevant to students who complete their degree; the average cannot be computed using a student who never nished her degree. Thus, only Duke undergraduate students who have graduated in the last ve years represent cases in the population under consideration. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.*



Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, “It’s one storm, in one region, of one country.”

Anecdotal evidence

Be careful of data collected in a haphazard fashion. Such evidence may be true and veri able, but it may only represent extraordinary cases.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Sampling from a Population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the population, and graduates who are selected for review are collectively called the sample. In general, we always seek to randomly select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates. Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

Example

Example 1.8 Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a biased sample, even if that bias is unintentional or difficult to discern.

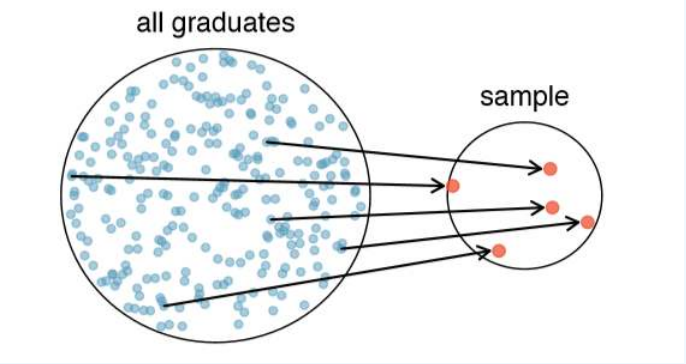


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

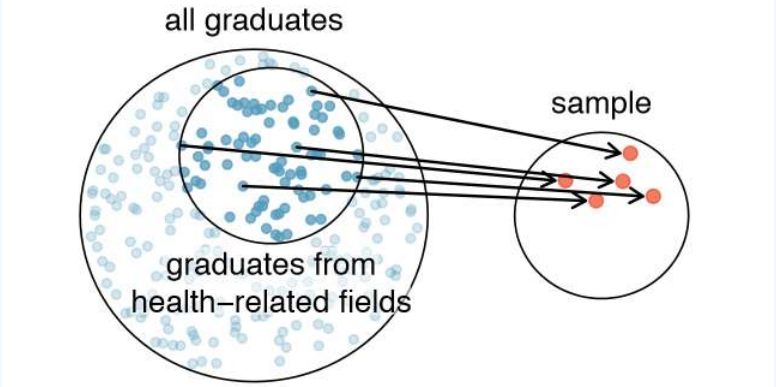


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health related majors disproportionally often.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces bias into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and it is the equivalent of using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

The act of taking a simple random sample helps minimize bias, however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

Another common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often diffcult to discern what sub-population a convenience sample represents.

Exercise

Exercise 1.9 We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[11]

[11]*Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind should data on the subject become available.*

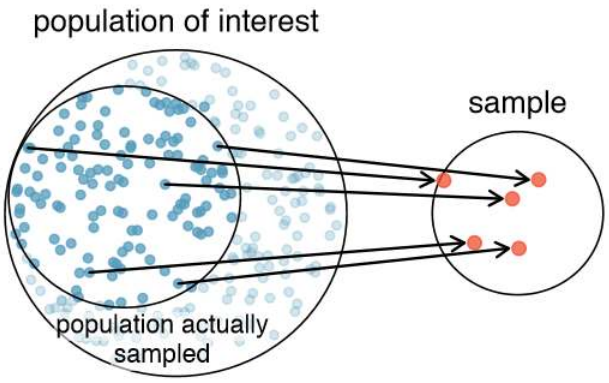


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely x this problem.

Explanatory and Response Variables

Consider the following question from page 7 for the county data set:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

If we suspect poverty might a ect spending in a county, then poverty is the explanatory variable and federal spending is the response variable in the relationship.[12] If there are many variables, it may be possible to consider a number of them as explanatory variables.

TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of a ecting the other and plan an appropriate analysis.

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable} \tag{1.4.1}$$

Caution: association does not imply causation

Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identi ed between the two variables. We use these labels only to keep track of which variable we suspect a ects the other.

In some cases, there is no explanatory or response variable. Consider the following question from page 7:

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

It is difficult to decide which of these variables should be considered the explanatory and response variable, i.e. the direction is ambiguous, so no explanatory or response labels are suggested here.

[12]*Sometimes the explanatory variable is called the independent variable and the response variable is called the dependent variable. However, this becomes confusing since a pair of variables might be independent or dependent, so we avoid this language.*

Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are assigned a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

TIP: association $\neq$ causation

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

This page titled 1.4: Data Collection and Sampling Techniques is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

- **1.4: Overview of Data Collection Principles** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.

1.5: Computers and Calculators

A Note About Technology

Many of the textbook examples are shown with graphing calculator steps. However, many different technologies can be used to perform statistics, including professional software packages Statisticians use (such as R, SPSS, or Minitab), and easily available spreadsheets like Microsoft Excel and Google Sheets.

Your instructor will guide you on what technology you should use for the class. The goal is to help you learn at least one technology for statistical analysis. If your instructor has not required a graphing calculator, then it is not expected that you know or understand how to enter commands in a graphing calculator to answer an example in the book that shows those commands. In using Excel or Sheets, your instructor will provide a lecture, video, or instructions on learning that application.

Beyond the class, one advantage to learning Excel or Sheets is that you will gain a marketable skill that you can include on your resume!

Please do not assume you must purchase a graphing calculator in order to complete this course. If your instructor requires a graphing calculator, the library has some to borrow and the Math Department has some to rent for the semester. Some instructors don't use graphing calculators at all, and some use them exclusively. Make sure you understand the technology requirements of your class by reviewing your syllabus and reaching out to your instructor for any clarification.

1.5: Computers and Calculators is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.5.1: Using Spreadsheets for Statistics

This content is from a Biological Statistics Textbook. The tips for using MS Excel (and also Google Sheets) are good tips for any statistics course.

Learning Objectives

- You can do most, maybe all of your statistics using a spreadsheet such as Excel. Here are some general tips.

Introduction

If you're like most biologists, you can do all of your statistics with spreadsheets such as Excel. You may spend months getting the most technologically sophisticated new biological techniques to work, but in the end you'll be able to analyze your data with a simple chi-squared test, ***t***–test, one-way anova or linear regression. The graphing abilities of spreadsheets make it easy to inspect data for errors and outliers, look for non-linear relationships and non-normal distributions, and display your final results. Even if you're going to use something like SAS or SPSS or ***R***, there will be many times when it's easier to enter your data into a spreadsheet first, inspect it for errors, sort and arrange it, then export it into a format suitable for your fancy-schmancy statistics package.

Some statisticians are contemptuous of Excel for statistics. One of their main complaints is that it can't do more sophisticated tests. While it is true that you can't do advanced statistics with Excel, that doesn't make it wrong to use it for simple statistics; that Excel can't do principal components analysis doesn't make its answer for a two-sample ***t***–test incorrect. If you are in a field that requires complicated multivariate analyses, such as epidemiology or ecology, you will definitely have to use something more advanced than spreadsheets. But if you are doing well designed, simple laboratory experiments, you may be able to analyze all of your data with the kinds of tests you can do in spreadsheets.

The more serious complaint about Excel is that some of the procedures gave incorrect results (McCullough and Heiser 2008, Yalta 2008). Most of these problems were with procedures more advanced than those covered in this handbook, such as exponential smoothing, or were errors in how Excel analyzes very unusual data sets that you're unlikely to get from a real experiment. After years of complaining, Microsoft finally fixed many of the problems in Excel 2010 (Keeling and Pavur 2011). So for the statistical tests I describe in this handbook, I feel confident that you can use Excel and get accurate results.

A free alternative to Excel is Calc, part of the free, open-source OpenOffice.org package. Calc does almost everything that Excel does, with just enough exceptions to be annoying. Calc will open Excel files and can save files in Excel format. The OpenOffice.org package is available for Windows, Mac, and Linux. OpenOffice.org also includes a word processor (like Word) and presentation software (like PowerPoint).

Gnumeric sounds like a good, free, open-source spreadsheet program; while it is primarily used by Linux users, it can be made to work with Mac. I haven't used it, so I don't know how well my spreadsheets will work with it.

The instructions on this web page apply to both Excel and Calc, unless otherwise noted.

Basic spreadsheet tasks

I'm going to assume you know how to enter data into a spreadsheet, copy and paste, insert and delete rows and columns, and other simple tasks. If you're a complete beginner, you may want to look at tutorials on using Excel here or here. Here are a few other things that will be useful for handling data.

Separate text into columns

Excel

When you copy columns of data from a web page or text document, then paste them into an Excel spreadsheet, all the data will be in one column. To put the data into multiple columns, select the cells you want to convert, then choose "Text to columns..." from the Data menu. If you choose "Delimited," you can tell it that the columns are separated by spaces, commas, or some other character. Check the "Treat consecutive delimiters as one" box (in Excel) or the "Merge Delimiters" box (in Calc) if numbers may

be separated by more than one space, more than one tab, etc. The data will be entered into the columns to the right of the original column, so make sure they're empty.

If you choose "Fixed width" instead of "Delimited", you can do things like tell it that the first 10 characters go in column 1, the next 7 characters go in column 2, and so on.

If you paste more text into the same Excel spreadsheet, it will automatically be separated into columns using the same delimiters. If you want to turn this off, select the column where you want to paste the data, choose "Text to columns..." from the Data menu, and choose "Delimited." Then unclick all the boxes for delimiters (spaces, commas, etc.) and click "Finish." Now paste your data into the column.

Series fill

You'll mainly use this for numbering a bunch of rows or columns. Numbering them will help you keep track of which row is which, and it will be especially useful if you want to sort the data, then put them back in their original order later. Put the first number of your series in a cell and select it, then choose "Fill: Series..." from the Edit menu. Choose "Rows" or "Columns" depending on whether you want the series to be in a row or a column, set the "Step value" (the amount the series goes up by; usually you'll use 1) and the "Stop value" (the last number in the series). So if you had a bunch of data in cells $B2$ through $E101$ and you wanted to number the rows, you'd put a 1 in cell $A2$, choose "Columns", set the "Step value" to 1 and the "Stop value" to 100, and the numbers 1 through 100 would be entered in cells $A2$ through $A101$.

Sorting

To sort a bunch of data, select the cells and choose "Sort" from the Data menu. If the first row of your data set has column headers identifying what is in each column, click on "My list has headers." You can sort by multiple columns; for example, you could sort data on a bunch of chickens by "Breed" in column A, "Sex" in column C, and "Weight" in column B, and it would sort the data by breeds, then within each breed have all the females first and then all the males, and within each breed/sex combination the chickens would be listed from smallest to largest.

If you've entered a bunch of data, it's a good idea to sort each column of numbers and look at the smallest and largest values. This may help you spot numbers with misplaced decimal points and other egregious typing errors, as they'll be much larger or much smaller than the correct numbers.

Graphing

See the web page on graphing with Excel. Drawing some quick graphs is another good way to check your data for weirdness. For example, if you've entered the height and leg length of a bunch of people, draw a quick graph with height on the X axis and leg length on the Y axis. The two variables should be pretty tightly correlated, so if you see some outlier who's 2.10 meters tall and has a leg that's only 0.65 meters long, you know to double-check the data for that person.

Absolute and relative cell references

In the formula "$=B1+C1$", $B1$ and $C1$ are relative cell references. If this formula is in cell $D1$, "$B1$" means "that cell that is two cells to the left." When you copy cell $D1$ into cell $D2$, the formula becomes "$=B2+C2$"; when you copy it into cell $G1$, it would become "$=E1+F1$". This is a great thing about spreadsheets; for example, if you have long columns of numbers in columns A and B and you want to know the sum of each pair, you don't need to type "$=B1+C1$" into cell $D1$, then type "$=B2+C2$" into cell $D2$, then type "$=B3+C3$" into cell $D3$, and so on; you just type "$=B1+C1$" once into cell $D1$, then copy and paste it into all the cells in column D at once.

Sometimes you don't want the cell references to change when you copy a cell; in that case, you should use absolute cell references, indicated with a dollar sign. A dollar sign before the letter means the column won't change when you copy and paste into a different cell. If you enter "$=\$B1+C1$" into cell $D1$, then copy it into cell $E1$, it will change to "$=\$B1+D1$"; the $C1$ will change to $D1$ because you've copied it one column over, but the $B1$ won't change because it has a dollar sign in front of it. A dollar sign before the number means the row won't change; if you enter "$=B\$1+C1$" into cell $D1$ and then copy it to cell $D2$, it will change to "$=B\$1+C2$". And a dollar sign before both the column and the row means that nothing will change; if you enter "$=\$B\$1+C1$" into cell $D2$ and then copy it into cell $E2$, it will change to "$=\$B\$1+D2$". So if you had 100 numbers in column B, you could enter

"$=B1-\text{AVERAGE}(B\$1:B\$100)$" in cell $C1$, copy it into cells $C2$ through $C100$, and each value in column B would have the average of the 100 numbers subtracted from it.

Paste Special

When a cell has a formula in it (such as "$=B1*C1+D1^2$"), you see the numerical result of the formula (such as "7.15") in the spreadsheet. If you copy and paste that cell, the formula will be pasted into the new cell; unless the formula only has absolute cell references, it will show a different numerical result. Even if you use only absolute cell references, the result of the formula will change every time you change the values in $B1$, $C1$ *or* $D1$. When you want to copy and paste the number that results from a function in **Excel**, choose "Paste Special" from the Edit menu and then click the button that says "Values." The number (7.15, in this example) will be pasted into the cell.

In **Calc**, choose "Paste Special" from the Edit menu, uncheck the boxes labeled "Paste All" and "Formulas," and check the box labeled "Numbers."

Change number format

The default format in Excel and Calc displays 9 digits to the right of the decimal point, if the column is wide enough. For example, the P value corresponding to a chi-square of 4.50 with 1 degree of freedom, found with "=CHIDIST(4.50, 1)", will be displayed as 0.033894854. This number of digits is almost always ridiculous. To change the number of decimal places that are displayed in a cell, choose "Cells..." from the Format menu, then choose the "Number" tab. Under "Category," choose "Number" and tell it how many decimal places you want to display. For the P value above, you'd probably just need three digits, 0.034. Note that this only changes the way the number is displayed; all of the digits are still in the cell, they're just invisible.

The disadvantage of setting the "Number" format to a fixed number of digits is that very small numbers will be rounded to 0. Thus if you set the format to three digits to the right of the decimal, "=CHIDIST(24.50,1)" will display as "0.000" when it's really 0.00000074. The default format ("General" format) automatically uses scientific notation for very small or large numbers, and will display $7.4309837243E-007$, which means 7.43×10^{-7}; that's better than just rounding to 0, but still has way too many digits. If you see a 0 in a spreadsheet where you expect a non-zero number (such as a P value), change the format to back to General.

For P values and other results in the spreadsheets linked to this handbook, I created a user-defined format that uses 6 digits right of the decimal point for larger numbers, and scientific notation for smaller numbers. I did this by choosing "Cells" from the Format menu and pasting the following into the box labeled "Format code":

$$[>0.00001]0.######;[<-0.00001]0.######;0.00E-00 \tag{1.5.1.1}$$

This will display 0 as $0.00E00$, but otherwise it works pretty well.

If a column is too narrow to display a number in the specified format, digits to the right of the decimal point will be rounded. If there are too many digits to the left of the decimal point to display them all, the cell will contain "###". Make sure your columns are wide enough to display all your numbers.

Useful spreadsheet functions

There are hundreds of functions in Excel and Calc; here are the ones that I find most useful for statistics and general data handling. Note that where the argument (the part in parentheses) of a function is "Y", it means a single number or a single cell in the spreadsheet. Where the argument says "Ys", it means more than one number or cell. See AVERAGE(Ys) for an example.

All of the examples here are given in Excel format. Calc uses a semicolon instead of a comma to separate multiple parameters; for example, Excel would use "=ROUND(A1, 2)" to return the value in cell $A1$ rounded to 2 decimal places, while Calc would use "=ROUND(A1; 2)". If you import an Excel file into Calc or export a Calc file to Excel format, Calc automatically converts between commas and semicolons. However, if you type a formula into Calc with a comma instead of a semicolon, Calc acts like it has no idea what you're talking about; all it says is "#NAME?".

I've typed the function names in all capital letters to make them stand out, but you can use lower case letters.

Math functions

ABS(Y) Returns the absolute value of a number.

EXP(Y) Returns e to the yth power. This is the inverse of LN, meaning that "=EXP(LN(Y))" equals Y.

LN(Y) Returns the natural logarithm (logarithm to the base e) of Y.

LOG10(Y) Returns the base-10 logarithm of Y. The inverse of LOG is raising 10 to the Yth power, meaning "=10^(LOG10(Y))" returns Y.

RAND() Returns a pseudorandom number, equal to or greater than zero and less than one. You must use empty parentheses so the spreadsheet knows that RAND is a function. For a pseudorandom number in some other range, just multiply; thus "=RAND()*79" would give you a number greater than or equal to 0 and less than 79. The value will change every time you enter something in any cell. One use of random numbers is for randomly assigning individuals to different treatments; you could enter "=RAND()" next to each individual, Copy and Paste Special the random numbers, Sort the individuals based on the column of random numbers, then assign the first 10 individuals to the placebo, the next 10 individuals to $10mg$ of the trial drug, etc.

A "pseudorandom" number is generated by a mathematical function; if you started with the same starting number (the "seed"), you'd get the same series of numbers. Excel's pseudorandom number generator bases its seed on the time given by the computer's internal clock, so you won't get the same seed twice. There are problems with Excel's pseudorandom number generator that make it inappropriate for serious Monte Carlo simulations, but the numbers it produces are random enough for anything you're likely to do as an experimental biologist.

ROUND(Y,digits) Returns Y rounded to the specified number of digits. For example, if cell $A1$ contains the number 37.38, "=ROUND(A1, 1)" returns 37.4, "=ROUND(A1, 0)" returns 37, and "=ROUND(A1, -1)" returns 40. Numbers ending in 5 are rounded up (away from zero), so "=ROUND(37.35,1)" returns 37.4 and "=ROUND(-37.35)" returns -37.4.

SQRT(Y) Returns the square root of Y.

SUM(Ys) Returns the sum of a set of numbers.

Logical functions

AND(logical_test1, logical_test2,...) Returns TRUE if logical_test1, logical_test2... are all true, otherwise returns FALSE. As an example, let's say that cells $A1$, $B1$ and $C1$ all contain numbers, and you want to know whether they're all greater than 100. One way to find out would be with the statement "=AND(A1>100, B1>100, C1>100)", which would return TRUE if all three were greater than 100 and FALSE if any one were not greater than 100.

IF(logical_test, A, B) Returns A if the logical test is true, B if it is false. As an example, let's say you have 1000 rows of data in columns A through E, with a unique ID number in column A, and you want to check for duplicates. Sort the data by column A, so if there are any duplicate ID numbers, they'll be adjacent. Then in cell $F1$, enter "=IF(A1=A2, "duplicate","ok"). This will enter the word "duplicate" if the number in $A1$ equals the number in $A2$; otherwise, it will enter the word "ok". Then copy this into cells $F2$ through $F999$. Now you can quickly scan through the rows and see where the duplicates are.

ISNUMBER(Y) Returns TRUE if Y is a number, otherwise returns FALSE. This can be useful for identifying cells with missing values. If you want to check the values in cells $A1$ to $A1000$ for missing data, you could enter "=IF(ISNUMBER(A1), "OK", "MISSING")" into cell $B1$, copy it into cells $B2$ to $B1000$, and then every cell in $A1$ that didn't contain a number would have "MISSING" next to it in column B.

OR(logical_test1, logical_test2,...) Returns TRUE if one or more of logical_test1, logical_test2... are true, otherwise returns FALSE. As an example, let's say that cells $A1$, $B1$ and $C1$ all contain numbers, and you want to know whether any is greater than 100. One way to find out would be with the statement "=OR(A1>100, B1>100, C1>100)", which would return TRUE if one or more were greater than 100 and FALSE if all three were not greater than 100.

Statistical functions

AVERAGE(Ys) Returns the arithmetic mean of a set of numbers. For example, "=AVERAGE(B1..B17)" would give the mean of the numbers in cells $B1..B17$, and "=AVERAGE(7, A1, B1..C17)" would give the mean of 7, the number in cell $A1$, and the numbers in the cells $B1..C17$. Note that Excel only counts those cells that have numbers in them; you could enter

"=AVERAGE(A1:A100)", put numbers in cells $A1$ to $A9$, and Excel would correctly compute the arithmetic mean of those 9 numbers. This is true for other functions that operate on a range of cells.

BINOMDIST(S, K, P, cumulative_probability) Returns the binomial probability of getting S "successes" in K trials, under the null hypothesis that the probability of a success is P. The argument "cumulative_probability" should be TRUE if you want the cumulative probability of getting S or fewer successes, while it should be FALSE if you want the probability of getting exactly S successes. (**Calc** uses 1 and 0 instead of TRUE and FALSE.) This has been renamed "BINOM.DIST" in newer versions of Excel, but you can still use "BINOMDIST".

CHIDIST(Y, df) Returns the probability associated with a variable, Y, that is chi-square distributed with df degrees of freedom. If you use SAS or some other program and it gives the result as "Chi-sq=78.34, 1 d.f., P<0.0001", you can use the CHIDIST function to figure out just how small your P value is; in this case, "=CHIDIST(78.34, 1)" yields 8.67×10^{-19}. This has been renamed CHISQ.DIST.RT in newer versions of Excel, but you can still use CHIDIST.

CONFIDENCE(alpha, standard-deviation, sample-size) Returns the confidence interval of a mean, *assuming you know the population standard deviation*. Because you don't know the population standard deviation, **you should never use this function**; instead, see the web page on confidence intervals for instructions on how to calculate the confidence interval correctly.

COUNT(Ys) Counts the number of cells in a range that contain numbers; if you've entered data into cells $A1$ through $A9$, $A11$, and $A17$, "=COUNT(A1:A100)" will yield 11.

COUNTIF(Ys, criterion) Counts the number of cells in a range that meet the given criterion.
"=COUNTIF(D1:E1100,50)" would count the number of cells in the range $D1:E100$ that were equal to 50;
"=COUNTIF(D1:E1100,">50")" would count the number of cells that had numbers greater than 50 (note the quotation marks around ">50");
"=COUNTIF(D1:E1100,F3)" would count the number of cells that had the same contents as cell $F3$;
"=COUNTIF(D1:E1100,"Bob")" would count the number of cells that contained just the word "Bob". You can use wildcards; "?" stands for exactly one character, so "Bo?" would count "Bob" or "Boo" but not "Bobble", while "Bo*" would count "Bob", "Boo", "Bobble" or "Bodacious".

DEVSQ(Ys) Returns the sum of squares of deviations of data points from the mean. This is what statisticians refer to as the "sum of squares." I use this in setting up spreadsheets to do anova, but you'll probably never need this.

FDIST(Y, df1, df2) Returns the probability value associated with a variable, Y, that is F-distributed with $df1$ degrees of freedom in the numerator and $df2$ degrees of freedom in the denominator. If you use SAS or some other program and it gives the result as "F=78.34, 1, 19 d.f., P<0.0001", you can use the FDIST function to figure out just how small your P value is; in this case, "=FDIST(78.34, 1, 19)" yields 3.62×10^{-8}. Newer versions of Excel call this function F.DIST.RT, but you can still use FDIST.

MEDIAN(Ys) Returns the median of a set of numbers. If the sample size is even, this returns the mean of the two middle numbers.

MIN(Ys) Returns the minimum of a set of numbers. Useful for finding the range, which is MAX(Ys)-MIN(Ys).

MAX(Ys) Returns the maximum of a set of numbers.

NORMINV(probability, mean, standard_deviation) Returns the inverse of the normal distribution for a given mean and standard deviation. This is useful for creating a set of random numbers that are normally distributed, which you can use for simulations and teaching demonstrations; if you paste "=NORMINV(RAND(),5,1.5)" into a range of cells, you'll get a set of random numbers that are normally distributed with a mean of 5 and a standard deviation of 1.5.

RANK.AVG(X, Ys, type) Returns the rank of X in the set of Ys. If *type* is set to 0, the largest number has a rank of 1; if *type* is set to 1, the smallest number has a rank of 1. For example, if cells $A1:A8$ contain the numbers $10, 12, 14, 14, 16, 17, 20, 21$ "=RANK(A2, A$1:A$8, 0)" returns 7 (the number 12 is the 7th largest in that list), and "=RANK(A2, A$1:A$8, 1)" returns 2 (it's the 2nd smallest).

The function "RANK.AVG" gives average ranks to ties; for the above set of numbers, "=RANK.AVG(A3, A$1:A$8, 0)" would return 5.5, because the two values of 14 are tied for fifth largest. Older versions of Excel and Calc don't have RANK.AVG; they have RANK, which handled ties incorrectly for statistical purposes. If you're using Calc or an older version of Excel, this formula shows how to get ranks with ties handled correctly:

$$= \text{AVERAGE}(\text{RANK}(A1,\ A\$1:A\$8,\ 0),\\ 1+\text{COUNT}(A\$1:A\$8)-\text{RANK}(A\$1,\ A\$1:A\$8,\ 1)) \tag{1.5.1.2}$$

STDEV(Ys) Returns an estimate of the standard deviation based on a population sample. This is the function you should use for standard deviation.

STDEVP(Ys) Returns the standard deviation of values from an entire population, not just a sample. **You should never use this function**.

SUM(Ys) Returns the sum of the Ys.

SUMSQ(Ys) Returns the sum of the squared values. Note that statisticians use "sum of squares" as a shorthand term for the sum of the squared deviations from the mean. SUMSQ does not give you the sum of squares in this statistical sense; for the statistical sum of squares, use DEVSQ. You will probably never use SUMSQ.

TDIST(Y, df, tails) Returns the probability value associated with a variable, Y, that is t-distributed with df degrees of freedom and *tails* equal to one or two (you'll almost always want the two-tailed test). If you use SAS or some other program and it gives the result as "t=78.34, 19 d.f., P<0.0001", you can use the TDIST function to figure out just how small your P value is; in this case, "=TDIST(78.34, 19, 2)" yields 2.55×10^{-25}. Newer versions of Excel have renamed this function T.DIST.2T, but you can still use TDIST.

VAR(Ys) Returns an estimate of the variance based on a population sample. This is the function you should use for variance.

VARP(Ys) Returns the variance of values from an entire population, not just a sample. **You should never use this function**.

References

Keeling, K.B., and R.J. Pavur. 2011. Statistical accuracy of spreadsheet software. American Statistician 65: 265-273.

McCullough, B.D., and D.A. Heiser. 2008. On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis 52: 4570-4578.

Yalta, A.T. 2008. The accuracy of statistical distributions in Microsoft Excel 2007. Computational Statistics and Data Analysis 52: 4579-4586.

This page titled 1.5.1: Using Spreadsheets for Statistics is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **7.1: Using Spreadsheets for Statistics** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

1.6: Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the response variable. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Example 1.6.1

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Answer

- The population is men aged 50 to 84

- The *population* is men aged 30 to 84.
- The *sample* is the 400 men who participated.
- The *experimental units* are the individual men in the study.
- The *explanatory variable* is oral medication.
- The *treatments* are aspirin and a placebo.
- The *response variable* is whether a subject had a heart attack.

Example 1.6.2

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Answer

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Example 1.6.3

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Answer

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Exercise 1.6.4

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Answer

a. Explanatory: presence of distraction from texting; response: response time measured in seconds

b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.
f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[2]

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.

- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

✔ Example 1.6.5

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Answer

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
c. It is never acceptable to fake data. Even though the responses she uses are "real" responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

? Exercise 1.6.6

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying "Most teens like Brand X as much as or more than Brand Y."

Answer

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

1. "Vitamin E and Health," Nutrition Source, Harvard School of Public Health, www.hsph.harvard.edu/nutritio...rce/vitamin-e/ (accessed May 1, 2013).
2. Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).
3. Ankita Mehta. "Daily Dose of Aspiring Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-as...s-study-300443 (accessed May 1, 2013).
4. The Data and Story Library, lib.stat.cmu.edu/DASL/Stories...dLearning.html (accessed May 1, 2013).
5. M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).
6. "Earthquake Information by Year," U.S. Geological Survey. earthquake.usgs.gov/earthquak...archives/year/ (accessed May 1, 2013).
7. "Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).
8. Data from www.businessweek.com (accessed May 1, 2013).
9. Data from www.forbes.com (accessed May 1, 2013).
10. "America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).
11. U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.
12. "April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), www.dot.gov/airconsumer/april...onsumer-report (accessed May 1, 2013).
13. Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).
14. Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

"An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule."[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

? Exercise 1.6.7

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

? Exercise 1.6.7

Discuss potential violations of the rule requiring informed consent.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children's allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Answer

a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

Footnotes

[1] McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

[2] Y.udhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/ma...src=dayp&_r=2& (accessed May 1, 2013).

[3] "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tillburg University, November 28, 2012, www.tilburguniversity.edu/upl...012_UK_web.pdf (accessed May 1, 2013).

[4] Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, http://www.stat.columbia.edu/~gelman...nceEthics1.pdf (accessed May 1, 2013).

Glossary

Explanatory Variable
the independent variable in an experiment; the value controlled by researchers

Treatments
different values or components of the explanatory variable applied in an experiment

Response Variable
the dependent variable in an experiment; the value that is measured for change at the end of an experiment

Experimental Unit
any individual or object to be measured

Lurking Variable
a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Random Assignment
the act of organizing experimental units into treatment groups using random methods

Control Group

a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Informed Consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board

a committee tasked with oversight of research programs that involve human subjects

Placebo

an inactive treatment that has no real effect on the explanatory variable

Blinding

not telling participants which treatment a subject is receiving

Double-blinding

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

---

This page titled 1.6: Experimental Design and Ethics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.6.1: More on Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g. using a coin ip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier nding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, inuence the response. Under these circumstances, they may rst group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might rst split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.15. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the rst three experimental design principles into any study, and this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.17 In particular, researchers wanted to know if the drug reduced deaths in patients.

[17]*Anturane Reinfarction Trial Research Group. 1980. Sul npyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.*

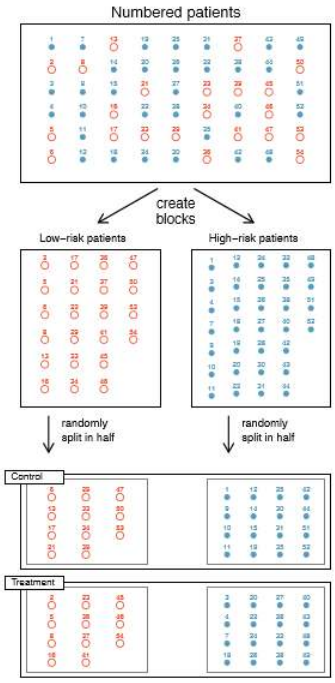


Figure 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly divided into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers18 were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[19]

Exercise 1.14 Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[20]

This page titled 1.6.1: More on Experiments is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

- **1.6: Experiments** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.

1.6.2: Observational Studies and Sampling Strategies

Observational Studies

Generally, data in observational studies are collected only by monitoring what occurs, what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers. Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations.

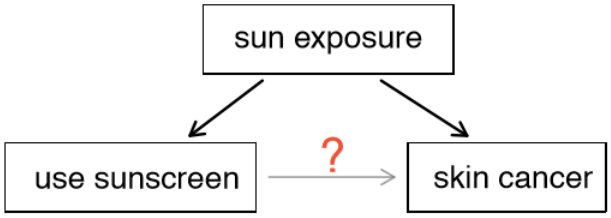
Exercise 1.6.2.1

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

Solution

No. See the paragraph following the exercise for an explanation.

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen and more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a **confounding variable** (also called a lurking variable, confounding factor, or a confounder), which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured. In the same way, the county data set is an observational study with confounding variables, and its data cannot easily be used to make causal conclusions.

Exercise 1.6.2.2

Figure 1.9 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in Figure 1.9.

Solution

Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses Health Study, started in 1976 and expanded in 1989. This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets, such as county, may contain both rospectively- and retrospectively-collected variables. Local governments

prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

Three Sampling Methods

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, these statistical methods are not reliable. Here we consider three random sampling techniques: simple, stratified, and cluster sampling. Figure 1.14 provides a graphical representation of these techniques.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's 828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the nal sample and knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called *strata*. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees). Then we might randomly sample 4 players from each team for a total of 120 players.

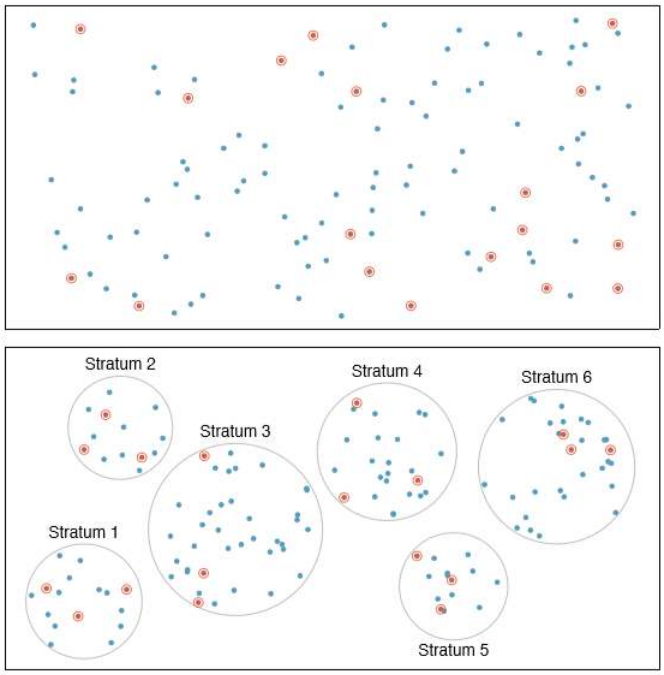


Figure 1.14: *Examples of simple random, stratified, and cluster sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the middle panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum. In the bottom panel, cluster sampling was used, where data were binned into nine clusters, three of the clusters were randomly selected, and six cases were randomly sampled in each of these clusters.*

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

Example 1.6.2.1

Why would it be good for cases within each stratum to be very similar?

Solution

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population.

A *cluster sample* is employed by taking all of the members from some of the groups. That is, we break up the population into many groups, called *clusters*. Then we sample a fixed number of clusters and use all the members within each cluster. If the population is large, instead of taking all members of the selected clusters, we can perform a simple random sample within the clusters. This technique is similar to stratified sampling in its process, except that there is no requirement in cluster sampling to sample from every cluster. Stratified sampling requires observations be sampled from every stratum.

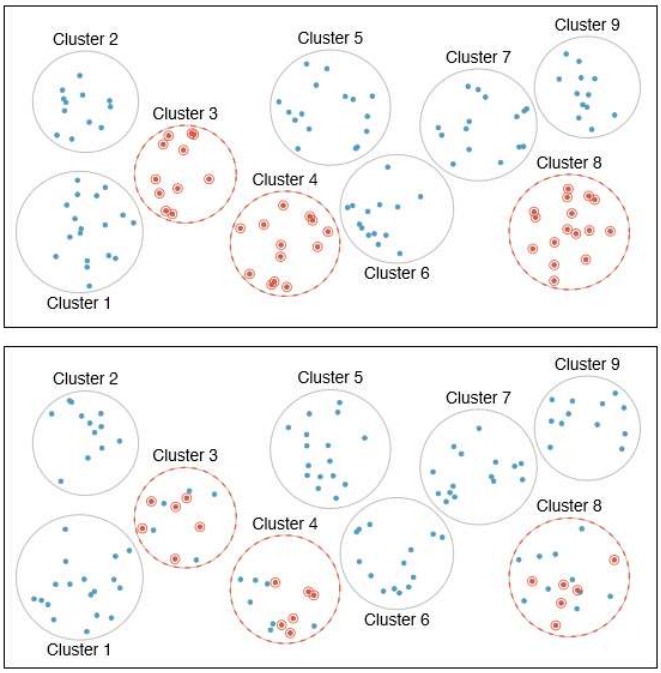


Figure 1.15: Examples of cluster and multistage sampling. In the top panel, cluster sampling was used. Here, data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used. It differs from cluster sampling in that of the clusters selected, we randomly select a subset of each cluster to be included in the sample.

Sometimes cluster sampling can be a more economical random sampling technique than the alternatives. Also, unlike stratified sampling, cluster sampling is most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then this sampling method works best when the neighborhoods are very diverse. A downside of cluster sampling is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

Example 1.6.2.3

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

Solution

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling seems like a very good idea. First, we might randomly select half the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us reliable information.

Contributors and Attributions

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

This page titled 1.6.2: Observational Studies and Sampling Strategies is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

1.E: Sampling and Data (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

1.1: Introduction

1.2: Definitions of Statistics, Probability, and Key Terms

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Q 1.2.1

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Q. 1.2.2

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

S 1.2.2

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. $X =$ the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

Q 1.2.3

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Q 1.2.4

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

S 1.2.5

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. $X =$ the health costs of one client
f. values for X, such as 34, 9, 82, and so on

Q 1.2.6

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Q 1.2.7

A marriage counselor is interested in the proportion of clients she counsels who stay married.

S 1.2.7

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. $X =$ the number of couples who stay married

f. yes, no

Q 1.2.8

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Q 1.2.9

A marketing company is interested in the proportion of people who will buy a particular product.

S 1.2.9

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. $X =$ the number of people who will buy it
f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Q 1.2.10

What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

Q 1.2.11

Consider the following:

$X =$ number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

S 1.2.12

a

Q 1.2.12

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

1.3: Data, Sampling, and Variation in Data and Sampling

Practice

Exercise 1.3.11

"Number of times per week" is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Use the following information to answer the next four exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

Exercise 1.3.12

The sampling method was

a. simple random
b. systematic
c. stratified
d. cluster

Answer

b

Exercise 1.3.13

"Duration (amount of time)" is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Exercise 1.3.14

The colors of the houses around the park are what kind of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Answer

a

Exercise 1.3.15

The population is _____

Exercise 1.3.16

Table contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Use Table to answer the following questions.

a. What is the proportion of deaths between 2007 and 2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that happened before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.

Answer

a. 0.5242
b. 0.03%
c. 6.86%
d. $\frac{823,088}{823,856}$
e. quantitative discrete
f. quantitative continuous
g. In both years, underwater earthquakes produced massive tsunamis.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

Exercise 1.3.17

A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

Exercise 1.3.18

A market researcher polls every tenth person who walks into a store.

Answer

systematic

Exercise 1.3.19

The first 50 people who walk into a sporting event are polled on their television preferences.

Exercise 1.3.20

A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Answer

simple random

Use the following information to answer the next seven exercises: Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Exercise 1.3.21

Complete the tables using the data provided:

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Researcher B

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Exercise 1.3.22

Determine what the key term data refers to in the above example for Researcher A.

Answer

values for X, such as 3, 4, 11, and so on

Exercise 1.3.23

List two reasons why the data may differ.

Exercise 1.3.24

Can you tell if one researcher is correct and the other one is incorrect? Why?

Answer

No, we do not have enough information to make such a claim.

Exercise 1.3.25

Would you expect the data to be identical? Why or why not?

Exercise 1.3.26

How might the researchers gather random data?

Answer

Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

Exercise 1.3.27

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise 1.3.28

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Answer

This would be convenience sampling and is not random.

Use the following data to answer the next five exercises: Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher B

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

Exercise 1.3.29

Give a reason why the data may differ.

Exercise 1.3.30

Would the sample size be large enough if the population is the students in the school?

Answer

Yes, the sample size of 150 would be large enough to reflect a population of one school.

Exercise 1.3.31

Would the sample size be large enough if the population is school-aged children and young adults in the United States?

Exercise 1.3.32

Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

Answer

Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

Exercise 1.3.33

As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in the first Table. The second study collected the data in the second Table.

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Exercise 1.3.34

Given what you know, which study is correct?

Answer

There is not enough information given to judge if either one is correct or incorrect.

Exercise 1.3.35

The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

Exercise 1.3.36

Both groups that performed the study concluded that the software works. Is this accurate?

Answer

The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

Exercise 1.3.37

The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

Exercise 1.3.38

Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from Exercise?

Answer

Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

Exercise 1.3.39

Is a sample size of 1,000 a reliable measure for a population of 5,000?

Exercise 1.3.40

Is a sample of 500 volunteers a reliable measure for a population of 2,500?

Answer

No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

Exercise 1.3.41

A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

Exercise 1.3.42

Is a sample size of two representative of a population of five?

Answer

No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

Exercise 1.3.43

Is it possible for two experiments to be well run with similar sample sizes to get different data?

Bringing It Together

Exercise 1.3.44

Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed below.

LBCC Distance Learning Survey Results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Exercise 1.3.45

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

Answer

Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

Q 1.3.1

number of tickets sold to a concert

S 1.3.1

quantitative discrete, 150

Q 1.3.2

percent of body fat

Q 1.3.3

favorite baseball team

S 1.3.3

qualitative, Oakland A's

Q 1.3.4

time in line to buy groceries

Q 1.3.5

number of students enrolled at Evergreen Valley College

S 1.3.5

quantitative discrete, 11,234 students

Q 1.3.6

most-watched television show

Q 1.3.7

brand of toothpaste

S 1.3.7

qualitative, Crest

Q 1.3.8

distance to the closest movie theater

Q 1.3.9

age of executives in Fortune 500 companies

S 1.3.9

quantitative continuous, 47.3 years

Q 1.3.10

number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Q 1.3.11

“Number of times per week” is what type of data?

1. qualitative
2. quantitative discrete
3. quantitative continuous

S 1.3.11

b

Q 1.3.12

“Duration (amount of time)” is what type of data?

1. qualitative
2. quantitative discrete
3. quantitative continuous

Q 1.3.13

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City

to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

S 1.3.13

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Q 1.3.14

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Q 1.3.15

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

S 1.3.15

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Q 1.3.16

List some practical difficulties involved in getting accurate results from a telephone survey.

Q 1.3.17

List some practical difficulties involved in getting accurate results from a mailed survey.

S 1.3.17

Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can’t be sure who is responding. In addition, mailing lists can be incomplete.

Q 1.3.18

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

Q 1.3.19

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

S 1.3.19

b

Q 1.3.20

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

Q 1.3.21

Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

S 1.3.21

a. convenience
b. cluster
c. stratified
d. systematic
e. simple random

Q 1.3.22

A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why? Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Q 1.3.23

The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

S 1.3.23

a. qualitative
b. quantitative discrete
c. quantitative discrete
d. qualitative

Q 1.3.24

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

Q 1.3.25

Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in [link] could explain this connection?

S 1.3.26

Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

Q 1.3.27

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?”[1]

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

Q 1.3.28

A scholarly article about response rates begins with the following quote:

“Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.”[2]

The Pew Research Center for People and the Press admits:

“The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more.”[3]

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

S 1.3.28

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

1.4: Frequency, Frequency Tables, and Levels of Measurement

Q 1.4.1

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

a. Fill in the blanks in Table.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Q 1.4.2

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in Table.

Flossing Frequency for Adults with Gum Disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

a. Fill in the blanks in Table.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

S 1.4.2

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

Q 1.4.3

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 45; 10 .

Table was produced.

Frequency of Immigrant Survey Responses

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	219219	0.1053
2	3	319319	0.2632
4	1	119119	0.3158
5	3	319319	0.4737
7	2	219219	0.5789
10	2	219219	0.6842
12	2	219219	0.7895
15	1	119119	0.8421
20	1	119119	1.0000

a. Fix the errors in Table. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Q 1.4.4

How much time does it take to travel to work? Table shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

S 1.4.4

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state’s travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

Q 1.4.5

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. Table shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?

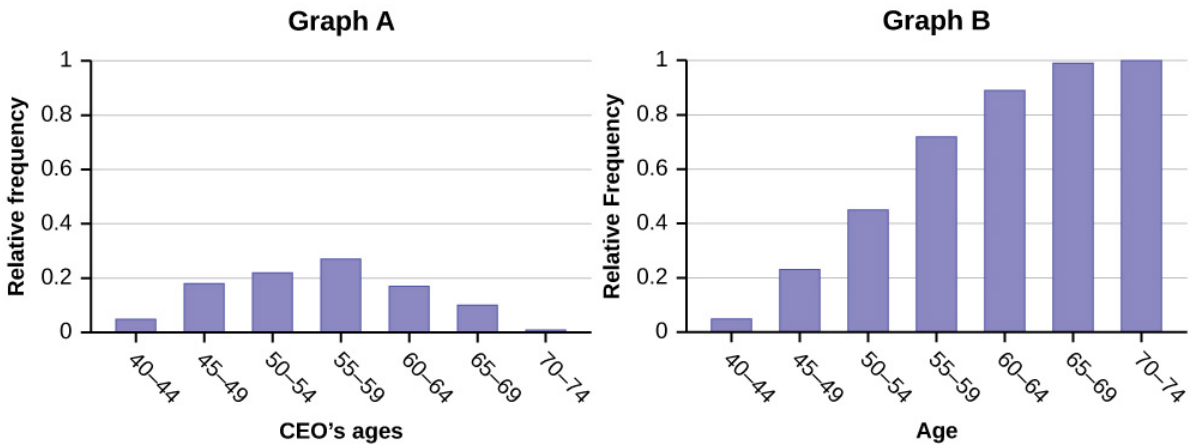


Figure 1.4.1. (a) Figure 1.4.1. (b)

Use the following information to answer the next two exercises: Table contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Frequency of Hurricane Direct Hits

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Q 1.4.6

What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

S 1.4.6

b

Q 1.4.7

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

1.5: Experimental Design and Ethics

Q 1.5.1

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

Use key terms from this module to describe the design of this experiment.

S 1.5.1

Explanatory variable: amount of sleep

Response variable: performance measured in assigned tasks

Treatments: normal sleep and 27 hours of total sleep deprivation

Experimental Units: 19 professional drivers

Lurking variables: none – all drivers participated in both treatments

Random assignment: treatments were assigned in random order; this eliminated the effect of any "learning" that may take place during the first experimental session

Control/Placebo: completing the experimental session under normal sleep conditions

Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time

Q 1.5.2

An advertisement for Acme Investments displays the two graphs in Figures to show the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

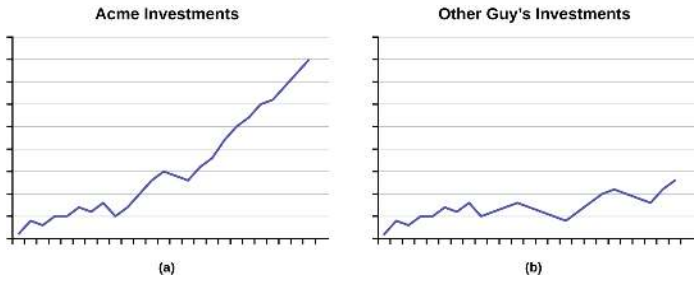


Figure 1.5.1. (a) Figure 1.5.1. (b)

As the graphs show, Acme consistently outperforms the Other Guys!

Q 1.5.3

The graph in Figure shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?

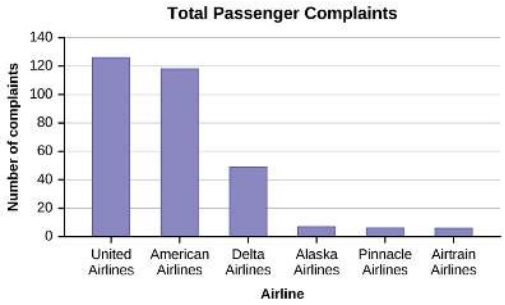


Figure 1.5.2.

https://stats.libretexts.org/@go/page/10914

S 1.5.3

You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

1.6: Data Collection Experiment

1.7: Sampling Experiment

This page titled 1.E: Sampling and Data (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **1.E: Sampling and Data (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

2: Frequency Distributions and Graphs

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics." You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

2.1: Prelude to Graphs

2.2: Histograms, Ogives, and Frequency Polygons

2.2.1: Frequency Polygons and Time Series Graphs

2.3: Other Types of Graphs

2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

2.3.2: Dot Plots

2.3.3: Guide to Fairly Good Graphs

2.3.4: Presenting Data in Tables

2.4: Organizing Data - Frequency Distributions

2.E: Graphs (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 2: Frequency Distributions and Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.1: Prelude to Graphs

Skills to Develop

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Create and interpret frequency distributions for categorical and grouped data.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.



Figure 2.1.1: When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

In this chapter and the next, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 2.1: Prelude to Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Histograms, Ogives, and Frequency Polygons

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A histogram consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The graph for quantitative data looks similar to a bar graph, except there are some major differences. First, in a bar graph the categories can be put in any order on the horizontal axis. There is no set order for these data values. You can't say how the data is distributed based on the shape, since the shape can change just by putting the categories in different orders. With quantitative data, the data are in specific orders, since you are dealing with numbers. With quantitative data, you can talk about a distribution, since the shape only changes a little bit depending on how many categories you set up. This is called a **frequency distribution**.

This leads to the second difference from bar graphs. In a bar graph, the categories that you made in the frequency table were determined by you. In quantitative data, the categories are numerical categories, and the numbers are determined by how many categories (or what are called classes) you choose. If two people have the same number of categories, then they will have the same frequency distribution. Whereas in qualitative data, there can be many different categories depending on the point of view of the author.

The third difference is that the categories touch with quantitative data, and there will be no gaps in the graph. The reason that bar graphs have gaps is to show that the categories do not continue on, like they do in quantitative data. Since the graph for quantitative data is different from qualitative data, it is given a new name. The name of the graph is a **histogram**. To create a histogram, you must first create the frequency distribution. The idea of a frequency distribution is to take the interval that the data spans and divide it up into equal subintervals called classes.

Summary of the steps involved in making a frequency distribution:

1. Find the range = largest value – smallest value
2. Pick the number of classes to use. Usually the number of classes is between five and twenty. Five classes are used if there are a small number of data points and twenty classes if there are a large number of data points (over 1000 data points). (Note: categories will now be called classes from now on.)
3. Class width = $\frac{\text{range}}{\#\text{ classes}}$ Always round up to the next integer (even if the answer is already a whole number go to the next integer). If you don't do this, your last class will not contain your largest data value, and you would have to add another class just for it. If you round up, then your largest data value will fall in the last class, and there are no issues.
4. Create the classes. Each class has limits that determine which values fall in each class. To find the class limits, set the smallest value as the lower class limit for the first class. Then add the class width to the lower class limit to get the next lower class limit. Repeat until you get all the classes. The upper class limit for a class is one less than the lower limit for the next class.
5. In order for the classes to actually touch, then one class needs to start where the previous one ends. This is known as the class boundary. To find the class boundaries, subtract 0.5 from the lower class limit and add 0.5 to the upper class limit.
6. Sometimes it is useful to find the class midpoint. The process is
   Midpoint $= \frac{\text{lower limit} + \text{upper limit}}{2}$
7. To figure out the number of data points that fall in each class, go through each data value and see which class boundaries it is between. Utilizing tally marks may be helpful in counting the data values. The frequency for a class is the number of data values that fall in the class.

Note

The above description is for data values that are whole numbers. If you data value has decimal places, then your class width should be rounded up to the nearest value with the same number of decimal places as the original data. In addition, your class boundaries should have one more decimal place than the original data. As an example, if your data have one decimal place,

then the class width would have one decimal place, and the class boundaries are formed by adding and subtracting 0.05 from each class limit.

Example 2.2.1 creating a frequency table

Table 2.2.1 contains the amount of rent paid every month for 24 students from a statistics course. Make a relative frequency distribution using 7 classes.

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

***Table 2.2.1**: Data of Monthly Rent*

Solution:

1. Find the range:
   largest value - smallest value $= 2550 - 350 = 2200$
2. Pick the number of classes:
   The directions to say to use 7 classes.
3. Find the class width:
   width $= \frac{\text{range}}{7} = \frac{2200}{7} \approx 314.286$
   Round up to 315
   Always round up to the next integer even if the width is already an integer.
4. Find the class limits:
   Start at the smallest value. This is the lower class limit for the first class. Add the width to get the lower limit of the next class. Keep adding the width to get all the lower limits.
   $350 + 315 = 665, 665 + 315 = 980, 980 + 315 = 1295 \rightleftharpoons,$
   The upper limit is one less than the next lower limit: so for the first class the upper class limit would be $665 - 1 = 664$.
   When you have all 7 classes, make sure the last number, in this case the 2550, is at least as large as the largest value in the data. If not, you made a mistake somewhere.
5. Find the class boundaries:
   Subtract 0.5 from the lower class limit to get the class boundaries. Add 0.5 to the upper class limit for the last class's boundary.
   $350 - 0.5 = 349.5, \quad 665 - 0.5 = 664.5, \quad 980 - 0.5 = 979.5, \quad 1295 - 0.5 = 1294.5 \rightleftharpoons$
   Every value in the data should fall into exactly one of the classes. No data values should fall right on the boundary of two classes.
6. Find the class midpoints:
   midpoint $= \frac{\text{lower limit} + \text{upper limit}}{2}$
   $\frac{350+664}{2} = 507, \frac{665+979}{2} = 822, \rightleftharpoons$
7. Tally and find the frequency of the data:
   Go through the data and put a tally mark in the appropriate class for each piece of data by looking to see which class boundaries the data value is between. Fill in the frequency by changing each of the tallies into a number.

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency
350-664	349.5-664.5	507	||||	4
665-979	664.5-979.5	822	𝍸 |||	8
980-1294	979.5-1294.5	1137	𝍸	5
1295-1609	1294.5-1609.5	1452	𝍸 |	6
1610-1924	1609.5-1924.5	1767		0
1925-2239	1924.5-2239.5	2082		0
2240-2554	2239.5-2554.5	2397	|	1

***Table 2.2.2**: Frequency Distribution for Monthly Rent*

Make sure the total of the frequencies is the same as the number of data points.

It is difficult to determine the basic shape of the distribution by looking at the frequency distribution. It would be easier to look at a graph. The graph of a frequency distribution for quantitative data is called a **frequency histogram** or just histogram for short.

Definition 2.2.1

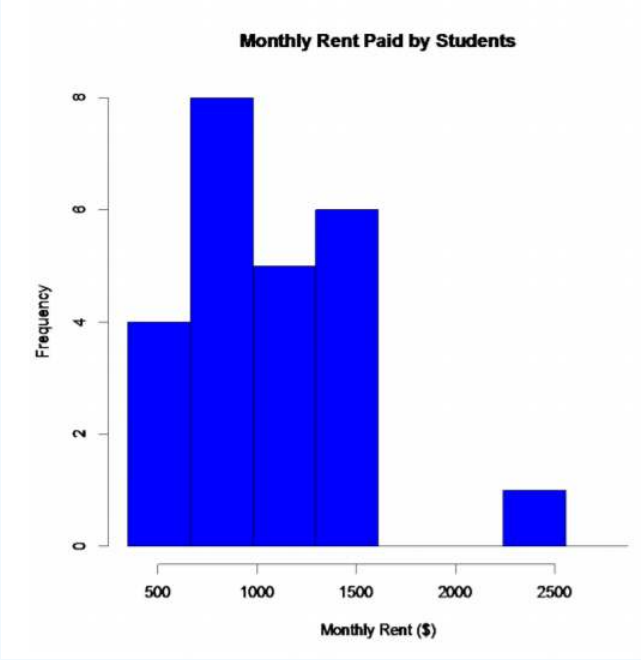
Histogram: a graph of the frequencies on the vertical axis and the class boundaries on the horizontal axis. Rectangles where the height is the frequency and the width is the class width are drawn for each class.

Example 2.2.2 drawing a histogram

Draw a histogram for the distribution from *Example 2.2.1*.

Solution:

The class boundaries are plotted on the horizontal axis and the frequencies are plotted on the vertical axis. You can plot the midpoints of the classes instead of the class boundaries. *Graph 2.2.1* was created using the midpoints because it was easier to do with the software that created the graph.



***Graph 2.2.1**: Histogram for Monthly Rent*

Notice the graph has the axes labeled, the tick marks are labeled on each axis, and there is a title. It is important that your graphs (all graphs) are clearly labeled.

Reviewing the graph you can see that most of the students pay around $750 per month for rent, with about $1500 being the other common value. You can see from the graph, that most students pay between $600 and $1600 per month for rent. Of course, these values are just estimates from the graph. There is a large gap between the $1500 class and the highest data value. This seems to say that one student is paying a great deal more than everyone else. This value could be considered an outlier. An **outlier** is a data value that is far from the rest of the values. It may be an unusual value or a mistake. It is a data value that should be investigated. In this case, the student lives in a very expensive part of town, thus the value is not a mistake, and is just very unusual. There are other aspects that can be discussed, but first some other concepts need to be introduced.

Frequencies are helpful, but understanding the relative size each class is to the total is also useful. To find this you can divide the frequency by the total to create a relative frequency. If you have the relative frequencies for all of the classes, then you have a relative frequency distribution.

Definition 2.2.2

Relative Frequency Distribution

A variation on a frequency distribution is a relative frequency distribution. Instead of giving the frequencies for each class, the relative frequencies are calculated.

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{\# of data points}}$$

This gives you percentages of data that fall in each class.

Example **2.2.3** creating a relative frequency table

Find the relative frequency for the grade data.

Solution:

From *Example 2.2.1*, the frequency distribution is reproduced in *Table 2.2.2*.

Class Limits	Class Boundaries	Class Midpoint	Frequency
350-664	349.5-664.5	507	4
665-979	664.5-979.5	822	8
980-1294	979.5-1294.5	1127	5
1295-1609	1294.5-1609.5	1452	6
1610-1924	1609.5-1924.5	1767	0
1925-2239	1924.5-2239.5	2082	0
2240-2554	2239.5-2554.5	2397	1

***Table 2.2.2**: Frequency Distribution for Monthly Rent*

Divide each frequency by the number of data points.

$$\frac{4}{24} = 0.17, \frac{8}{24} = 0.33, \frac{5}{24} = 0.21, \rightleftharpoons$$

Class Limits	Class Boundaries	Class Midpoint	Frequency	Relative Frequency
350-664	349.5-664.5	507	4	0.17
665-979	664.5-979.5	822	8	0.33
980-1294	979.5-1294.5	1127	5	0.21
1295-1609	1294.5-1609.5	1452	6	0.25
1610-1924	1609.5-1924.5	1767	0	0
1925-2239	1924.5-2239.5	2082	0	0
2240-2554	2239.5-2554.5	2397	1	0.04
Total			24	1

***Table 2.2.3**: Relative Frequency Distribution for Monthly Rent*

The relative frequencies should add up to 1 or 100%. (This might be off a little due to rounding errors.)

The graph of the relative frequency is known as a relative frequency histogram. It looks identical to the frequency histogram, but the vertical axis is relative frequency instead of just frequencies.

Example 2.2.4 drawing a relative frequency histogram

Draw a relative frequency histogram for the grade distribution from *Example 2.2.1*.

Solution:

The class boundaries are plotted on the horizontal axis and the relative frequencies are plotted on the vertical axis. (This is not easy to do in R, so use another technology to graph a relative frequency histogram.)



***Graph 2.2.2**: Relative Frequency Histogram for Monthly Rent*

Notice the shape is the same as the frequency distribution.

Another useful piece of information is how many data points fall below a particular class boundary. As an example, a teacher may want to know how many students received below an 80%, a doctor may want to know how many adults have cholesterol below 160, or a manager may want to know how many stores gross less than $2000 per day. This is known as a **cumulative frequency**. If you want to know what percent of the data falls below a certain class boundary, then this would be a **cumulative relative frequency**. For cumulative frequencies you are finding how many data values fall below the upper class limit.

To create a **cumulative frequency distribution**, count the number of data points that are below the upper class boundary, starting with the first class and working up to the top class. The last upper class boundary should have all of the data points below it. Also include the number of data points below the lowest class boundary, which is zero.

Example 2.2.5 creating a cumulative frequency distribution

Create a cumulative frequency distribution for the data in *Example 2.2.1*.

Solution:

The frequency distribution for the data is in *Table 2.2.2*.

Class Limits	Class Boundaries	Class Midpoint	Frequency
350-664	349.5-664.5	507	4
665-979	664.5-979.5	822	8
980-1294	979.5-1294.5	1127	5
1295-1609	1294.5-1609.5	1452	6
1610-1924	1609.5-1924.5	1767	0
1925-2239	1924.5-2239.5	2082	0
2240-2554	2239.5-2554.5	2397	1

***Table 2.2.2**: Frequency Distribution for Monthly Rent*

Now ask yourself how many data points fall below each class boundary. Below 349.5, there are 0 data points. Below 664.5 there are 4 data points, below 979.5, there are 4 + 8 = 12 data points, below 1294.5 there are 4 + 8 + 5 = 17 data points, and continue this process until you reach the upper class boundary. This is summarized in *Table 2.2.4*.

Class Limits	Class Boundaries	Class Midpoint	Frequency	Cumulative Frequency
350-664	349.5-664.5	507	4	4
665-979	664.5-979.5	822	8	12
980-1294	979.5-1294.5	1127	5	17
1295-1609	1294.5-1609.5	1452	6	23
1610-1924	1609.5-1924.5	1767	0	23
1925-2239	1924.5-2239.5	2082	0	23
2240-2554	2239.5-2554.5	2397	1	24

***Table 2.2.4**: Cumulative Distribution for Monthly Rent*

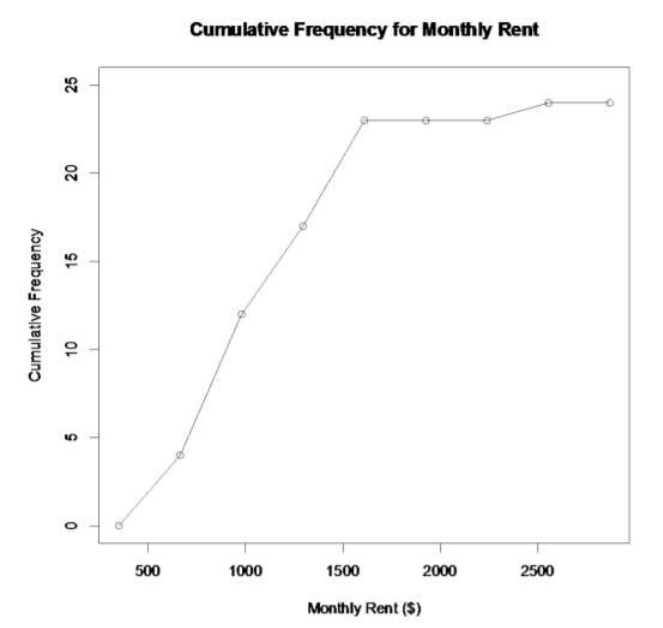
Again, it is hard to look at the data the way it is. A graph would be useful. The graph for cumulative frequency is called an **ogive** (o-jive). To create an ogive, first create a scale on both the horizontal and vertical axes that will fit the data. Then plot the points of the class upper class boundary versus the cumulative frequency. Make sure you include the point with the lowest class boundary and the 0 cumulative frequency. Then just connect the dots.

Example 2.2.6 drawing an ogive

Draw an ogive for the data in *Example 2.2.1*.

Solution:

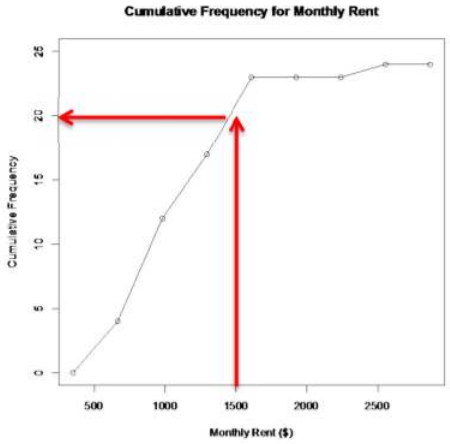
Using the upper class boundary and its corresponding cumulative frequency, plot the points as ordered pairs on the axes. Then connect the dots. You should have a line graph that rises as you move from left to right.



***Graph 2.2.3**: Ogive for Monthly Rent*

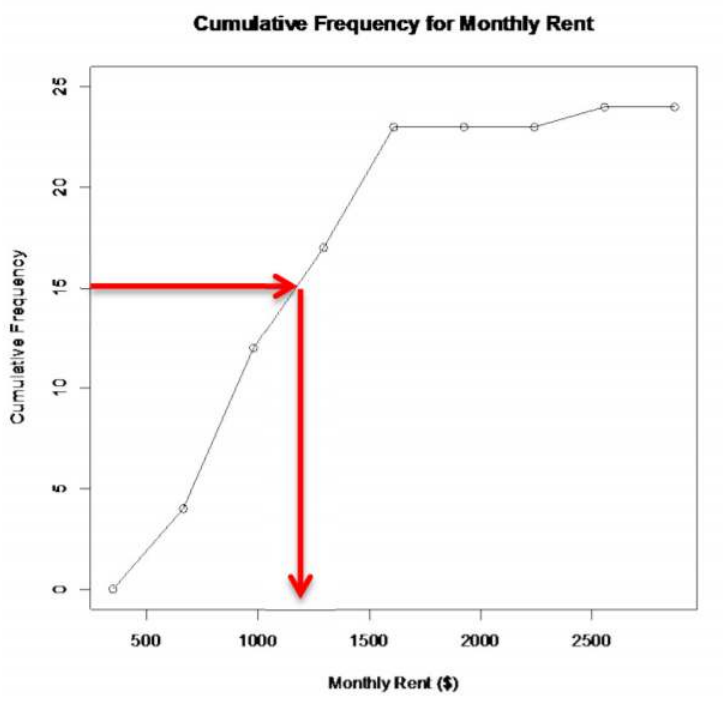
The usefulness of a ogive is to allow the reader to find out how many students pay less than a certain value, and also what amount of monthly rent is paid by a certain number of students. As an example, suppose you want to know how many students pay less than $1500 a month in rent, then you can go up from the $1500 until you hit the graph and then you go over to the cumulative

frequency axes to see what value corresponds to this value. It appears that around 20 students pay less than $1500. (See *Graph 2.2.4*.)



***Graph 2.2.4**: Ogive for Monthly Rent with Example*

Also, if you want to know the amount that 15 students pay less than, then you start at 15 on the vertical axis and then go over to the graph and down to the horizontal axis where the line intersects the graph. You can see that 15 students pay less than about $1200 a month. (See *Graph 2.2.5*.)



***Graph 2.2.5**: Ogive for Monthly Rent with Example*

If you graph the cumulative relative frequency then you can find out what percentage is below a certain number instead of just the number of people below a certain value.

Shapes of the distribution:

When you look at a distribution, look at the basic shape. There are some basic shapes that are seen in histograms. Realize though that some distributions have no shape. The common shapes are symmetric, skewed, and uniform. Another interest is how many peaks a graph may have. This is known as modal.

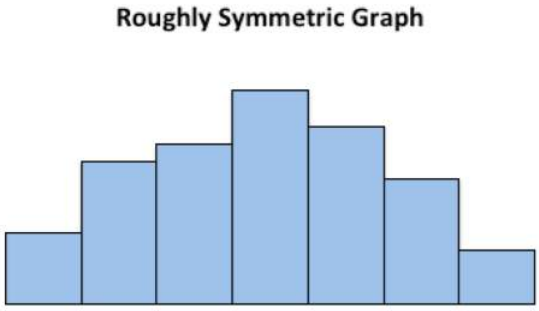
Symmetric means that you can fold the graph in half down the middle and the two sides will line up. You can think of the two sides as being mirror images of each other. Skewed means one “tail” of the graph is longer than the other. The graph is skewed in the direction of the longer tail (backwards from what you would expect). A uniform graph has all the bars the same height.

Modal refers to the number of peaks. Unimodal has one peak and bimodal has two peaks. Usually if a graph has more than two peaks, the modal information is not longer of interest.

Other important features to consider are gaps between bars, a repetitive pattern, how spread out is the data, and where the center of the graph is.

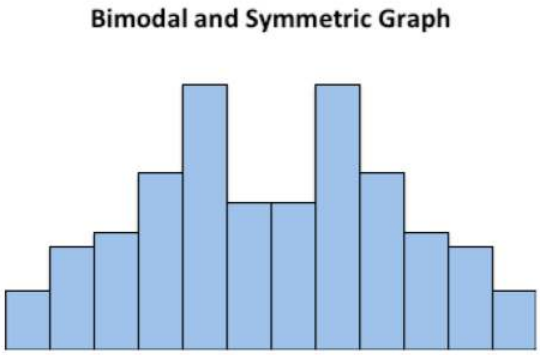
Examples of Graphs:

This graph is roughly symmetric and unimodal:



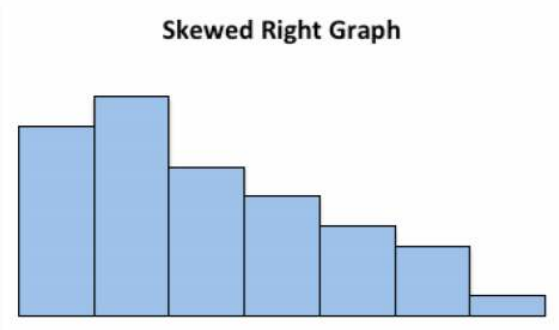
***Graph 2.2.6**: Symmetric, Unimodal Graph*

This graph is symmetric and bimodal:



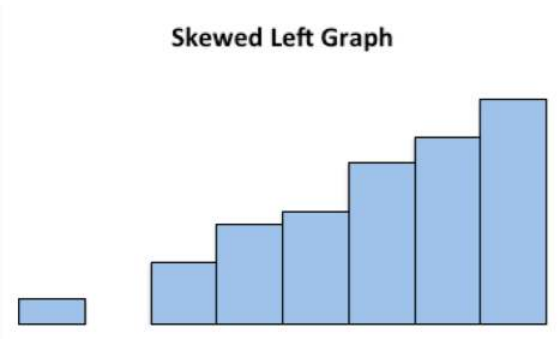
***Graph 2.2.7**: Symmetric, Bimodal Graph*

This graph is skewed to the right:



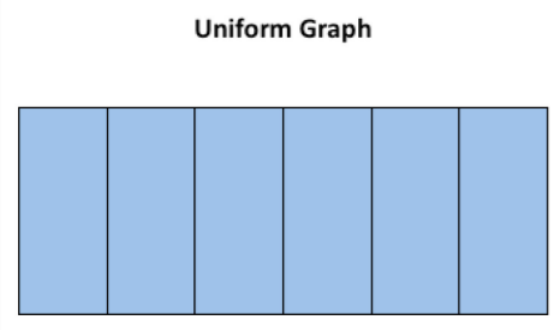
***Graph 2.2.8**: Skewed Right Graph*

This graph is skewed to the left and has a gap:



***Graph 2.2.9**: Skewed Left Graph*

This graph is uniform since all the bars are the same height:



***Graph 2.2.10**: Uniform Graph*

Example 2.2.7 creating a frequency distribution, histogram, and ogive

The following data represents the percent change in tuition levels at public, fouryear colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). Create a frequency distribution, histogram, and ogive for the data.

19.5%	40.8%	57.0%	15.1%	17.4%	5.2%	13.0%
15.6%	51.5%	15.6%	14.5%	22.4%	19.5%	31.3%
21.7%	27.0%	13.1%	26.8%	24.3%	38.0%	21.1%
9.3%	46.7%	14.5%	78.4%	67.3%	21.1%	22.4%
5.3%	17.3%	17.5%	36.6%	72.0%	63.2%	15.1%
2.2%	17.5%	36.7%	2.8%	16.2%	20.5%	17.8%
30.1%	63.6%	17.8%	23.2%	25.3%	21.4%	28.5%
9.4%						

***Table 2.2.5**: Data of Tuition Levels at Public, Four-Year Colleges*

Solution:

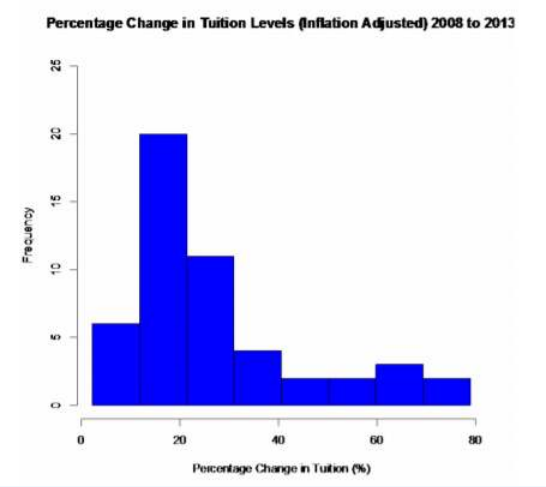
1. Find the range:
   largest value - smallest value = $78.4\% - -2.2\% = 76.2\%$
2. Pick the number of classes:
   Since there are 50 data points, then around 6 to 8 classes should be used. Let's use 8.

3. Find the class width:
   width $= \frac{\text{range}}{8} = \frac{76.2\%}{8} \approx 9.525\%$
   Since the data has one decimal place, then the class width should round to one decimal place. Make sure you round up.
   width $= 9.6\%$
4. Find the class limits:
   $2.2\% + 9.6\% = 11.8\%, 11.8\% + 9.6\% = 21.4\%, 21.4\% + 9.6\% = 31.0\%, \leftrightharpoons$
5. Find the class boundaries:
   Since the data has one decimal place, the class boundaries should have two decimal places, so subtract 0.05 from the lower class limit to get the class boundaries. Add 0.05 to the upper class limit for the last class's boundary.
   $2.2 - 0.05 = 2.15\%, 11.8 - 0.05 = 11.75\%, 21.4 - 0.05 = 21.35\% \leftrightharpoons$
   Every value in the data should fall into exactly one of the classes. No data values should fall right on the boundary of two classes.
6. Find the class midpoints:
   midpoint $= \frac{\text{lower limt + upper limit}}{2}$
   $\frac{2.2+11.7}{2} = 6.95\%, \frac{11.8+21.3}{2} = 16.55\%, \leftrightharpoons$
7. Tally and find the frequency of the data:

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency	Relative Frequency	Cumulative Frequency
2.2-11.7	2.15-11.75	6.95	𝍸 I	6	0.12	6
11.8-21.3	11.75-21.35	16.55	𝍸 𝍸 𝍸 𝍸	20	0.40	26
21.4-30.9	21.35-30.95	26.15	𝍸 𝍸 I	11	0.22	37
31.0-45.0	30.95-40.55	35.75	IIII	4	0.08	41
40.6-50.1	40.55-50.15	45.35	II	2	0.04	43
50.2-59.7	50.15-59.75	54.95	II	2	0.04	45
59.8-69.3	59.75-69.35	64.55	III	3	0.06	48
69.4-78.9	69.35-78.95	74.15	II	2	0.04	50

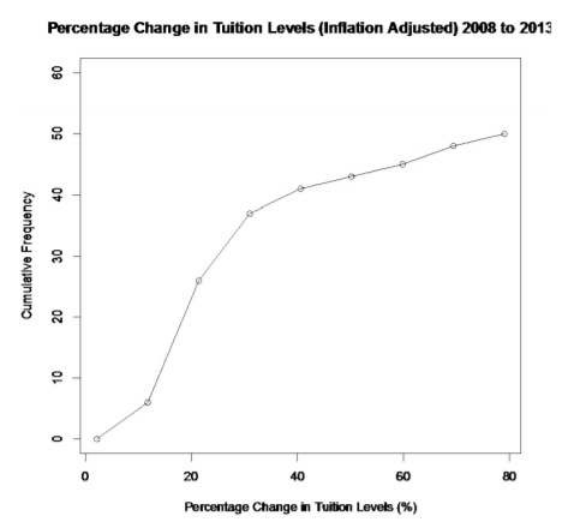
***Table 2.2.6**: Frequency Distribution for Tuition Levels at Public, Four-Year Colleges*

Make sure the total of the frequencies is the same as the number of data points.



***Graph 2.2.11**: Histogram for Tuition Levels at Public, Four-Year Colleges*

This graph is skewed right, with no gaps. This says that most percent increases in tuition were around 16.55%, with very few states having a percent increase greater than 45.35%.



***Graph 2.2.12**: Ogive for Tuition Levels at Public, Four-Year Colleges*

Looking at the ogive, you can see that 30 states had a percent change in tuition levels of about 25% or less.

There are occasions where the class limits in the frequency distribution are predetermined. *Example 2.2.8* demonstrates this situation.

Example 2.2.8 creating a frequency distribution and histogram

The following are the percentage grades of 25 students from a statistics course. Make a frequency distribution and histogram.

62	87	81	69	87	62	45	95	76	76
62	71	65	67	72	80	40	77	87	58
84	73	93	64	89					

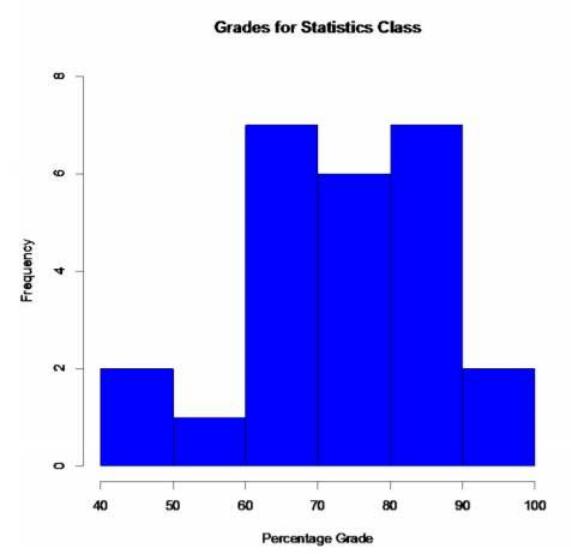
***Table 2.2.7**: Data of Test Grades*

Solution:

Since this data is percent grades, it makes more sense to make the classes in multiples of 10, since grades are usually 90 to 100%, 80 to 90%, and so forth. It is easier to not use the class boundaries, but instead use the class limits and think of the upper class limit being up to but not including the next classes lower limit. As an example the class 80 – 90 means a grade of 80% up to but not including a 90%. A student with an 89.9% would be in the 80-90 class.

Class Limit	Class Midpoint	Tally	Freqeuncy
40-50	45	||	2
50-60	55			1
60-70	65	𝍸 ||	7
70-80	75	𝍸		6
80-90	85	𝍸 ||	7
90-100	95	||	2

***Table 2.2.8**: Frequency Distribution for Test Grades*



***Graph 2.2.13**: Histogram for Test Grades*

It appears that most of the students had between 60 to 90%. This graph looks somewhat symmetric and also bimodal. The same number of students earned between 60 to 70% and 80 to 90%.

There are other types of graphs for quantitative data. They will be explored in the next section.

This page titled 2.2: Histograms, Ogives, and Frequency Polygons is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Histograms, Frequency Polygons, and Time Series Graphs** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

2.2.1: Frequency Polygons and Time Series Graphs

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons. To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

You can create a Frequency Polygon from a grouped frequency distribution by using the class midpoints for the *x*-axis values. Frequency polygons are "anchored" on the *x*-axis on both ends, as you can see below.

Example 2.2.1.4

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

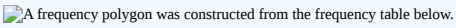


Figure **2.2.1.4**.

The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first “real” interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Exercise 2.2.1.4

Construct a frequency polygon of U.S. Presidents’ ages at inauguration shown in Table.

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Answer

The first label on the *x*-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 44 represents the next interval, or the first "real" interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

.

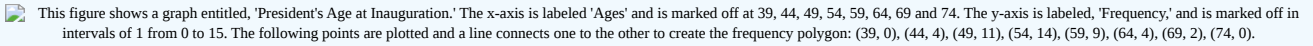

This figure shows a graph entitled, 'President's Age at Inauguration.' The x-axis is labeled 'Ages' and is marked off at 39, 44, 49, 54, 59, 64, 69 and 74. The y-axis is labeled, 'Frequency,' and is marked off in intervals of 1 from 0 to 15. The following points are plotted and a line connects one to the other to create the frequency polygon: (39, 0), (44, 4), (49, 11), (54, 14), (59, 9), (64, 4), (69, 2), (74, 0).

***Figure* 2.2.1.5.**

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 2.2.1.5

We will construct an overlay frequency polygon comparing the scores from Example with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

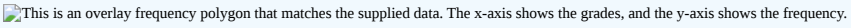

This is an overlay frequency polygon that matches the supplied data. The x-axis shows the grades, and the y-axis shows the frequency.

Figure **2.2.1.6**.

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example 2.2.1.6

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Answer

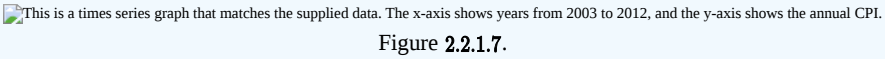


Figure 2.2.1.7.

Exercise 2.2.1.5

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO_2 Emissions

	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

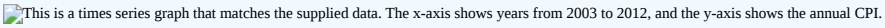

This is a times series graph that matches the supplied data. The x-axis shows years from 2003 to 2012, and the y-axis shows the annual CPI.

Figure 2.2.1.8.

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A **frequency polygon** can be used instead of a histogram when graphing large data sets with data points that repeat. The data usually goes on *x*-axis with the frequency being graphed on the *y*-axis. **Time series graphs** can be helpful when looking at large amounts of data for one variable over a period of time.

References

1. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker
2. "Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at www.scholastic.com/teachers/a...-us-presidents (accessed April 3, 2013).
3. "Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).
4. "Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).
5. "Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).
6. "CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).
7. "Births Time Series Data." General Register Office For Scotland, 2013. Available online at www.gro-scotland.gov.uk/stati...me-series.html (accessed April 3, 2013).
8. "Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

9. Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.
10. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Frequency

the number of times a value of the data occurs

Histogram

a graphical representation in $x-y$ form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 2.2.1: Frequency Polygons and Time Series Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Histograms, Frequency Polygons, and Time Series Graphs** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

2.3: Other Types of Graphs

Remember, qualitative data are words describing a characteristic of the individual. There are several different graphs that are used for qualitative data. These graphs include bar graphs, Pareto charts, and pie charts.

Pie charts and bar graphs are the most common ways of displaying qualitative data. A spreadsheet program like Excel can make both of them. The first step for either graph is to make a **frequency or relative frequency table**. A frequency table is a summary of the data with counts of how often a data value (or category) occurs.

Example 2.3.1

Suppose you have the following data for which type of car students at a college drive. How can we analyze the data?

> Ford, Chevy, Honda, Toyota, Toyota, Nissan, Kia, Nissan, Chevy, Toyota, Honda, Chevy, Toyota, Nissan, Ford, Toyota, Nissan, Mercedes, Chevy, Ford, Nissan, Toyota, Nissan, Ford, Chevy, Toyota, Nissan, Honda, Porsche, Hyundai, Chevy, Chevy, Honda, Toyota, Chevy, Ford, Nissan, Toyota, Chevy, Honda, Chevy, Saturn, Toyota, Chevy, Chevy, Nissan, Honda, Toyota, Toyota, Nissan

Solution:

A listing of data is too hard to look at and analyze, so you need to summarize it. First you need to decide the categories. In this case it is relatively easy; just use the car type. However, there are several cars that only have one car in the list. In that case it is easier to make a category called other for the ones with low values. Now just count how many of each type of cars there are. For example, there are 5 Fords, 12 Chevys, and 6 Hondas. This can be put in a frequency distribution:

Cateogry	Frequency
Ford	5
Chevy	12
Honda	6
Toyota	12
Nissan	10
Other	5
Total	50

***Table 2.1.1**: Frequency Table for Type of Car Data*

The total of the frequency column should be the number of observations in the data.

Since raw numbers are not as useful to tell other people it is better to create a third column that gives the relative frequency of each category. This is just the frequency divided by the total. As an example for Ford category:

$$\text{relative frequency} = \frac{5}{50} = 0.10$$

This can be written as a decimal, fraction, or percent. You now have a relative frequency distribution:

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

***Table 2.1.2**: Relative Frequency Table for Type of Car Data*

The relative frequency column should add up to 1.00. It might be off a little due to rounding errors.

Now that you have the frequency and relative frequency table, it would be good to display this data using a graph. There are several different types of graphs that can be used: bar chart, pie chart, and Pareto charts.

Bar graphs or charts consist of the frequencies on one axis and the categories on the other axis. Then you draw rectangles for each category with a height (if frequency is on the vertical axis) or length (if frequency is on the horizontal axis) that is equal to the frequency. All of the rectangles should be the same width, and there should be equally width gaps between each bar.

Example 2.3.2 drawing a bar graph

Draw a bar graph of the data in *Example 2.1.1*.

Solution:

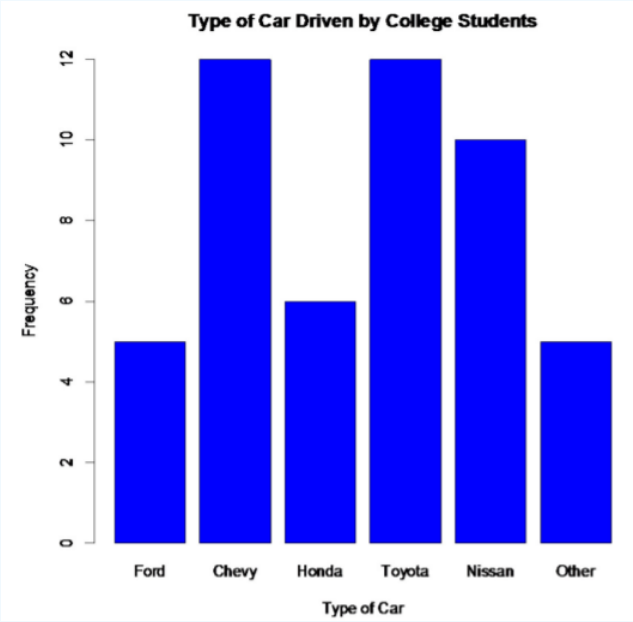
Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

***Table 2.1.2**: Relative Frequency Table for Type of Car Data*

Put the frequency on the vertical axis and the category on the horizontal axis.

Then just draw a box above each category whose height is the frequency.

You can also use MS Excel or Google Sheets to create a bar graph from the frequency table.



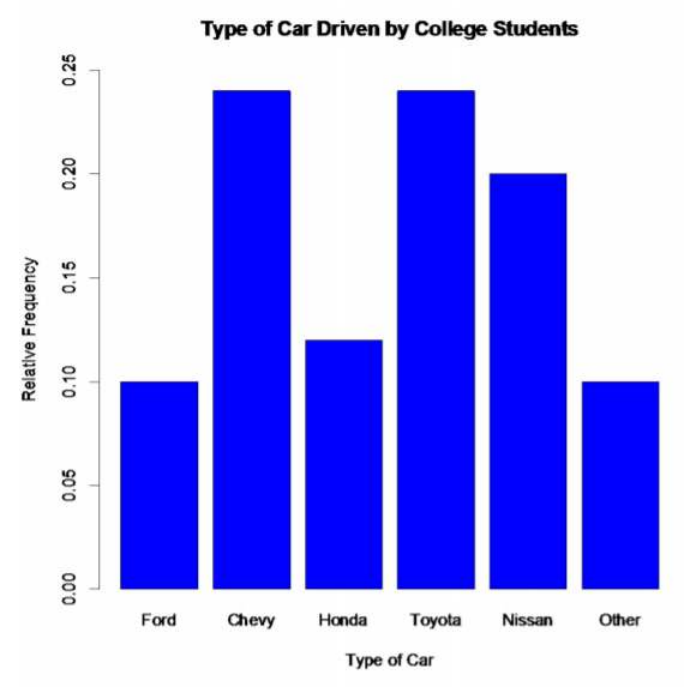
***Graph 2.1.1**: Bar Graph for Type of Car Data*

Notice from the graph, you can see that Toyota and Chevy are the more popular car, with Nissan not far behind. Ford seems to be the type of car that you can tell was the least liked, though the cars in the other category would be liked less than a Ford.

Some key features of a bar graph:

- Equal spacing on each axis.
- Bars are the same width.
- There should be labels on each axis and a title for the graph.
- There should be a scaling on the frequency axis and the categories should be listed on the category axis.
- The bars don't touch.

You can also draw a bar graph using relative frequency on the vertical axis. This is useful when you want to compare two samples with different sample sizes. The relative frequency graph and the frequency graph should look the same, except for the scaling on the frequency axis.



***Graph 2.1.2**: Relative Frequency Bar Graph for Type of Car Data*

Another type of graph for qualitative data is a pie chart. A pie chart is where you have a circle and you divide pieces of the circle into pie shapes that are proportional to the size of the relative frequency. There are 360 degrees in a full circle. Relative frequency is just the percentage as a decimal. All you have to do to find the angle by multiplying the relative frequency by 360 degrees. Remember that 180 degrees is half a circle and 90 degrees is a quarter of a circle.

Example 2.3.3 drawing a pie chart

Draw a pie chart of the data in *Example 2.1.1*.

First you need the relative frequencies.

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

***Table 2.1.2**: Relative Frequency Table for Type of Car Data*

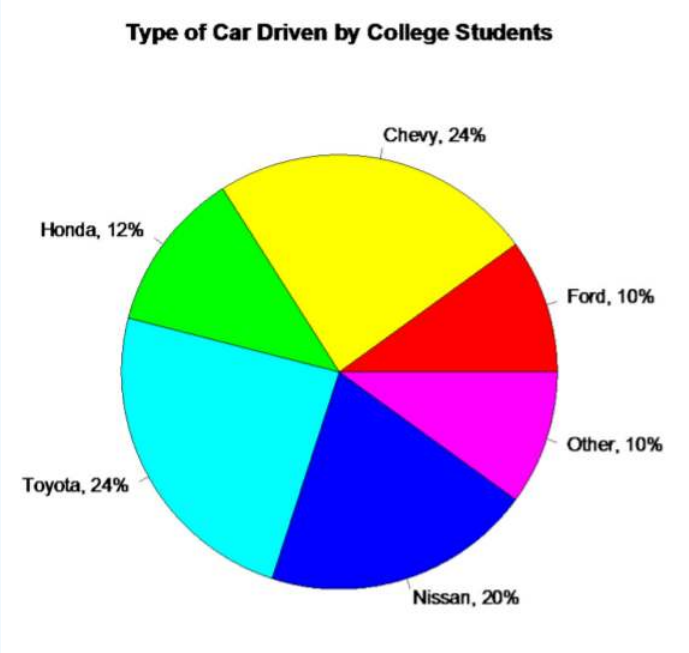
Solution:

Then you multiply each relative frequency by 360° to obtain the angle measure for each category.

Category	Relative Frequency	Angle (in degrees (°))
Ford	0.10	36.0
Chevy	0.24	86.4
Honda	0.12	43.2
Toyota	0.24	86.4
Nissan	0.20	72.0
Other	0.10	36.0
Total	1.00	360.0

***Table 2.1.3**: Pie Chart Angles for Type of Car Data*

Now draw the pie chart using a compass, protractor, and straight edge. Technology is preferred. If you use technology, there is no need for the relative frequencies or the angles. Technology like MS Excel or Google Sheets will create pie charts very quickly.

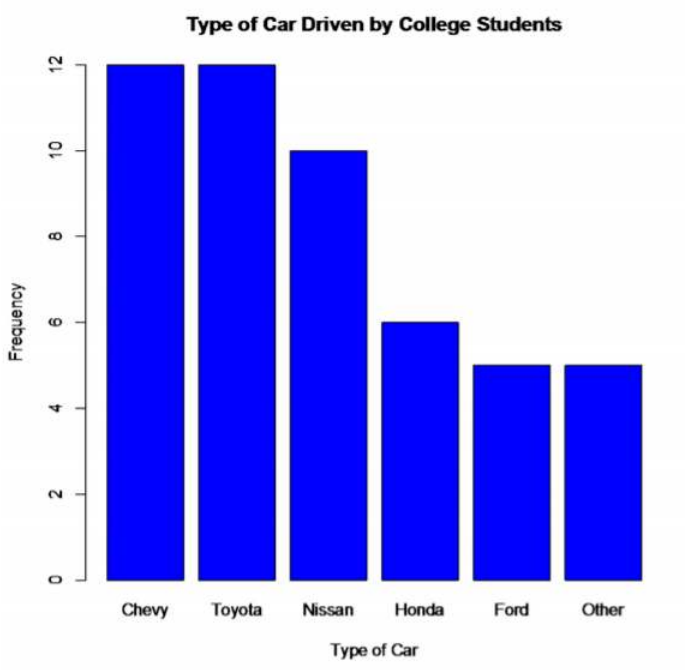


***Graph 2.1.3**: Pie Chart for Type of Car Data*

As you can see from the graph, Toyota and Chevy are more popular, while the cars in the other category are liked the least. Of the cars that you can determine from the graph, Ford is liked less than the others.

Pie charts are useful for comparing sizes of categories. Bar charts show similar information. It really doesn't matter which one you use. It really is a personal preference and also what information you are trying to address. However, pie charts are best when you only have a few categories and the data can be expressed as a percentage. The data doesn't have to be percentages to draw the pie chart, but if a data value can fit into multiple categories, you cannot use a pie chart. As an example, if you are asking people about what their favorite national park is, and you say to pick the top three choices, then the total number of answers can add up to more than 100% of the people involved. So you cannot use a pie chart to display the favorite national park.

A third type of qualitative data graph is a **Pareto chart**, which is just a bar chart with the bars sorted with the highest frequencies on the left. Here is the Pareto chart for the data in *Example 2.1.1*.



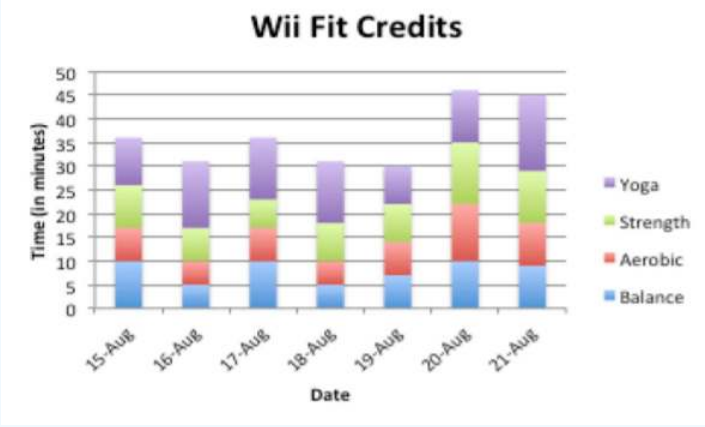
***Graph 2.1.4**: Pareto Chart for Type of Car Data*

The advantage of Pareto charts is that you can visually see the more popular answer to the least popular. This is especially useful in business applications, where you want to know what services your customers like the most, what processes result in more injuries, which issues employees find more important, and other type of questions like these.

There are many other types of graphs that can be used on qualitative data. There are spreadsheet software packages that will create most of them, and it is better to look at them to see what can be done. It depends on your data as to which may be useful. The next example illustrates one of these types known as a multiple bar graph.

Example 2.3.4 multiple bar graph

In the Wii Fit game, you can do four different types of exercises: yoga, strength, aerobic, and balance. The Wii system keeps track of how many minutes you spend on each of the exercises everyday. The following graph is the data for Dylan over one week time period. Discuss any indication you can infer from the graph.



***Graph 2.1.5**: Multiple Bar Chart for Wii Fit Data*

Solution:

It appears that Dylan spends more time on yoga exercises than on any other exercises on any given day. He seems to spend less time on strength exercises on a given day. There are several days when the amount of exercise in the different categories is

almost equal.

The usefulness of a multiple bar graph is the ability to compare several different categories over another variable, in *Example 2.1.4* the variable would be time. This allows a person to interpret the data with a little more ease.

Homework

1. Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in *Table 2.1.4*.

Activity	Grind	Multicoat	Assemble	Make frames	Receive finished	Unknown
Number of lenses	18872	12105	4333	25880	26991	1508

***Table 2.1.4**: Data for Eyeglassomatic*

Grind means that they ground the lenses and put them in frames, multicoat means that they put tinting or scratch resistance coatings on lenses and then put them in frames, assemble means that they receive frames and lenses from other sources and put them together, make frames means that they make the frames and put lenses in from other sources, receive finished means that they received glasses from other source, and unknown means they do not know where the lenses came from. Make a bar chart and a pie chart of this data. State any findings you can see from the graphs.

2. To analyze how Arizona workers ages 16 or older travel to work the percentage of workers using carpool, private vehicle (alone), and public transportation was collected. Create a bar chart and pie chart of the data in *Table 2.1.5*. State any findings you can see from the graphs.

Transportation type	Percentage
Carpool	11.6%
Private Vehicle (Alone)	75.8%
Public Transportation	2.0%
Other	10.6%

***Table 2.1.5**: Data of Travel Mode for Arizona Workers*

3. The number of deaths in the US due to carbon monoxide (CO) poisoning from generators from the years 1999 to 2011 are in table #2.1.6 (Hinatov, 2012). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Region	Number of Deaths from CO While Using a Generator
Urban Core	401
Sub-Urban	97
Large Rural	86
Small Rural/Isolated	111

***Table 2.1.6**: Data of Number of Deaths Due to CO Poisoning*

4. In Connecticut households use gas, fuel oil, or electricity as a heating source. *Table 2.1.7* shows the percentage of households that use one of these as their principle heating sources ("Electricity usage," 2013), ("Fuel oil usage," 2013), ("Gas usage," 2013). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Heating Source	Percentage
Electricity	15.3%
Fuel Oil	46.3%
Gas	35.6%
Other	2.85

***Table 2.1.7**: Data of Household Heating Sources*

5. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made during the time period of January 1 to March 31. *Table 2.1.8* gives the defect and the number of defects. Create a Pareto chart of the data and then describe what this tells you about what causes the most defects.

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

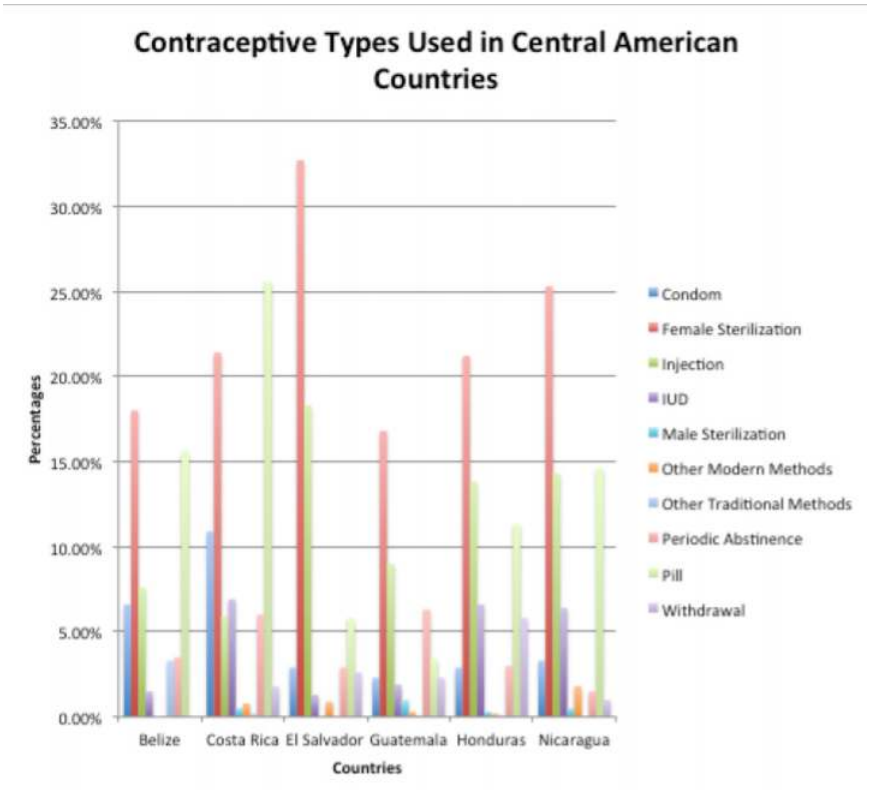
***Table 2.1.8**: Data of Defect Type*

6. People in Bangladesh were asked to state what type of birth control method they use. The percentages are given in *Table 2.1.9* ("Contraceptive use," 2013). Create a Pareto chart of the data and then state any findings you can from the graph.

Method	Percentage
Condom	4.50%
Pill	28.50%
Periodic Abstinence	4.90%
Injection	7.00%
Female Sterilization	5.00%
IUD	0.90%
Male Sterilization	0.70%
Withdrawal	2.90%
Other Modern Methods	0.70%
Other Traditional Methods	0.60%

***Table 2.1.9**: Data of Birth Control Type*

7. The percentages of people who use certain contraceptives in Central American countries are displayed in *Graph 2.1.6* ("Contraceptive use," 2013). State any findings you can from the graph.



***Graph 2.1.6**: Multiple Bar Chart for Contraceptive Types*

Answer

See solutions

This page titled 2.3: Other Types of Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the *stem-and-leaf graph* or *stemplot*, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.3.1.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Graph

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Exercise 2.3.1.2

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Answer

Stem	Leaf
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down

50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example 2.3.1.3

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

HINT: The leaves are to the right of the decimal.

Answer

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Exercise 2.3.1.4

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Answer

Stem	Leaf
0	5 7
1	1 2 2 3 3 5 5 7 7 8 9
2	0 2 5 6 8 8 8
3	5 8
4	4 8 9
5	2 5 7 8
6	
7	
8	0

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Example 2.3.1.5: Side-by-Side Stem-and-Leaf plot

A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Tables **2.3.1.1** and **2.3.1.2** show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Table **2.3.1.1**: Presidential Ages at Inauguration

President	Ageat Inauguration	President	Age	President	Age
Pierce	48	Harding	55	Obama	47
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Fillmore	50	Wilson	56	G. W. Bush	54
Tyler	51	McKinley	54	Reagan	69
Van Buren	54	B. Harrison	55	Ford	61
Washington	57	Lincoln	52	Hoover	54
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
J. Q. Adams	57	Arthur	51	L. Johnson	55
Monroe	58	Garfield	49	Kennedy	43
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jackson	61	Cleveland	47	Nixon	56
Taylor	64	Taft	51	Clinton	47
Buchanan	65	Coolidge	51	Trump	70
W. H. Harrison	68	Cleveland	55	Carter	52

2.3.1.2 *Presidential Age at Death*

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Answer

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 3 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

? Exercise 2.3.1.6

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Answer

Table **2.3.1.3**: Atlanta Hawks Wins and Losses

Number of Wins		Number of Losses
3	1	9
9 8 8 6 5	2	5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3	0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 7 6 6 6 3 3 3 2 2 1 1 0	4	0 0 1 1 2 4 5 6 6 7 7 8 9
7 7 6 3 2 0 0 0 0	5	1 1 1 2 3 4 4 6 7
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example 2.3.1.7

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table and in Figure.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Answer

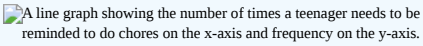


Figure 2.3.1.1: A line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis and frequency on the y-axis.

Exercise 2.3.1.8

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Answer

Figure 2.3.1.2: A line graph showing the number of times a car is in the shop on the x-axis and frequency on the y-axis.

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in Example 2.3.1.9 has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.3.1.9

By the end of 2011, Facebook had over 146 million users in the United States. Table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Answer

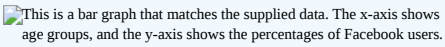


Figure **2.3.1.3**: This is a bar graph that matches the supplied data. The x-axis shows age groups and the y-axis show the percentages of Facebook users

Exercise 2.3.1.10

The population in Park City is made up of children, working-age adults, and retirees. Table shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Answer

Figure **2.3.1.4**: This is a bar graph that matches the supplied data. The x-axis shows age groups, and the y-axis shows the percentages of Park City's population.

Example 2.3.1.11

The columns in Table contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution

Figure **2.3.1.5**: This is a bar graph that matches the supplied data. The x-axis shows race and ethnicity, and the y-axis shows the percentages of AP examinees.

Exercise 2.3.1.12

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Answer

Figure **2.3.1.6**: This is a bar graph that matches the supplied data. The x-axis shows Park City voting districts, and the y-axis shows the percentages of the registered voter population.

Summary

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

References

1. Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at www.kenburbary.com/2011/03/fa...-statistics-2/ (accessed August 21, 2013).
2. “9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goa...omoting-equity (accessed September 13, 2013).
3. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

This page titled 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.3.2: Dot Plots

Learning Objectives

- Create and interpret dot plots
- Judge whether a dot plot would be appropriate for a given data set

Dot plots can be used to display various types of information. Figure **2.3.2.1** uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. Each dot represents a single M & M. From the figure, you can see that there were **3** blue M & M's, **19** brown M & M's, etc.

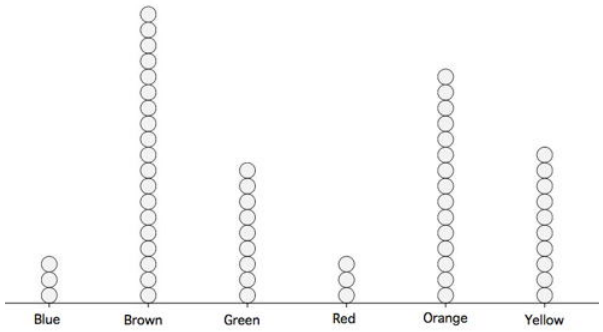


Figure **2.3.2.1**: A dot plot showing the number of M & M's of various colors in a bag of M & M's

The dot plot in Figure **2.3.2.2** shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure **2.3.2.1**, the location rather than the number of dots represents the frequency.

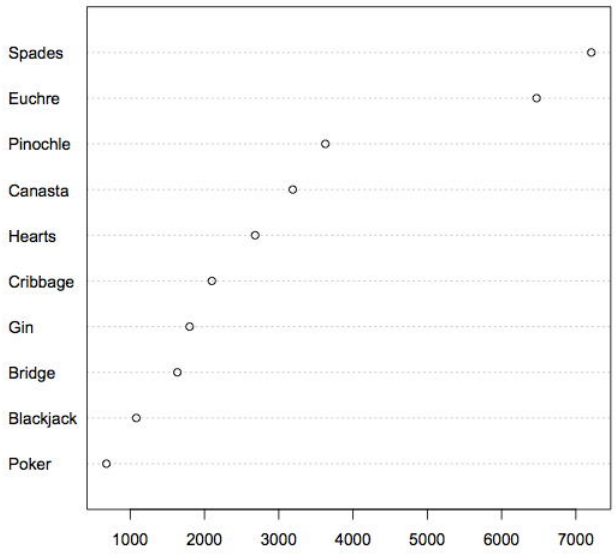


Figure **2.3.2.2**: A dot plot showing the number of people playing various card games on a Wednesday

The dot plot in Figure **2.3.2.3** shows the number of people playing on a Sunday and on a Wednesday. This graph makes it easy to compare the popularity of the games separately for the two days, but does not make it easy to compare the popularity of a given game on the two days.

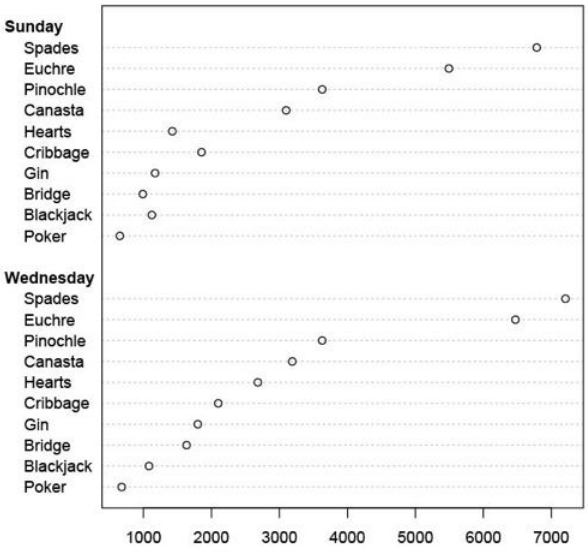


Figure **2.3.2.3**: A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday

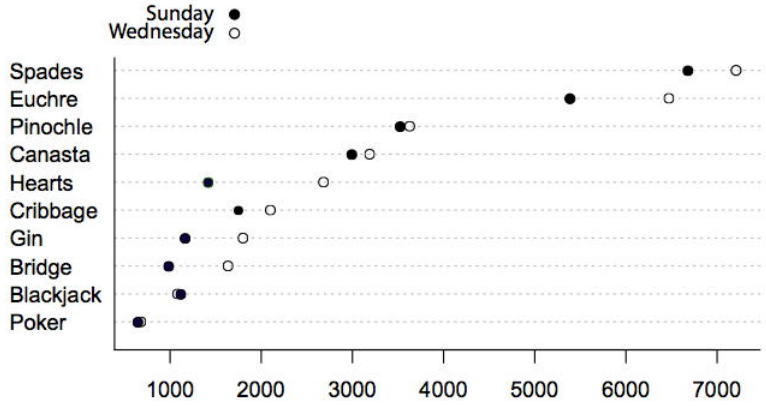


Figure **2.3.2.4**: An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday

The dot plot in Figure **2.3.2.4** makes it easy to compare the days of the week for specific games while still portraying differences among games.

This page titled 2.3.2: Dot Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **2.10: Dot Plots** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

2.3.3: Guide to Fairly Good Graphs

Learning Objectives

- It's not easy, but you can force spreadsheets to make publication-quality scientific graphs. This page explains how.

Introduction

Drawing graphs is an important part of analyzing your data and presenting the results of your research. Here I describe the features of clear, effective graphs, and I outline techniques for generating graphs using Excel. Most of these instructions also apply if you're using Calc, part of the free OpenOffice.org suite of programs, instead of Excel.

Many of the default conditions for Excel graphs are annoying, but with a little work, you can get it to produce graphs that are good enough for presentations and web pages. With a little more work, you can make publication-quality graphs. If you're drawing a lot of graphs, you may find it easier to use a specialized scientific graphing program.

General tips for all graphs

- Don't clutter up your graph with unnecessary junk. Grid lines, background patterns, $3-D$ effects, unnecessary legends, excessive tick marks, etc. all distract from the message of your graph.
- Do include all necessary information. Clearly label both axes of your graph, including measurement units if appropriate. You should identify symbols and patterns in a legend on the graph, or in the caption. If the graph has "error bars," you should say in the caption whether they're 95% confidence interval, standard error, standard deviation, comparison interval, or something else.
- Don't use color in graphs for publication. If your paper is a success, many people will be reading photocopies or will print it on a black-and-white printer. If the caption of a graph says "Red bars are mean HDL levels for patients taking 2000 mg niacin/day, while blue bars are patients taking the placebo," some of your readers will just see gray bars and will be confused and angry. For bars, use solid black, empty, gray, cross-hatching, vertical stripes, horizontal stripes, etc. Don't use different shades of gray, they may be hard to distinguish in photocopies. There are enough different symbols that you shouldn't need to use colors.
- Do use color in graphs for presentations. It's pretty, and it makes it easier to distinguish different categories of bars or symbols. But don't use red type on a blue background (or vice-versa), as the eye has a hard time focusing on both colors at once and it creates a distracting $3-D$ effect. And don't use both red and green bars or symbols on the same graph; from 5% to 10% of the men in your audience (and less than 1% of the women) have red-green colorblindness and can't distinguish red from green.

Choosing the right kind of graph

There are many kinds of graphs—bubble graphs, pie graphs, doughnut graphs, radar graphs—and each may be the best for some kinds of data. But by far the most common graphs in scientific publications are scatter graphs and bar graphs, so that's all that I'll talk about here.

Use a **scatter graph** (also known as an $X-Y$ graph) for graphing data sets consisting of pairs of numbers. These could be measurement variables, or they could be nominal variables summarized as percentages. Plot the independent variable on the X axis (the horizontal axis), and plot the dependent variable on the Y axis.

The independent variable is the one that you manipulate, and the dependent variable is the one that you observe. For example, you might manipulate salt content in the diet and observe the effect this has on blood pressure. Sometimes you don't really manipulate either variable, you observe them both. In that case, if you are testing the hypothesis that changes in one variable cause changes in the other, put the variable that you think causes the changes on the X axis. For example, you might plot "height, in cm" on the X axis and "number of head-bumps per week" on the Y axis if you are investigating whether being tall causes people to bump their heads more often. Finally, there are times when there is no cause-and-effect relationship, in which case you can plot either variable on the X axis; an example would be a graph showing the correlation between arm length and leg length.

There are a few situations where it is common to put the independent variable on the Y axis. For example, oceanographers often put "distance below the surface of the ocean" on the Y axis, with the top of the ocean at the top of the graph, and the dependent variable (such as chlorophyll concentration, salinity, fish abundance, etc.) on the X axis. Don't do this unless you're really sure that it's a strong tradition in your field.

Use a **bar graph** for plotting means or percentages for different values of a nominal variable, such as mean blood pressure for people on four different diets. Usually, the mean or percentage is on the *Y* axis, and the different values of the nominal variable are on the *X* axis, yielding vertical bars.

In general, I recommend using a bar graph when the variable on the *X* axis is nominal, and a scatter graph when the variable on the *X* axis is measurement. Sometimes it is not clear whether the variable on the *X* axis is a measurement or nominal variable, and thus whether the graph should be a scatter graph or a bar graph. This is most common with measurements taken at different times. In this case, I think a good rule is that if you could have had additional data points in between the values on your *X* axis, then you should use a scatter graph; if you couldn't have additional data points, a bar graph is appropriate. For example, if you sample the pollen content of the air on January 15, February 15, March 15, etc., you should use a scatter graph, with "day of the year" on the *X* axis. Each point represents the pollen content on a single day, and you could have sampled on other days; there could be points in between January 15 and February 15. However, if you sampled the pollen every day of the year and then calculated the mean pollen content for each month, you should plot a bar graph, with a separate bar for each month. This is because you have one mean for January, and one mean for February, and of course there are no months between January and February. This is just a recommendation on my part; if most people in your field plot this kind of data with a scatter graph, you probably should too.

Drawing scatter graphs with Excel

1. Put your independent variable in one column, with the dependent variable in the column to its right. You can have more than one dependent variable, each in its own column; each will be plotted with a different symbol.
2. If you are plotting 95% confidence intervals, standard errors, standard deviation, or some other kind of error bar, put the values in the next column. These should be intervals, not limits; thus if your first data point has an *X* value of 7 and a *Y* value of 4 ± 1.5, you'd have 7 in the first column, 4 in the second column, and 1.5 in the third column. For limits that are asymmetrical, such as the confidence limits on a binomial percentage, you'll need two columns, one for the difference between the percentage and the lower confidence limit, and one for the difference between the percentage and the upper confidence limit.

	A	B	C
1	Latitude	Species	CI
2	39.22	125.17	6.13
3	38.8	138.17	4.76
4	39.47	105.17	5.37
5	38.96	114.5	8.67
6	38.6	134.5	1.29
7	38.58	94.33	4.23
8	39.73	98.83	8.09

Fig. 7.2.1 An Excel spreadsheet set up for a scatter graph. Latitude is the X variable, Species is the Y variable, and CI is the confidence intervals.

3. Select the cells that have the data in them. Don't select the cells that contain the confidence intervals. In the above example, you'd select cells *A2* through *B8*.
4. From the Insert menu, choose "Chart". Choose "Scatter" (called "*X Y*" in some versions of Excel) as your chart type, then "Marked Scatter" (the one with just dots, not lines) as your chart subtype. Do *not* choose "Line"; the little picture may look like a scatter graph, but it isn't. And don't choose the other types of scatter graphs, even if you're going to put lines on your graph; you'll add the lines to your "Marked Scatter" graph later.
5. As you can see, the default graph looks horrible, so you need to fix it by formatting the various parts of the graph. Depending on which version of Excel you're using, you may need to click on the "Chart Layout" tab, or choose "Formatting Palette" from the View menu. If you don't see those, you can usually click once on the part of the graph you want to format, then choose it from the Format menu.
6. You can enter a "Chart Title", which you will need for a presentation graph. You probably won't want a title for a publication graph (since the graph will have a detailed caption there). Then enter titles for the *X* axis and *Y* axis, and be sure to include the units. By clicking on an axis title and then choosing "Axis Title..." from the Format menu, you can format the font and other aspects of the titles.

7. Use the "Legend" tab to get rid of the legend if you only have one set of Y values. If you have more than one set of Y values, get rid of the legend if you're going to explain the different symbols in the figure caption; leave the legend on if you think that's the most effective way to explain the symbols.
8. Click on the "Axes" tab, choose the Y axis, and choose "Axis options". Modify the "Scale" (the minimum and maximum values of the Y axis). The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y scale greater than **100%**. If you're going to be adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary over a fairly narrow range. A good rule of thumb (that I made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.
9. Also use the "Axes" tab to format the "Number" (the format for the numbers on the Y axis), "Ticks" (the position of tick marks, and whether you want "minor" tick marks in between the "major" ones). Use "Font" to set the font of the labels. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures, and you should use the same font for axis labels, titles, and any other text on your graph.
10. Format your X axis the same way you formatted your Y axis.
11. Use the "Gridlines" tab get rid of the gridlines; they're ugly and unnecessary.
12. If you want to add a regression line to your graph, click on one of the symbols, then choose "Add Trendline..." from the Chart menu. You will almost always want the linear trendline. Only add a regression line if it conveys useful information about your data; don't just automatically add one as decoration to all scatter graphs.
13. If you want to add error bars, ignore the "Error Bars" tab; instead, click on one of the symbols on the graph, and choose "Data Series" from the Format menu. Click on "Error Bars" on the left side, and then choose "Y Error Bars". Ignore "Error Bars with Standard Error" and "Error Bars with Standard Deviation", because they are **not** what they sound like; click on "Custom" instead. Click on the "Specify value" button and click on the little picture to the right of the "Positive Error Value". Then drag to select the range of cells that contains your positive error intervals. In the above example, you would select cells $C2$ to $C8$. Click on the picture next to the box, and use the same procedure to select the cells containing your negative error intervals (which will be the same range of cells as the positive intervals, unless your error bars are asymmetrical). If you want horizontal (X axis) error bars as well, repeat this procedure.

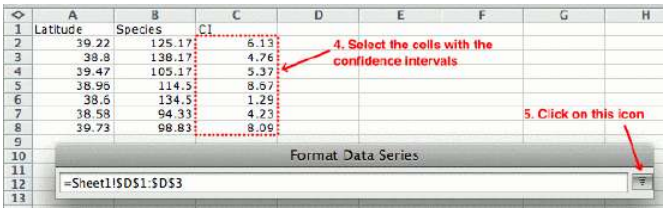


Fig. 7.2.2 Adding error bars to a graph. Repeat steps 3, 4 and 5 for the box labeled "-".

14. To format the symbols, click on one, and choose "Data Series" from the Format menu. Use "Marker Style" to set the shape of the markers, "Marker Line" to set the color and thickness of the line around the symbols, and "Marker Fill" to set the color that fills the marker. Repeat this for each set of symbols.
15. Click in the graph area, *inside* the graph, to select the whole graph. Choose "Plot Area" from the Format menu. Choose "Line" and set the color to black, to draw a black line on all four sides of the graph.
16. Click in the graph area, *outside* the graph, to select the whole box that includes the graph and the labels. Choose "Chart Area" from the Format menu. Choose "Line" and set the color to "No Line". On the "Properties" tab, choose "Don't move or size with cells," so the graph won't change size if you adjust the column widths of the spreadsheet.
17. You should now have a beautiful, beautiful graph. You can click once on the graph area (in the blank area outside the actual graph), copy it, and paste it into a word processing document, graphics program or presentation.

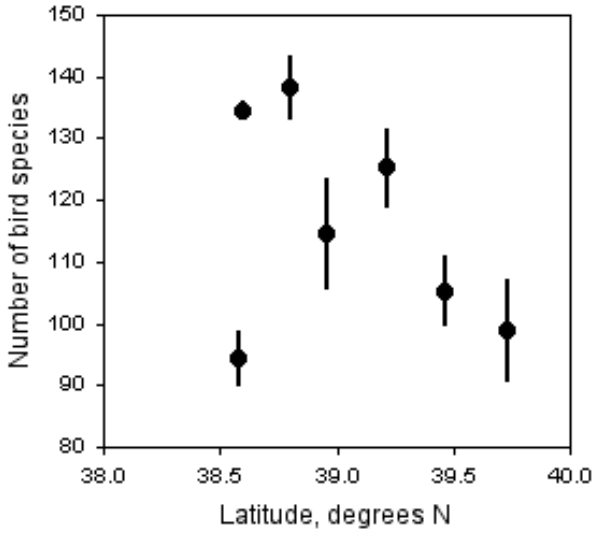


Fig. 7.2.3 The number of bird species observed in the Christmas Bird Count vs. latitude at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95% confidence intervals.

Drawing bar graphs with Excel

1. Put the values of the independent variable (the nominal variable) in one column, with the dependent variable in the column to its right. The first column will be used to label the bars or clusters of bars. You can have more than one dependent variable, each in its own column; each will be plotted with a different pattern of bar.
2. If you are plotting 95% confidence intervals or some other kind of error bar, put the values in the next column. These should be confidence intervals, not confidence limits; thus if your first row has a Y value of 4 ± 1.5, you'd have Control in the first column, 4 in the second column, and 1.5 in the third column. For confidence limits that are asymmetrical, such as the confidence intervals on a binomial percentage, you'll need two columns, one for the lower confidence interval, and one for the upper confidence interval.

	A	B	C
1	Location	Species	CI
2	Bombay Hook	125.17	6.13
3	Cape Henlopen	138.17	4.76
4	Middletown	105.17	5.37
5	Milford	114.5	8.67
6	Rehoboth	134.5	1.29
7	Seaford-Nanticoke	94.33	4.23
8	Wilmington	98.83	8.09

Fig. 7.2.4 An Excel spreadsheet set up for a bar graph including confidence intervals.

3. Select the cells that have the data in them. Include the first column, with the values of the nominal variable, but don't select cells that contain confidence intervals.
4. From the Insert menu, choose "Chart". Choose "Column" as your chart type, and then "Clustered Column" under "$2-D$ Column." Do not choose the three-dimensional bars, as they just add a bunch of clutter to your graph without conveying any additional information.
5. The default graph looks horrible, so you need to fix it by formatting the various parts of the graph. Depending on which version of Excel you're using, you may need to click on the "Chart Layout" tab, or choose "Formatting Palette" from the View menu. If you don't see those, you can usually click once on the part of the graph you want to format, then choose it from the Format menu.
6. You can enter a "Chart Title", which you will need for a presentation, but probably not for a publication (since the graph will have a detailed caption there). Then enter a title for the Y axis, including the units. You may or may not need an X axis title,

depending on how self-explanatory the column labels are. By clicking on "Axis title options..." you can format the font and other aspects of the titles.

7. Use the "Legend" tab to get rid of the legend if you only have one set of bars. If you have more than one set of bars, get rid of the legend if you're going to explain the different patterns in the figure caption; leave the legend on if you think that's the most effective way to explain the bar patterns.
8. Click on the "Axes" tab, choose the Y axis, and choose "Axis options". Modify the "Scale" (the minimum and maximum values of the Y axis). The maximum should be a nice round number, somewhat larger than the highest point on the graph. If you're plotting a binomial percentage, don't make the Y scale greater than 100%. If you're going to be adding error bars, the maximum Y should be high enough to include them. The minimum value on the Y scale should usually be zero, unless your observed values vary over a fairly narrow range. A good rule of thumb (that I made up, so don't take it too seriously) is that if your maximum observed Y is more than twice as large as your minimum observed Y, your Y scale should go down to zero. If you're plotting multiple graphs of similar data, they should all have the same scales for easier comparison.
9. Also use the "Axes" tab to format the "Number" (the format for the numbers on the Y axis), Ticks (the position of tick marks, and whether you want "minor" tick marks in between the "major" ones). Use "Font" to set the font of the labels. Most publications recommend sans-serif fonts (such as Arial, Geneva, or Helvetica) for figures, and you should use the same font for axis labels, titles, and any other text on your graph.
10. Format your X axis the same way you formatted your Y axis.
11. Use the "Gridlines" tab get rid of the gridlines; they're ugly and unnecessary.
12. If you want to add error bars, ignore the "Error Bars" tab; instead, click on one of the bars on the graph, and choose "Data Series" from the Format menu. Click on "Error Bars" on the left side. Ignore "Standard Error" and "Standard Deviation", because they are **not** what they sound like; click on "Custom" instead. Click on the "Specify value" button and click on the little picture to the right of the "Positive Error Value". Then drag to select the range of cells that contains your positive error intervals. In the above example, you would select cells $C2$ to $C8$. Click on the picture next to the box, and use the same procedure to select the cells containing your negative error intervals (which will be the same range of cells as the positive intervals, unless your error bars are asymmetrical).
13. To format the bars, click on one, and choose "Data Series" from the "Format" menu. Use "Line" to set the color and thickness of the lines around the bars, and "Fill" to set the color and pattern that fills the bars. Repeat this for each set of bars. Use "Options" to adjust the "Gap width," the space between sets of bars, and "Overlap" to adjust the space between bars within a set. Negative values for "Overlap" will produce a gap between bars within the same group.
14. Click in the graph area, inside the graph, to select the whole graph. Choose "Plot Area" from the Format menu. Choose "Line" and set the color to black, to draw a black line on all four sides of the graph.
15. Click in the graph area, outside the graph, to select the whole box that includes the graph and the labels. Choose "Chart Area" from the Format menu. Choose "Line" and set the color to "No Line". On the "Properties" tab, choose "Don't move or size with cells," so the graph won't change size if you adjust the column widths of the spreadsheet.
16. You should now have a beautiful, beautiful graph.

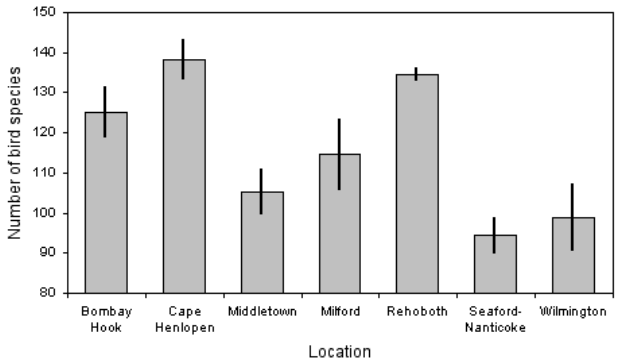


Fig. 7.2.5 The number of bird species observed in the Christmas Bird Count at seven locations in Delaware. Data points are the mean number of species for the counts in 2001 through 2006, with 95% confidence intervals.

Exporting Excel graphs to other formats

Once you've produced a graph, you'll probably want to export it to another program. You may want to put the graph in a presentation (Powerpoint, Keynote, Impress, etc.) or a word processing document. This is easy; click in the graph area to select the whole thing, copy it, then paste it into your presentation or word processing document. Sometimes, this will be good enough quality for your purposes.

Sometimes, you'll want to put the graph in a graphics program, so you can refine the graphics in ways that aren't possible in Excel, or so you can export the graph as a separate graphics file. This is particularly important for publications, where you need each figure to be a separate graphics file in the format and high resolution demanded by the publisher. To do this, right-click on the graph area (control-click on a Mac) somewhere **outside** the graph, then choose "Save as Picture". Change the format to PDF and you will create a pdf file containing just your graph. You can then open the pdf in a vector graphics program such as Adobe Illustrator or the free program Inkscape, ungroup the different elements of the graph, modify it, and export it in whatever format you need.

This page titled 2.3.3: Guide to Fairly Good Graphs is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: Guide to Fairly Good Graphs** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

2.3.4: Presenting Data in Tables

Learning Objectives

- Here are some tips for presenting scientific information in tables.

Graph or table

For a presentation, you should almost always use a graph, rather than a table, to present your data. It's easier to compare numbers to each other if they're represented by bars or symbols on a graph, rather than numbers. Here's data from the one-way anova page presented in both a graph and a table:

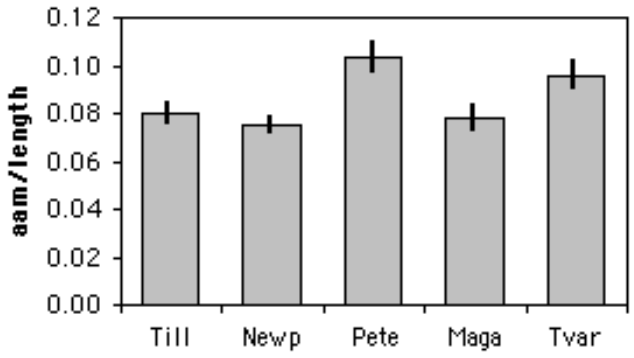


Fig. 7.3.1 Length of the anterior adductor muscle scar divided by total length in Mytilus trossulus. Means ±one standard error are shown for five locations.

Length of the anterior adductor muscle scar divided by total length in Mytilus trossulus. SE: standard error, N: sample size

Location	Mean AAM/ length	SE	N
Tillamook	0.080	0.0038	10
Newport	0.075	0.0030	8
Petersburg	0.103	0.0061	7
Magadan	0.078	0.0046	8
Tvarminne	0.096	0.0053	6

It's a lot easier to look at the graph and quickly see that the AAM/length ratio is highest at Petersburg and Tvarminne, while the other three locations are lower and about the same as each other. If you put this table in a presentation, you would have to point your laser frantically at one of the **15** numbers and say, "Here! Look at this number!" as your audience's attention slowly drifted away from your science and towards the refreshments table. "Would it be piggish to take a couple of cookies on the way out of the seminar, to eat later?" they'd be thinking. "Mmmmm, cookies...."

In a publication, the choice between a graph and a table is trickier. A graph is still easier to read and understand, but a table provides more detail. Most of your readers will probably be happy with a graph, but a few people who are deeply interested in your results may want more detail than you can show in a graph. If anyone is going to do a meta-analysis of your data, for example, they'll want means, sample sizes, and some measure of variation (standard error, standard deviation, or confidence limits). If you've done a bunch of statistical tests and someone wants to reanalyze your data using a correction for multiple comparisons, they'll need the exact ***P*** values, not just stars on a graph indicating significance. Someone who is planning a similar experiment to yours who is doing power analysis will need some measure of variation, as well.

Editors generally won't let you show a graph with the exact same information that you're also presenting in a table. What you can do for many journals, however, is put graphs in the main body of the paper, then put tables as supplemental material. Because these supplemental tables are online-only, you can put as much detail in them as you want; you could even have the individual measurements, not just means, if you thought it might be useful to someone.

Making a good table

Whatever word processor you're using probably has the ability to make good tables. Here are some tips:

- Each column should have a heading. It should include the units, if applicable.
- Don't separate columns with vertical lines. In the olden days of lead type, it was difficult for printers to make good-looking vertical lines; it would be easy now, but most journals still prohibit them.
- When you have a column of numbers, make sure the decimal points are aligned vertically with each other.
- Use a reasonable number of digits. For nominal variables summarized as proportions, use two digits for n less than 101, three digits for n from 101 to 1000, etc. This way, someone can use the proportion and the n and calculate your original numbers. For example, if n is 143 and you give the proportion as 0.22, it could be $31/143$ or $32/143$; reporting it as 0.217 lets anyone who's interested calculate that it was $31/143$. For measurement variables, you should usually report the mean using one more digit than the individual measurement has; for example, if you've measured hip extension to the nearest degree, report the mean to the nearest tenth of a degree. The standard error or other measure of variation should have two or three digits. P values are usually reported with two digits ($P = 0.44$, $P = 0.032$, $P = 2.7 \times 10^{-5}$, etc.).
- Don't use excessive numbers of horizontal lines. You'll want horizontal lines at the top and bottom of the table, and a line separating the heading from the main body, but that's probably about it. The exception is when you have multiple lines that should be grouped together. If the table of AAM/length ratios above had separate numbers for male and female mussels at each location, it might be acceptable to separate the locations with horizontal lines.
- Table formats sometimes don't translate well from one computer program to another; if you prepare a beautiful table using a Brand X word processor, then save it in Microsoft Word format or as a pdf to send to your collaborators or submit to a journal, it may not look so beautiful. So don't wait until the last minute; try out any format conversions you'll need, well before your deadline.

This page titled 2.3.4: Presenting Data in Tables is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **7.3: Presenting Data in Tables** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

2.4: Organizing Data - Frequency Distributions

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. In Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction. Use your instructor's guidance regarding whether to reduce fractions.

Categorical Frequency Distribution

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table lists the different data values in ascending order and their frequencies.

Table **2.4.1**: Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

A frequency is the number of times a value of the data occurs. According to Table Table **2.4.1**, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

Definition: Categorical Frequency Distribution

A categorical frequency distribution is a table to organize data that can be placed in specific categories, such as nominal- or ordinal-level data.

Definition: Relative frequencies

A *relative frequency* is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Table **2.4.2**: Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

The sum of the values in the relative frequency column of Table **2.4.2** is $\frac{20}{20}$, or 1.

Definition: Cumulative relative frequency

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table **2.4.3**.

Table **2.4.3**: Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Grouped Frequency Distribution

Definition: Grouped Frequency Distribution

A grouped frequency distribution is a table to organize data in which the data are grouped into classes with more than one unit in width. Used when the data is large, or it makes sense to group the data.

Table **2.4.4** represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table **2.4.4**: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

The data in this table have been **grouped** into the following intervals:

- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

This example is used again in Descriptive Statistics.

The next section will explain in detail how to create a grouped frequency distribution given a raw data set.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Exercise 2.4.1

a. From the Table, find the percentage of heights that are less than 65.95 inches.
b. Find the percentage of heights that fall between 61.95 and 65.95 inches.

c.
d.

e.
f.

Answer

a. If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.
b. Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

Exercise 2.4.2

Table **2.4.5** shows the amount, in inches, of annual rainfall in a sample of towns.

Table **2.4.5**:

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97–6.99	7	$\frac{7}{50} = 0.14$	$0.12 + 0.14 = 0.26$
6.99–9.01	15	$\frac{15}{50} = 0.30$	$0.26 + 0.30 = 0.56$
9.01–11.03	8	$\frac{8}{50} = 0.16$	$0.56 + 0.16 = 0.72$
11.03–13.05	9	$\frac{9}{50} = 0.18$	$0.72 + 0.18 = 0.90$
13.05–15.07	5	$\frac{5}{50} = 0.10$	$0.90 + 0.10 = 1.00$
	Total = 50	Total = 1.00	

a. Find the percentage of rainfall that is less than 9.01 inches.
b. Find the percentage of rainfall that is between 6.99 and 13.05 inches.

Answer

a. 0.56 or 56
b. $0.30 + 0.16 + 0.18 = 0.64$ or 64

Exercise 2.4.3

Use the heights of the 100 male semiprofessional soccer players in Table **2.4.4**. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Answer

a. 29%
b. 36%
c. 77%

d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Exercise 2.4.4

From Table **2.4.5**, find the number of towns that have rainfall between 2.95 and 9.01 inches.

Answer

$6 + 7 + 15 = 28$ towns

COLLABORATIVE EXERCISE 2.4.7

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

a. What percentage of the students in your class have no siblings?
b. What percentage of the students have from one to three siblings?
c. What percentage of the students have fewer than three siblings?

Example 2.4.7

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table was produced:

Table **2.4.6**: Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 2.4.8

Table represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Answer

$\frac{9}{50}$

Example 2.4.9

Table contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Table **2.4.7**:

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Answer

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

Exercise 2.4.10

Table contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Table **2.4.8**:

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Answer

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

1. “State & County QuickFacts,” U.S. Census Bureau. quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).
2. “State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).
3. “Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).
4. “Levels of Measurement,” infinity.cos.edu/faculty/wood...ata_Levels.htm (accessed May 1, 2013).

5. Courtney Taylor, “Levels of Measurement,” about.com, http://statistics.about.com/od/Helpa...easurement.htm (accessed May 1, 2013).
6. David Lane. “Levels of Measurement,” Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Glossary

Categorical Frequency Distribution
A table to organize data that can be placed in specific categories, such as nominal- or ordinal-level data.

Cumulative Relative Frequency
The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency
the number of times a value of the data occurs

Relative Frequency
the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 2.4: Organizing Data - Frequency Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.E: Graphs (Optional Exercises)

2.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Q 2.2.1

Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Q 2.2.2

Table contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

S 2.2.2

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.
   - Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
   - Press MATH
   - Arrow over to PRB
   - Press 5:randInt(

- Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

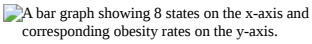


Figure 2.2.1 (a)

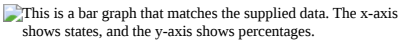


Figure 2.2.2 (b)

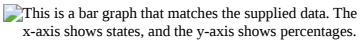


Figure 2.2.2 (c)

For each of the following data sets, create a stem plot and identify any outliers.

Exercise 2.2.7

The miles per gallon rating for 30 cars are shown below (lowest to highest).

19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Answer

Stem	Leaf
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Answer

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

The data are daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

For the next three exercises, use the data to construct a line graph.

Exercise 2.2.8

In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in Table.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Answer

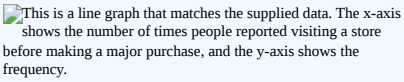


Figure 2.2.7

Exercise 2.2.9

In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in Table.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

Exercise 2.2.10

Several children were asked how many TV shows they watch each day. The results of the survey are shown in Table.

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Answer

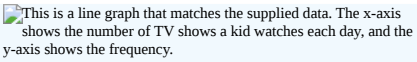


Figure 2.2.8

Exercise 2.2.11

The students in Ms. Ramirez's math class have birthdays in each of the four seasons. Table shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Using the data from Mrs. Ramirez's math class supplied in Exercise, construct a bar graph showing the percentages.

Answer

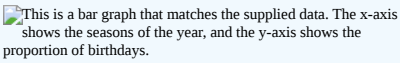


Figure 2.2.9

Exercise 2.2.12

David County has six high schools. Each school sent students to participate in a county-wide science competition. Table shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Use the data from the David County science competition supplied in Exercise. Construct a bar graph that shows the county-wide population percentage of students at each school.

Answer

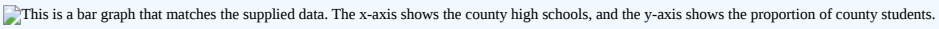


Figure 2.2.10

2.3: Histograms, Frequency, Polygons, and Time Series Graphs

Q 2.3.1

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the

previous month. The results are as follows:

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher C

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.

f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Q 2.3.2

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Singles

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Couples

Amount($)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
   i. List two similarities between the graphs.
   ii. List two differences between the graphs.
   iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
   i. List two similarities between the graphs.
   ii. Overall, are the graphs more similar or different?

g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

S 2.3.2

Singles

Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Couples

Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

a. See Table and Table.
b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

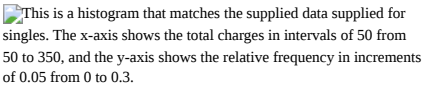

This is a histogram that matches the supplied data supplied for singles. The x-axis shows the total charges in intervals of 50 from 50 to 350, and the y-axis shows the relative frequency in increments of 0.05 from 0 to 0.3.

Figure 2.3.2.1.

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

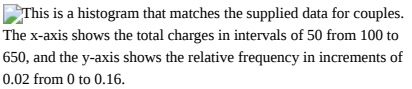

This is a histogram that matches the supplied data for couples. The x-axis shows the total charges in intervals of 50 from 100 to 650, and the y-axis shows the relative frequency in increments of 0.02 from 0 to 0.16.

Figure 2.3.2.2.

d. Compare the two graphs:
   i. Answers may vary. Possible answers include:

- Both graphs have a single peak.
- Both graphs use class intervals with width equal to $50.

ii. Answers may vary. Possible answers include:

- The couples graph has a class interval with no values.
- It takes almost twice as many class intervals to display the data for couples.

iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the Singles with the new graph for the Couples:

i.
- Both graphs have a single peak.
- Both graphs display 6 class intervals.
- Both graphs show the same general pattern.

ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Q 2.3.3

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.

b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.

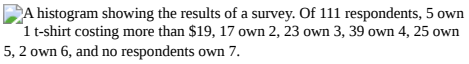


Figure 2.3.3.1.

Q 2.3.4

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21

b. 59

c. 41

d. Cannot be determined

S 2.3.4

c

Q 2.3.5

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

Q 2.3.6

Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.

S 2.3.7

Answers will vary.

Exercise 2.3.6

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Exercise 2.3.7

What does the frequency column in Table sum to? Why?

Answer

65

Exercise 2.3.8

What does the relative frequency column in Table sum to? Why?

Exercise 2.3.9

What is the difference between relative frequency and frequency for each data value in Table?

Answer

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

Exercise 2.3.10

What is the difference between cumulative relative frequency and relative frequency for each data value?

Exercise 2.3.11

To construct the histogram for the data in Table, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.



Figure 2.3.9.

Answer

Answers will vary. One possible histogram is shown:

Figure 2.3.10.

Exercise 2.3.12

Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Exercise 2.3.13

Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Answer

Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.

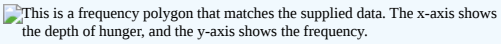


Figure 2.3.11.

Exercise 2.3.14

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Exercise 2.3.15

Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

Answer

Figure 2.3.12.

https://stats.libretexts.org/@go/page/11830

Exercise 2.3.16

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

This page titled 2.E: Graphs (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

3: Data Description

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "Descriptive Statistics." You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

3.1: Prelude to Descriptive Statistics
3.2: Measures of the Center of the Data
3.2.1: Skewness and the Mean, Median, and Mode
3.3: Measures of Variation
3.3.1: Coefficient of Variation
3.3.2: The Empirical Rule and Chebyshev's Theorem
3.4: Measures of Position
3.4.1: Measures of Location- Deciles
3.4.2: Z-scores
3.5: Exploratory Data Analysis
3.E: Descriptive Statistics (Optional Exercises)
1.E: Measures of Position (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 3: Data Description is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.1: Prelude to Descriptive Statistics

Skills to Develop

By the end of this chapter, the student should be able to:

- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.
- Recognize, describe, and calculate the measures of location of data: quartiles, percentiles, and deciles.
- Use quartiles to create a boxplot.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.



Figure 2.1.1: When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Measures of center, spread and location are additional ways you can describe data. We will explore important measures and another graph called a box plot.

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 3.1: Prelude to Descriptive Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.2: Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the median. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

> The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{3.2.1}$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \tag{3.2.2}$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression

$$\frac{n+1}{2} \tag{3.2.3}$$

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \tag{3.2.4}$$

The median is the 49th value in the ordered data. If the total number of data values is 100, then

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \tag{3.2.5}$$

The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 3.2.1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

Calculate the mean and the median.

Answer

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\ldots+35+37+40+(44)(2)+47]}{40} = 23.6 \quad (3.2.6)$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \quad (3.2.7)$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

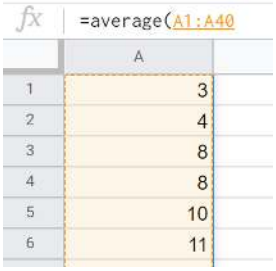
$$M = \frac{24+24}{2} = 24 \quad (3.2.8)$$

Technology

To find the mean and the median, you can use technology to assist you. Make sure you use the appropriate technology for your class.

Spreadsheets (Microsoft Excel/Google Sheets):

1. Enter each datum into its own cell. Usually we use one column for the data.
2. To find the mean, in the cell below the data, type: =average(
3. Select your data with your mouse. Make sure you have all your data selected, and no more. This should auto-populate the formula with the cell locations.
   For example, here is a small screenshot.



4. Hit enter. The spreadsheet should replace your formula with the mean of the data set.
5. To find the median, repeat the process, but use the formula: =median(

TI-83 or TI-84 Graphing Calculator:

1. Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.
2. Enter data into the list editor. Press STAT 1:EDIT.
3. Put the data values into list L1.
4. Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
5. Press the down and up arrow keys to scroll.

$\bar{x}$ = 23.6, M = 24

Exercise 3.2.1

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Answer

Mean: $3+4+5+7+7+7+7+8+8+9+9+10+10+10+10+10+11+12+12+13+14+14+15+15+17+17+18+19+19+19+21+21+22+22+23+24+24+24=544$

$$\frac{544}{39}=13.95 \tag{3.2.9}$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example 3.2.2

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution

$$\bar{x}=\frac{5,000,000+49(30,000)}{50}=129,400 \tag{3.2.10}$$

$M=30,000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Exercise 3.2.2

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Answer

The median is the better measure of the “center” than the mean because 59 of the values are $280,000 and one is $2,500,000. The $2,500,000 is an outlier. Either $280,000 or $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 3.2.3

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Answer

The most frequent score is 72, which occurs five times. Mode = 72.

Exercise 3.2.3

The number of books checked out from the library from 25 students are as follows:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 7; 8; 8; 8; 9; 10; 10; 11; 11; 12; 12

Find the mode.

Answer

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example 3.2.4

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

> The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Exercise 3.2.4

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Answer

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{3.2.11}$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary+upper boundary}}{2}. \tag{3.2.12}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{3.2.13}$$

where f is the frequency of the interval and m is the midpoint of the interval.

Example 3.2.5

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm 53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Exercise 3.2.5

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Answer

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.

$$Mean = (1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9) = 409.75 \quad (3.2.14)$$

References

1. Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).
2. “Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$$\mu = \frac{\sum fm}{\sum f} \tag{3.2.15}$$

where f = interval frequencies and m = interval midpoints.

Exercise 2.6.6

Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

Exercise 2.6.7

Calculate the mean.

Answer

Mean: $16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738$;

$\frac{738}{27} = 27.33$

Exercise 2.6.8

Identify the median.

Exercise 2.6.9

Identify the mode.

Answer

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise 2.6.10

sample mean = $\bar{x}$ = _____

Exercise 2.6.11

median = _____

Answer

4

Bringing It Together

Exercise 2.6.12

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

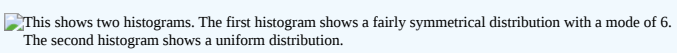


Figure **3.2.1**

1. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know? <figure >

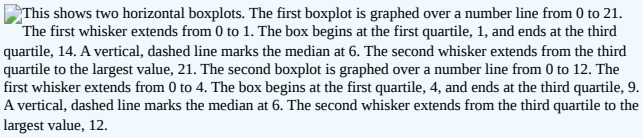


Figure **3.2.2**

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise 2.6.13

What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Answer

a

Exercise 2.6.14

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise 2.6.15

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Answer

b

Glossary

Frequency Table

a data representation in which grouped data is displayed along with the corresponding frequencies

Mean

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 3.2: Measures of the Center of the Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **2.6: Measures of the Center of the Data** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

3.2.1: Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

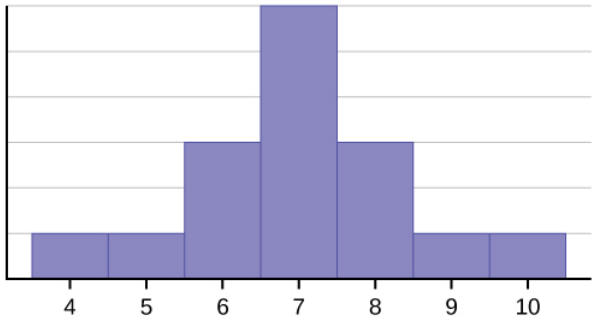


Figure **3.2.1.1**

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

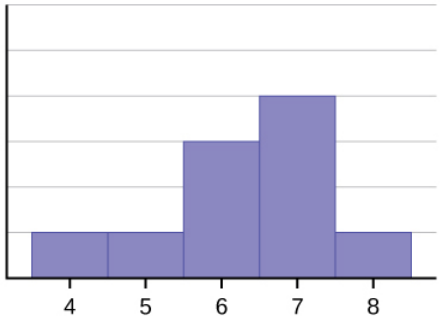


Figure **3.2.1.2**

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.

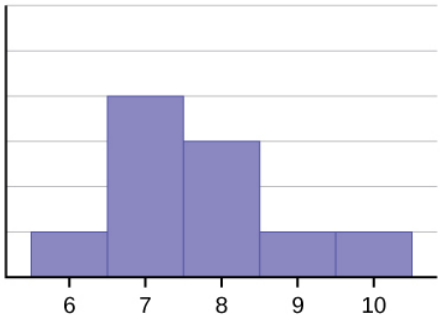


Figure **3.2.1.3**

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

> *Generally, if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.*

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example 3.2.1.1

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

- Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2
- Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1
- Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

Solution

a.

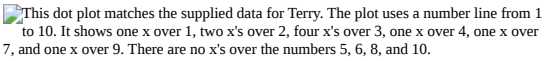


Figure **3.2.1.4**: This dot plot matches the supplied data for Terry. The plot uses a number line from 1 to 10. It shows one x over 1, two x's over 2, four x's over 3, one x over 4, one x over 7, and one x over 9. There are no x's over the numbers 5, 6, 8, and 10.

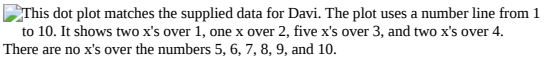


Figure **3.2.1.5**: Copy and Paste Caption here. (Copyright; author via source)

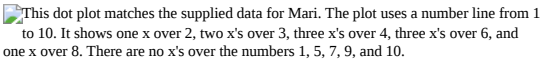


Figure **3.2.1.6**: Copy and Paste Caption here. (Copyright; author via source)

- Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
- Terry's median is three, Davis' median is three. Maris' median is four.
- It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Exercise 3.2.1.1

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

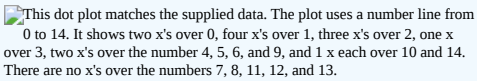


Figure **3.2.1.7**: This dot plot matches the supplied data. The plot uses a number line from 0 to 14. It shows two x's over 0, four x's over 1, three x's over 2, one x over 3, two x's over the number 4, 5, 6, and 9, and 1 x each over 10 and 14. There are no x's over the numbers 7, 8, 11, 12, and 13.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.

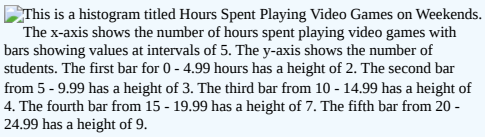


Figure **3.2.1.8**: This is a histogram titled Hours Spent Playing Video Games on Weekends. The x-axis shows the number of hours spent playing video games with bars showing values at intervals of 5. The y-axis shows the number of students. The first bar for 0 - 4.99 hours has a height of 2. The second bar from 5 - 9.99 has a height of 3. The third bar from 10 - 14.99 has a height of 4. The fourth bar from 15 - 19.99 has a height of 7. The fifth bar from 20 - 24.99 has a height of 9.

Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **left (or negative) skewed** distribution has a shape like Figure **3.2.1.2**. A **right (or positive) skewed** distribution has a shape like Figure **3.2.1.3**. A **symmetrical** distribution looks like Figure **3.2.1.1**.

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise 2.7.2

1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

Answer

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise 2.7.3

16; 17; 19; 22; 22; 22; 22; 22; 23

Exercise 2.7.4

87; 87; 87; 87; 87; 88; 89; 89; 90; 91

Answer

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise 2.7.5

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise 2.7.6

When the data are symmetrical, what is the typical relationship between the mean and median?

Answer

When the data are symmetrical, the mean and median are close or the same.

Exercise 2.7.7

What word describes a distribution that has two modes?

Exercise 2.7.8

Describe the shape of this distribution.

Figure **3.2.1.9**: This is a historgram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right.

Answer

The distribution is skewed right because it looks pulled out to the right.

Exercise 2.7.9

Describe the relationship between the mode and the median of this distribution.

Figure **3.2.1.10**: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heighs from left to right are: 8, 4, 2, 2, 1.

Exercise 2.7.10

Describe the relationship between the mean and the median of this distribution.

Figure **3.2.1.11**: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Answer

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise 2.7.11

Describe the shape of this distribution.

Figure **3.2.1.12**

Exercise 2.7.12

Describe the relationship between the mode and the median of this distribution.

Figure **3.2.1.13**

Answer

The mode and the median are the same. In this case, they are both five.

? Exercise 2.7.13

Are the mean and the median the exact same in this distribution? Why or why not?

Figure **3.2.1.14**

? Exercise 2.7.14

Describe the shape of this distribution.

Figure **3.2.1.15**

Answer

The distribution is skewed left because it looks pulled out to the left.

? Exercise 2.7.15

Describe the relationship between the mode and the median of this distribution.

Figure **3.2.1.16**: Copy and Paste Caption here. (Copyright; author via source)

? Exercise 2.7.16

Describe the relationship between the mean and the median of this distribution.

Figure **3.2.1.17**

Answer

The mean and the median are both six.

? Exercise 2.7.17

The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

? Exercise 2.7.18

Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

Answer

The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

? Exercise 2.7.19

Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

? Exercise 2.7.20

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Answer

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise 2.7.21

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

This page titled 3.2.1: Skewness and the Mean, Median, and Mode is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.3: Measures of Variation

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.

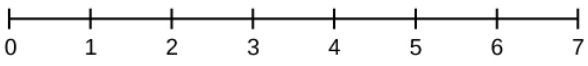


Figure **3.3.1**

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**

- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- sample:

Misplaced '#'

- Population:

Misplaced '#'

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\bar{x}$ is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{3.3.1}$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \tag{3.3.2}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{3.3.3}$$

or

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}} \tag{3.3.4}$$

For the population standard deviation, the denominator is N, the number of items in the population.

In Equations 3.3.2 and 3.3.4, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed previously in chapter 2. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the **sampling variability of a statistic** by its standard error.

The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in Chapter 7. The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Technology

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. If you are using a spreadsheet (Microsoft Excel or Google Sheets), you should use the appropriate formula =stdev.p(or =stdev.s(.We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The technology instructions appear at the end of this example.)

Example 3.3.1

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of n = 20 fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Freq.	Deviations	$Deviations^2$	(Freq.)($Deviations^2$)
x	f	$(x - \bar{x})$	$(x - \bar{x})2$	$(f)(x - \bar{x})2$
9	1	9 – 10.525 = –1.525	(–1.525)2 = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891$$

and this is rounded to two decimal places, $s = 0.72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: $x = \bar{x} + (\#ofSTDEVs)(s)$
- For a population: $x = \mu + (\#ofSTDEVs)\sigma$
- For this example, use $x = \bar{x} + (\#ofSTDEVs)(s)$ because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution: Spreadsheet (MS Excel/Google Sheets) (Part a only)

- Using raw data is easier for spreadsheets, because we can just use the standard deviation formulas =stdev.s(or =stdev.p(, depending on our data.
- This example can help us get ready for finding standard deviations of frequency distributions, so we'll emulate what was done above in the spreadsheet. Using the table above instead of the raw data, put the data values (9, 9.5, 10, 10.5, 11, 11.5) into the first column and the frequencies (1, 2, 4, 4, 6, 3) into the second column.
- We can take advantage of cell references to avoid typing repeated numbers and possibly making mistakes. We'll essentially copy the table above in the spreadsheet, but select the cells instead of typing them in. We can make the Spreadsheet do the calculations for us.
- For a number we don't want to change (the mean in this case), we can "lock" the cell reference using dollar signs around the letter. In this example, the mean is located in cell A9.

Formulas for use in Spreadsheets

Data (Column A)	Frequency (Column B)	Deviations (Column C)	Deviations^2 (Column D)	Freq*(Deviations)^2 (Column E)
9	1	=A2-A9	=C2^3	=B2*D2
9.5	2	=A3-A9	=C3^3	=B3*D3
10	4	=A4-A9	=C4^3	=B4*D4
10.5	4	=A5-A9	=C5^3	=B5*D5
11	6	=A6-A9	=C6^3	=B6*D6
11.5	3	=A7-A9	=C7^3	=B7*D7
	=sum(B2:B7)			=sum(E2:E7)

Then, just as above, divide the sum of Column E, 9.7375, by (20-1): 9.7375/19=0.5125.

Solution: TI Graphing Calculator

a.
- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x}$ = 10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$

c. $(\bar{x} - 2s) = 10.53–(2)(0.72) = 9.09$

d.
- $(\bar{x} - 1.5s) = 10.53–(1.5)(0.72) = 9.45$
- $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Exercise 2.8.1

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Answer

μ = 30.68

$s = 6.09$

$(\bar{x} + 2s = 30.68 + (2)(6.09) = 42.86.$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value

nine. **If you add the deviations, the sum is always zero**. (For Example 3.3.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one $(n - 1)$. Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

> Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example 3.3.2

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. *IQR*
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Answer

a. See Table
b. i. The sample mean = 73.5
   ii. The sample standard deviation = 17.9
   iii. The median = 73
   iv. The first quartile = 61
   v. The third quartile = 90
   vi. *IQR* = 90 – 61 = 29

c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5$ = the ending value; No data values fall on an interval boundary.

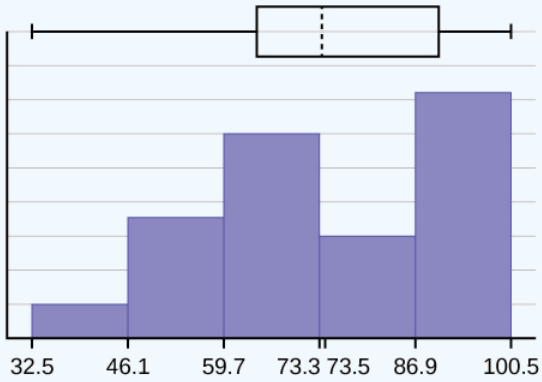


Figure 3.3.2.

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater $(73 - 33 = 40)$ than the spread in the upper 50% $(100 - 73 = 27)$. The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Exercise 3.3.2

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Answer

$\mu = 9.3$ and $s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{3.3.5}$$

where f interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example 3.3.3

Find the standard deviation for the data in Table 3.3.3.

Table 3.3.3

Class	Frequency, f	Midpoint, m	m^2	$\bar{x}$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\bar{x}$ = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\bar{x})^2}{n-1}}$ where s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Spreadsheets

For the previous example, we can use the spreadsheet to calculate the values in the table above, then plug the appropriate sums into the formula for sample standard deviation.

Graphing Calculator

Find the standard deviation for the data from the previous example

Class	0-2	3-5	6-8	9–11	12–14	15–17
Frequency, f	1	6	10	7	0	2

First, press the **STAT** key and select **1:Edit**



Figure 3.3.3

Input the midpoint values into **L1** and the frequencies into **L2**



Figure 3.3.4

Select **STAT**, **CALC**, and **1: 1-Var Stats**



Figure 3.3.5

Select 2nd then **1** then , 2nd then **2 Enter**



Figure **3.3.6**

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- Misplaced '#'
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example 3.3.4

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

Misplaced '#'

For John,

Misplaced '#'

For Ali,

Misplaced '#'

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Exercise 3.3.4

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Answer

For Angie:

$$z = \left(\frac{26.2 - 27.2}{0.8}\right) = -1.25$$

For Beth:

$$z = \left(\frac{27.3 - 30.1}{1.4}\right) = -2$$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

1. Data from Microsoft Bookshelf.
2. King, Bill.“Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$. .

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{3.3.6}$$

where s_x = sample standard deviation and $\bar{x}$ = sample mean

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Exercise 2.8.4

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Answer

s = 34.5

Exercise 2.8.5

Find the value that is one standard deviation below the mean.

Exercise 2.8.6

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Answer

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's *z*-score of –0.67 is higher than Karl's *z*-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise 2.8.7

Use Table to find the value that is three standard deviations:

- above the mean
- below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise 2.8.5

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Answer

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Bringing It Together

Exercise 2.8.7

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

a. Find the sample mean $\bar{x}$.
b. Find the approximate sample standard deviation, s.

Answer

a. 1.48
b. 1.12

Exercise 2.8.8

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let $X =$ the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

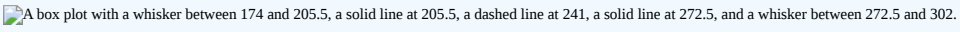
Exercise 2.8.9

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _____ to _____.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, σ.
   iii. the weight that is two standard deviations below the mean.
   iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Answer

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5
e. 

f. 205.5, 272.5
g. sample
h. population
i. i. 236.34
   ii. 37.50
   iii. 161.34
   iv. 0.84 std. dev. below the mean
j. Young

Exercise 2.8.10

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise 2.8.11

Refer to Figure determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

<figure >

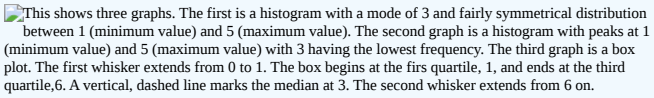


Figure **3.3.6**.

</figure>

a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

Answer

a. True
b. True
c. True
d. False

Exercise 2.8.12

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _____ days to _____ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise 2.8.13

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."

b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Answer

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. $X =$ the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise 2.8.14

The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise 2.8.15

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Answer

a

Exercise 2.8.16

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the Table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR*to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Contributors and Attributions

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 3.3: Measures of Variation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **2.8: Measures of the Spread of the Data** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

3.3.1: Coefficient of Variation

Coefficient of Variation

Coefficient of variation is the standard deviation divided by the mean; it summarizes the amount of variation as a percentage or proportion of the total. It is useful when comparing the amount of variation for one variable among groups with different means, or among different measurement variables. For example, the United States military measured foot length and foot width in 1774 American men. The standard deviation of foot length was $13.1mm$ and the standard deviation for foot width was $5.26mm$, which makes it seem as if foot length is more variable than foot width. However, feet are longer than they are wide. Dividing by the means ($269.7mm$ for length, $100.6mm$ for width), the coefficients of variation is actually slightly smaller for length (4.9%) than for width (5.2%), which for most purposes would be a more useful measure of variation.

Coefficient of Variation Formulas

The coefficient of variation, denoted by CVar or CV, is used to compare standard deviations from different populations.

For samples:

$$CV = \frac{s}{\bar{X}} \cdot 100 \tag{3.3.1.1}$$

For populations:

$$CV = \frac{\sigma}{\mu} \cdot 100 \tag{3.3.1.2}$$

Example

According to FuelEconomy.gov, for the year 2014, automatic Sport-Utility Vehicles with 4-wheel drive have an average fuel economy of 21 miles per gallon (mpg), with a standard deviation of 2.3 mpg. Standard trucks with 4-wheel drive and automatic transmission have an average fuel economy of 17 mpg and standard deviation of 2.0 mpg.

Compare the variations of the two.

Solution:

SUVs: 2.3/21*100% = 11.0%

Trucks: 2.0/17*100% = 11.8%

Comparing the coefficients of variation for the SUVs and the Trucks, the truck fuel economy is more variable than the SUVs.

Source: FuelEconomy.gov

This page titled 3.3.1: Coefficient of Variation is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

3.3.2: The Empirical Rule and Chebyshev's Theorem

Learning Objectives

- To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the *Empirical Rule* and *Chebyshev's Theorem*.
- To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

You probably have a good intuitive grasp of what the average of a data set says about that data set. In this section we begin to learn what the standard deviation has to tell us about the nature of the data set.

The Empirical Rule

We start by examining a specific set of data. Table **3.3.2.1** shows the heights in inches of **100** randomly selected adult men. A relative frequency histogram for the data is shown in Figure **3.3.2.1**. The mean and standard deviation of the data are, rounded to two decimal places, $\bar{x} = 69.92$ and $\sigma = 1.70$.

Table **3.3.2.1**: Heights of Men

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $69.92 - 1.70 = 68.22$ and $69.92 + 1.70 = 71.62$ inches, there are **69** of them. If we count the number of observations that are within two standard deviations of the mean, that is, that are between $69.92 - 2(1.70) = 66.52$ and $69.92 + 2(1.70) = 73.32$ inches, there are **95** of them. All of the measurements are within three standard deviations of the mean, that is, between $69.92 - 3(1.70) = 64.822$ and $69.92 + 3(1.70) = 75.02$ inches. These tallies are not coincidences, but are in agreement with the following result that has been found to be widely applicable.

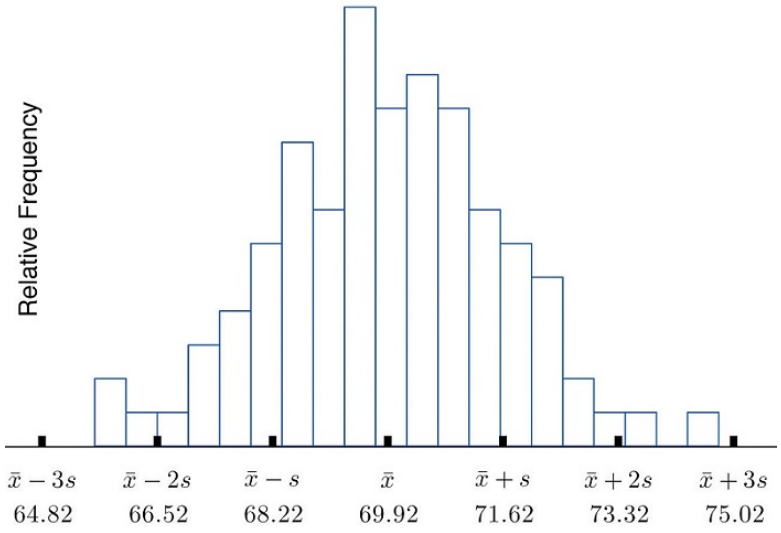


Figure 3.3.2.1: Heights of Adult Men

The Empirical Rule

Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations; if a data set has an approximately bell-shaped relative frequency histogram, then (Figure 3.3.2.2)

- approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations; and
- approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

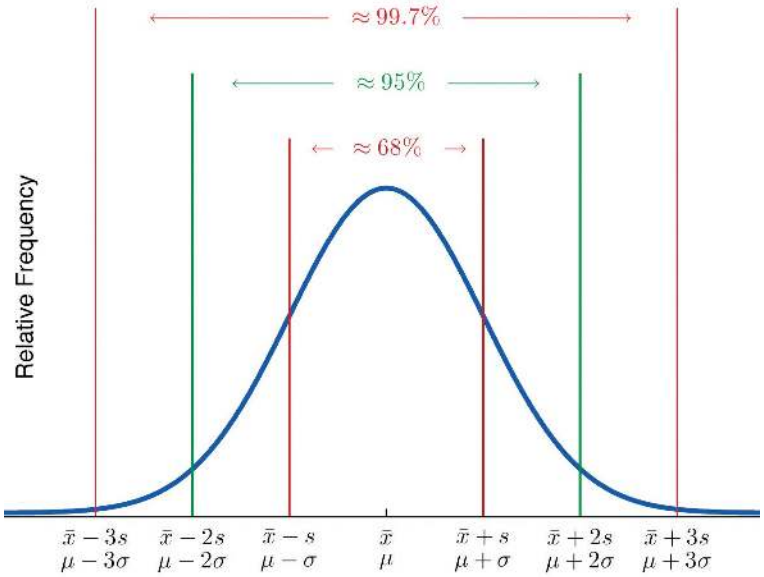


Figure 3.3.2.2: The Empirical Rule

Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule. We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches, but we counted 69.

Example 3.3.2.1

Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

1. About what proportion of all such men are between 68.2 and 71 inches tall?
2. What interval centered on the mean should contain about 95% of all such men?

Solution

A sketch of the distribution of heights is given in Figure 3.3.2.3.

1. Since the interval from 68.2 to 71.0 has endpoints $\bar{x} - s$ and $\bar{x} + s$, by the Empirical Rule about 68% of all 18-year-old males should have heights in this range.
2. By the Empirical Rule the shortest such interval has endpoints $\bar{x} - 2s$ and $\bar{x} + 2s$. Since

$$\bar{x} - 2s = 69.6 - 2(1.4) = 66.8$$

and

$$\bar{x} + 2s = 69.6 + 2(1.4) = 72.4$$

the interval in question is the interval from 66.8 inches to 72.4 inches.

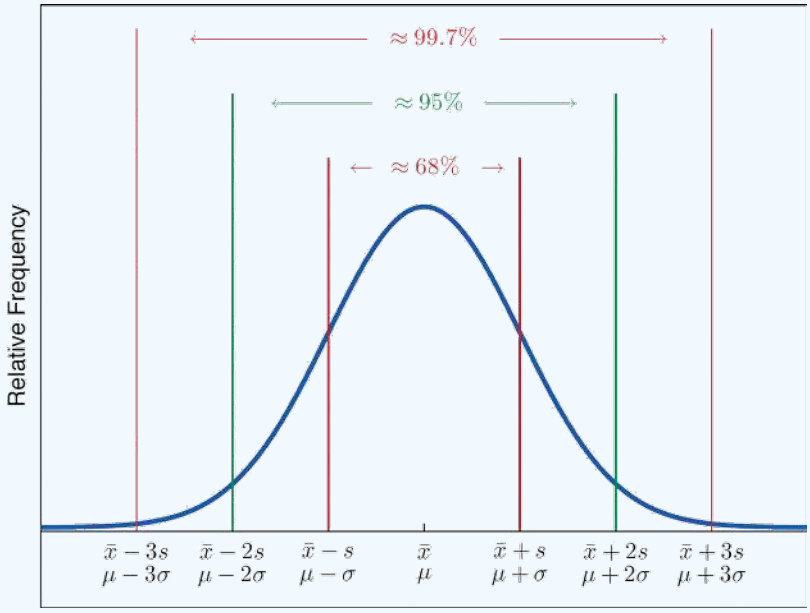


Figure 3.3.2.3: Distribution of Heights

Example 3.3.2.2

Scores on IQ tests have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution

A sketch of the IQ distribution is given in Figure 3.3.2.3. The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.

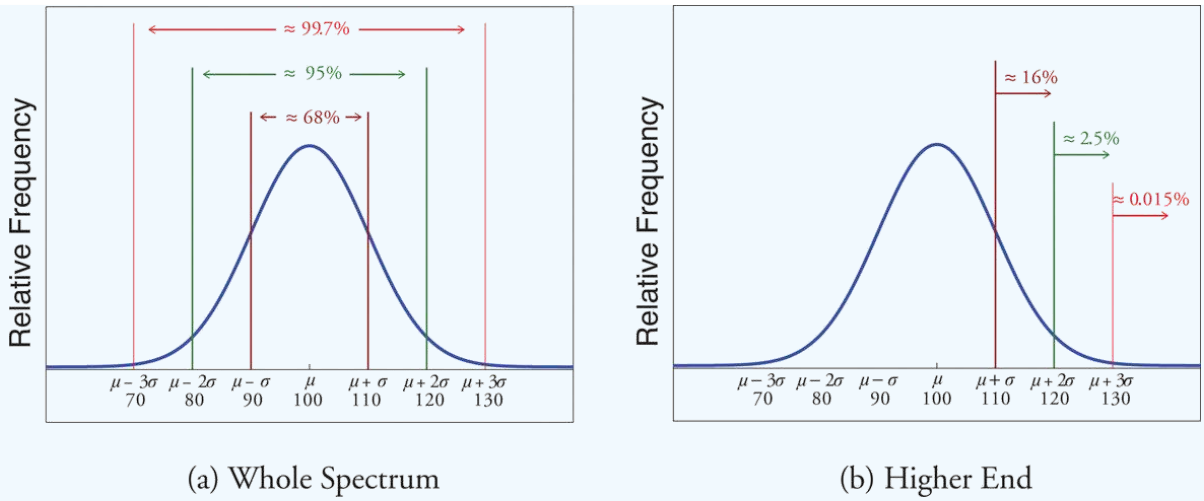


Figure **3.3.2.3**: Distribution of IQ Scores.

1. Since **68%** of the IQ scores lie *within* the interval from **90** to **110**, it must be the case that **32%** lie *outside* that interval. By symmetry approximately half of that **32%**, or **16%** of all IQ scores, will lie above **110**. If **16%** lie above **110**, then **84%** lie below. We conclude that the IQ score **110** is the 84th percentile.
2. The same analysis applies to the score **120**. Since approximately **95%** of all IQ scores lie within the interval form **80** to **120**, only **5%** lie outside it, and half of them, or **2.5%** of all scores, are above **120**. The IQ score **120** is thus higher than **97.5%** of all IQ scores, and is quite a high score.
3. By a similar argument, only **15/100** of **1%** of all adults, or about one or two in every thousand, would have an IQ score above **130**. This fact makes the score **130** extremely high.

Chebyshev's Theorem

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.

Chebyshev's Theorem

For any numerical data set,

- at least **3/4** of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations;
- at least **8/9** of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations;
- at least $1 - 1/k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than **1**.

Figure **3.3.2.4** gives a visual illustration of Chebyshev's Theorem.

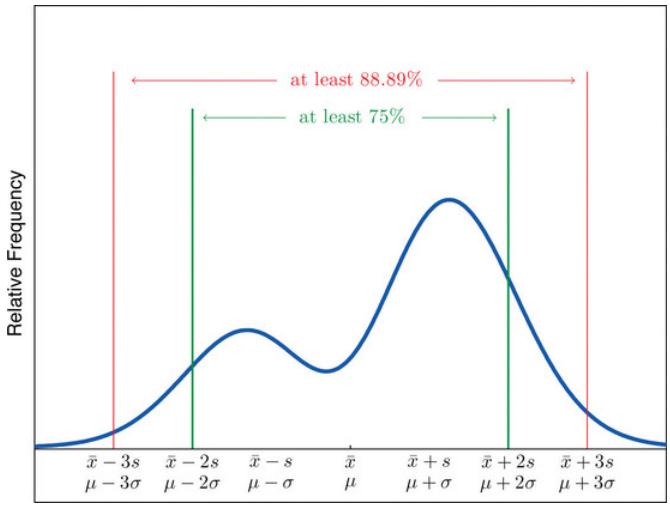


Figure **3.3.2.4**: Chebyshev's Theorem

It is important to pay careful attention to the words **"at least"** at the beginning of each of the three parts of Chebyshev's Theorem. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

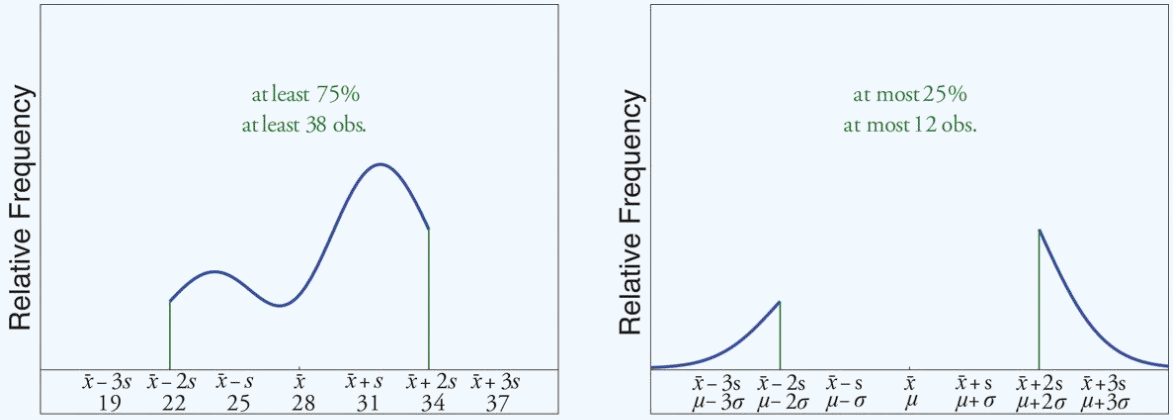
Example 3.3.2.3

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval $(22, 34)$? What can be said about the number of observations that lie outside that interval?

Solution

The interval $(22, 34)$ is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least $3/4$ of the data are within this interval. Since $3/4$ of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval $(22, 34)$.

If at least $3/4$ of the observations are in the interval, then at most $1/4$ of them are outside it. Since $1/4$ of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x $(22, 34)$.



(a) Within $\bar{x} \pm 2s$ (b) Outside $\bar{x} \pm 2$

Example 3.3.2.4

The number of vehicles passing through a busy intersection between $8:00\ a.m.$ and $10:00\ a.m.$ was observed and recorded on every weekday morning of the last year. The data set contains $n = 251$ numbers. The sample mean is $\bar{x} = 725$ and the sample standard deviation is $s = 25$. Identify which of the following statements *must* be true.

1. On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was either less than 675 or greater than 775.
5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.
6. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.

Solution

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.
2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $\bar{x} - 2s$, $\bar{x} + 2s = (675, 775)$. It must be correct.
3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct.
4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.
5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval $(675, 775)$ are less than 75. Thus statement (5) might not be correct.
6. Statement (4) is definitely correct and statement (4) implies statement (6): even if every measurement that is outside the interval $(675, 775)$ is less than 675 (which is conceivable, since symmetry is not known to hold), even so at most 25% of all observations are less than 675. Thus statement (6) must definitely be correct.

Key Takeaway

- The Empirical Rule is an approximation that applies only to data sets with a bell-shaped relative frequency histogram. It estimates the proportion of the measurements that lie within one, two, and three standard deviations of the mean.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

3.3.2: The Empirical Rule and Chebyshev's Theorem is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

- **2.5: The Empirical Rule and Chebyshev's Theorem** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

3.4: Measures of Position

The common measures of location are **quartiles** and **percentiles.** Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, M, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7 \tag{3.4.1}$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 – Q_1 \tag{2.4.1}$$

The *IQR* can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(*IQR*) below the first quartile or more than (1.5)(*IQR*) above the third quartile**. Potential outliers always require further investigation.

Definition: Outliers

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.4.1

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Answer

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Exercise 3.4.1

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

$33,000; $64,500; $28,000; $54,000; $72,000; $68,500; $69,000; $42,000; $54,000; $120,000; $40,500

Answer

Order the data from smallest to largest.

$28,000; $33,000; $40,500; $42,000; $54,000; $54,000; $64,500; $68,500; $69,000; $72,000; $120,000

Median = $54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

$$IQR = \$69,000 - \$40,500 = \$28,500$$

$$(1.5)(IQR) = (1.5)(\$28,500) = \$42,750$$

$$Q_1 - (1.5)(IQR) = \$40,500 - \$42,750 = -\$2,250$$

$$Q_3 + (1.5)(IQR) = \$69,000 + \$42,750 = \$111,750$$

No salary is less than –$2,250. However, $120,000 is more than $11,750, so $120,000 is a potential outlier.

Example 2.4.2

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Answer

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The *IQR* for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The *IQR* for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56–26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78–11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Exercise 3.4.2

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class *B*

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Answer

Class *A*

Order the data from smallest to largest.

65; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99

$Median = \frac{80 + 81}{2} = 80.5$

$Q_1 = \frac{69 + 76}{2} = 72.5$

$Q_3 = \frac{90 + 91}{2} = 90.5$

$IQR = 90.5 - 72.5 = 18$

Class B

Order the data from smallest to largest.

68; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 99; 100

$Median = \frac{80+80}{2} = 80$

$Q_1 = \frac{72+73}{2} = 72.5$

$Q_3 = \frac{92+95}{2} = 93.5$

$IQR = 93.5 - 72.5 = 21$

The data for Class B has a larger IQR, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class B are more spread out and not clustered about the median.

Example 3.4.3

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Exercise 3.4.3

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Answer

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example 2.4.4

Using the table above in Example **3.4.3**

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

Exercise 3.4.4

Refer to the table above in Exercise **3.4.3**. Find the third quartile. What is another name for the third quartile?

Answer

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

COLLABORATIVE STATISTICS

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

a. How many students were surveyed?
b. What kind of sampling did you do?
c. Construct two different histograms. For each, starting value = _____ ending value = ____.
d. Find the median, first quartile, and third quartile.
e. Construct a table of the data to find the following:
   i. the 10th percentile
   ii. the 70th percentile
   iii. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the kth percentile. Here is one of them.

- $k =$ the kth percentile. It may or may not be part of the data.
- $i =$ the index (ranking or position of a data value)
- $n =$ the total number of data

Order the data from smallest to largest.

Calculate $i = \frac{k}{100}(n+1)$

If i is an integer, then the kth percentile is the data value in the ith position in the ordered set of data.

If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.4.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution

a.
- $k = 70$
- $i =$ the index
- $n = 29$

$i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b.
- $k =$ 83rd percentile
- $i =$ the index
- $n = 29$

$i = \frac{k}{100}(n+1) = (\frac{83}{100})(29+1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Exercise 3.4.5

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Calculate the 20th percentile and the 55th percentile.

Answer

$k = 20$. Index $= i = \frac{k}{100}(n+1) = \frac{20}{100}(29+1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index $= i = \frac{k}{100}(n+1) = \frac{55}{100}(29+1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

Note 2.4.2

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- $x =$ the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- $y =$ the number of data values equal to the data value for which you want to find the percentile.
- $n =$ the total number of data.
- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.4.6

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

$x = 18$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{18 + 0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

$x = 3$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{3 + 0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12thpercentile.

Exercise 3.4.6

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Answer

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$x = 15$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(1)}{30}(100) = 51.67$. 47 is the 52nd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$x = 8$ and $y = 2$. $\frac{x + 0.5y}{n}(100) = \frac{8 + 0.5(2)}{30}(100) = 30$. 31 is the 30th percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Answer

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

? Exercise 3.4.7

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Answer

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Answer

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

? Exercise 3.4.8

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Answer

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example 2.4.9

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Answer

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Exercise 3.4.9

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Answer

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Example 2.4.10

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \tag{3.4.2}$$

.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

1. Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at usatoday30.usatoday.com/news/...sus/55029100/1 (accessed April 3, 2013).
2. Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).
3. “1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).
4. Data from *San Jose Mercury News*.
5. Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \frac{k}{100}(n+1)$$

where i = the ranking or position of a data value,

- k = the kth percentile,
- n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x+0.5y}{n}\right)(100)$

where $x =$ the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

$y =$ the number of data values equal to the data value for which you want to find the percentile,

$n =$ total number of data

Glossary

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

This page titled 3.4: Measures of Position is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **2.4: Measures of the Location of the Data** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

3.4.1: Measures of Location- Deciles

Deciles are another way we can consider location of data, where the data is separated into 10 groups. Just as quartiles correspond to specific percentiles, so do deciles. That is, the first decile is equivalent to the 10th percentile, the 5th decile is equivalent to the 2nd quartile and the 50th percentile.

If you are asked to find a specific decile, you are really looking for the corresponding percentile (multiply the decile by 10 to get the equivalent percentile). The 8th decile is equal to the 80th percentile.

You can also think of these measures of location like money: 2 quarters (2nd quartile) = 5 dimes (5th decile) = 50 pennies (50th percentile).

3.4.1: Measures of Location- Deciles is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.4.2: Z-scores

A z-score is a standardized version of a raw score (x) that gives information about the relative location of that score within its distribution. The formula for converting a raw score into a z-score is:

$$z = \frac{x - \mu}{\sigma} \tag{3.4.2.1}$$

for values from a population and for values from a sample:

$$z = \frac{x - \overline{X}}{s} \tag{3.4.2.2}$$

As you can see, z-scores combine information about where the distribution is located (the mean/center) with how wide the distribution is (the standard deviation/spread) to interpret a raw score (x). Specifically, z-scores will tell us how far the score is away from the mean in units of standard deviations and in what direction.

The value of a z-score has two parts: the sign (positive or negative) and the magnitude (the actual number). The sign of the z-score tells you in which half of the distribution the z-score falls: a positive sign (or no sign) indicates that the score is above the mean and on the right hand-side or upper end of the distribution, and a negative sign tells you the score is below the mean and on the left-hand side or lower end of the distribution. The magnitude of the number tells you, in units of standard deviations, how far away the score is from the center or mean. The magnitude can take on any value between negative and positive infinity, but for reasons we will see soon, they generally fall between -3 and 3.

Let's look at some examples. A z-score value of -1.0 tells us that this z-score is 1 standard deviation (because of the magnitude 1.0) below (because of the negative sign) the mean. Similarly, a z-score value of 1.0 tells us that this z-score is 1 standard deviation above the mean. Thus, these two scores are the same distance away from the mean but in opposite directions. A z-score of -2.5 is two-and-a-half standard deviations below the mean and is therefore farther from the center than both of the previous scores, and a z -score of 0.25 is closer than all of the ones before. In Unit 2, we will learn to formalize the distinction between what we consider "close" to the center or "far" from the center. For now, we will use a rough cut-off of 1.5 standard deviations in either direction as the difference between close scores (those within 1.5 standard deviations or between z = -1.5 and z = 1.5) and extreme scores (those farther than 1.5 standard deviations – below z = -1.5 or above z = 1.5).

We can also convert raw scores into z-scores to get a better idea of where in the distribution those scores fall. Let's say we get a score of 68 on an exam. We may be disappointed to have scored so low, but perhaps it was just a very hard exam. Having information about the distribution of all scores in the class would be helpful to put some perspective on ours. We find out that the class got an average score of 54 with a standard deviation of 8. To find out our relative location within this distribution, we simply convert our test score into a z-score.

$$z = \frac{X - \mu}{\sigma} = \frac{68 - 54}{8} = 1.75$$

We find that we are 1.75 standard deviations above the average, above our rough cut off for close and far. Suddenly our 68 is looking pretty good!

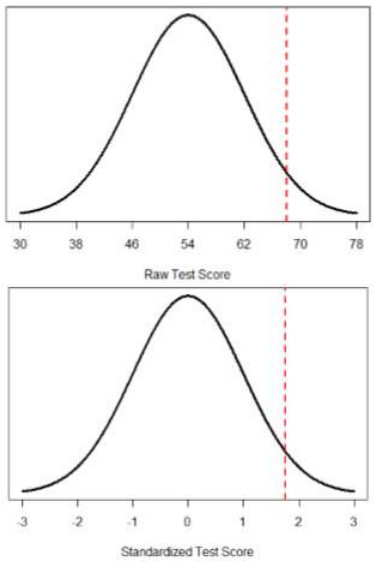


Figure **3.4.2.1**: Raw and standardized versions of a single score

Figure **3.4.2.1** shows both the raw score and the z-score on their respective distributions. Notice that the red line indicating where each score lies is in the same relative spot for both. This is because transforming a raw score into a z-score does not change its relative location, it only makes it easier to know precisely where it is.

Z-scores are also useful for comparing scores from different distributions. Let's say we take the SAT and score 501 on both the math and critical reading sections. Does that mean we did equally well on both? Scores on the math portion are distributed normally with a mean of 511 and standard deviation of 120, so our z-score on the math section is

$$z_{math} = \frac{501 - 511}{120} = -0.08$$

which is just slightly below average (note that use of "math" as a subscript; subscripts are used when presenting multiple versions of the same statistic in order to know which one is which and have no bearing on the actual calculation). The critical reading section has a mean of 495 and standard deviation of 116, so

$$z_{CR} = \frac{501 - 495}{116} = 0.05$$

So even though we were almost exactly average on both tests, we did a little bit better on the critical reading portion relative to other people.

Finally, z-scores are incredibly useful if we need to combine information from different measures that are on different scales. Let's say we give a set of employees a series of tests on things like job knowledge, personality, and leadership. We may want to combine these into a single score we can use to rate employees for development or promotion, but look what happens when we take the average of raw scores from different scales, as shown in Table **3.4.2.1**:

Table **3.4.2.1**: Raw test scores on different scales (ranges in parentheses).

Raw Scores	Job Knowledge (0 – 100)	Personality (1 –5)	Leadership (1 – 5)	Average
Employee 1	98	4.2	1.1	34.43
Employee 2	96	3.1	4.5	34.53
Employee 3	97	2.9	3.6	34.50

Because the job knowledge scores were so big and the scores were so similar, they overpowered the other scores and removed almost all variability in the average. However, if we standardize these scores into z-scores, our averages retain more variability and it is easier to assess differences between employees, as shown in Table **3.4.2.2**.

Table 3.4.2.2: Standardized scores.

z-Scores	Job Knowledge (0 – 100)	Personality (1 –5)	Leadership (1 – 5)	Average
Employee 1	1.00	1.14	-1.12	0.34
Employee 2	-1.00	-0.43	0.81	-0.20
Employee 3	0.00	-0.71	0.30	-0.14

Setting the scale of a distribution

Another convenient characteristic of z-scores is that they can be converted into any "scale" that we would like. Here, the term scale means how far apart the scores are (their spread) and where they are located (their central tendency). This can be very useful if we don't want to work with negative numbers or if we have a specific range we would like to present. The formulas for transforming z to x are:

$$x = z\sigma + \mu \tag{3.4.2.3}$$

for a population and

$$x = zs + \overline{X} \tag{3.4.2.4}$$

for a sample. Notice that these are just simple rearrangements of the original formulas for calculating z from raw scores.

Let's say we create a new measure of intelligence, and initial calibration finds that our scores have a mean of 40 and standard deviation of 7. Three people who have scores of 52, 43, and 34 want to know how well they did on the measure. We can convert their raw scores into z-scores:

$$z = \frac{52 - 40}{7} = 1.71$$
$$z = \frac{43 - 40}{7} = 0.43$$
$$z = \frac{34 - 40}{7} = -0.80$$

A problem is that these new z-scores aren't exactly intuitive for many people. We can give people information about their relative location in the distribution (for instance, the first person scored well above average), or we can translate these z scores into the more familiar metric of IQ scores, which have a mean of 100 and standard deviation of 16:

$$\text{IQ} = 1.71 * 16 + 100 = 127.36$$
$$\text{IQ} = 0.43 * 16 + 100 = 106.88$$
$$\text{IQ} = -0.80 * 16 + 100 = 87.20$$

We would also likely round these values to 127, 107, and 87, respectively, for convenience.

This page titled 3.4.2: Z-scores is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.2: Z-scores** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

3.5: Exploratory Data Analysis

Box plots (also called *box-and-whisker plots* or *box-whisker plots*) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately *the middle 50 percent of the data fall inside the box*. The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

> You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6;.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

> See the calculator instructions on the TI web site or in the appendix.



Figure **3.5.1**

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

> It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example 3.5.1

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70

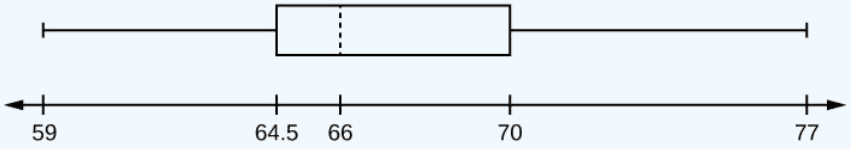


Figure **3.5.2**

a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. **Range = maximum value − the minimum value = 77 − 59 = 18**
d. Interquartile Range: $IQR = Q_3 – Q_1 = 70 − 64.5 = 5.5$.
e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Calculator

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4: Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE, and use the arrow keys to examine the box plot.

? Exercise 3.5.1

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

Answer

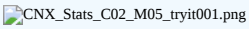


Figure 3.5.3

$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:

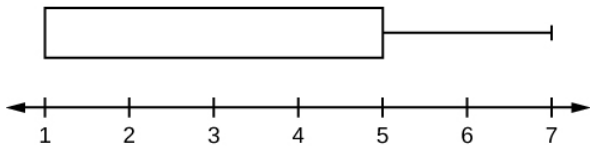


Figure 3.5.4

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example 3.5.2

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

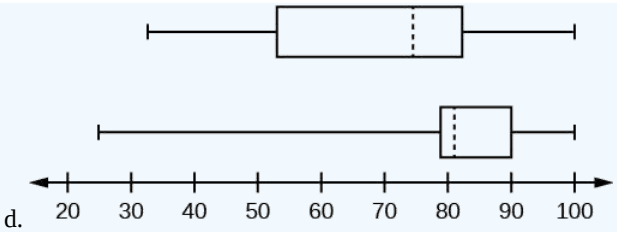
a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5
- Q_3 = 82.5
- Max = 99

b.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:



d.

e. The first data set has the wider spread for the middle 50% of the data. The *IQR*for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Exercise 3.5.2

The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Answer

CNX_Stats_C02_M05_tryit002.png

Figure 3.5.6

IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

Example 3.5.3

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.

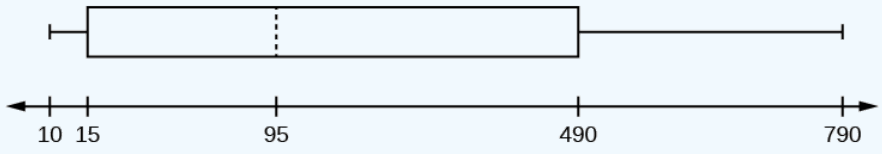


Figure 3.5.7

Exercise 3.5.3

Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

Answer

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330

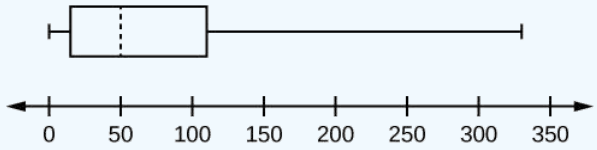


Figure 3.5.8

References

1. Data from *West Magazine*.

Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise 2.5.4

Construct a box plot below. Use a ruler to measure and scale accurately.

Exercise 2.5.5

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Answer

More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Bringing It Together

Exercise 2.5.6

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?

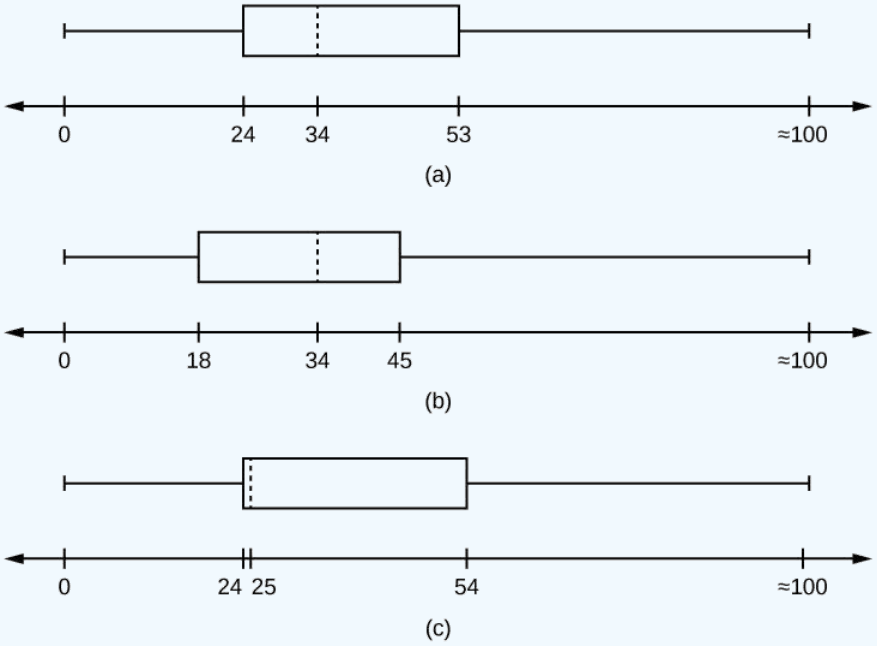


Figure **3.5.9**.

Answer

a. For graph, check student's solution.
b. 49.7% of the community is under the age of 35.
c. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

This page titled 3.5: Exploratory Data Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.E: Descriptive Statistics (Optional Exercises)

2.4: Measures of the Location of the Data

Q 2.4.1

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Q 2.4.2

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in Table. Also, include left endpoint, but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
   i. Should all bars have the same width, based on the data? Why or why not?
   ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40th and 80th percentiles
e. Construct a bar graph of the data

S 2.4.2

a. $1-(0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01)=0.06$
b. $0.19+0.26+0.18=0.63$
c. Check student's solution.
d. 40th percentile will fall between 30,000 and 40,000

   80th percentile will fall between 50,000 and 75,000

e. Check student's solution.

Q 2.4.3

Given the following box plot:

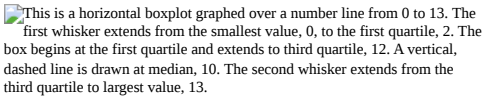


Figure 2.4.1.

a. which quarter has the smallest spread of data? What is that spread?
b. which quarter has the largest spread of data? What is that spread?
c. find the interquartile range (*IQR*).
d. are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
e. which interval has the fewest data in it? How do you know this?
   i. 0–2
   ii. 2–4
   iii. 10–12
   iv. 12–13
   v. need more information

Q 2.4.4

The following box plot shows the U.S. population for 1990, the latest available year.

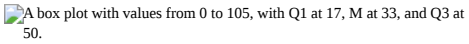


Figure 2.4.2.

a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

S 2.4.4

a. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
b. 62.4%

2.5: Box Plots

Q 2.5.1

In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.

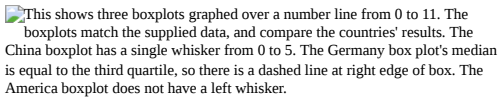


Figure 2.5.1.

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

Q 2.5.2

Given the following box plot, answer the questions.

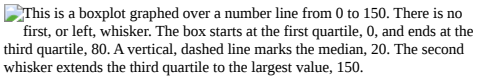


Figure 2.5.2.

a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

S 2.5.2

a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

Q 2.5.3

Given the following box plots, answer the questions.

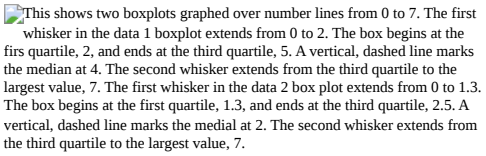


Figure 2.5.3.

a. In complete sentences, explain why each statement is false.
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.
b. For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences.

Q 2.5.4

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.

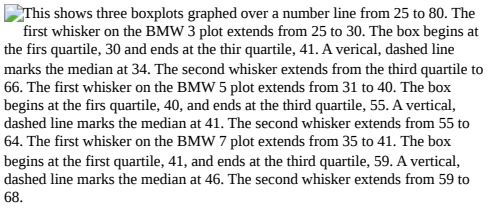


Figure 2.5.4.

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 series. Estimate the interquartile range (IQR).
g. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
   i. 31–35
   ii. 38–41
   iii. 41–64

S 2.5.4

a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.
b. The BMW 3 series is most likely to have an outlier. It has the longest whisker.

c. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
f. *IQR* ~ 17 years
g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

Q 2.5.5

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Construct a box plot of the data.

2.6: Measures of the Center of the Data

Q 2.6.1

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

Q 2.6.2

Table gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

S 2.6.2

The mean percentage, $\bar{x} = \frac{1328.65}{50} = 26.75$

2.7: Skewness and the Mean, Median, and Mode

Q 2.7.1

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.8: Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median = 1,014 FTES
- $\sigma = 474$ FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- $n = 29$ years

Q 2.8.1

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

S 2.8.1

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Q 2.8.2

75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

Q 2.8.3

The population standard deviation = _____

S 2.8.3

474 FTES

Q 2.8.4

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Q 2.8.5

What is the *IQR*? What does the *IQR* represent?

S 2.8.5

919

Q 2.8.6

How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Q 2.8.7

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the*IQR*. Round to one decimal place.

S 2.8.7

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Q 2.8.8

Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.

Q 2.8.9

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

S 2.8.10

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Q 2.8.11

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Q 2.8.12

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

S 2.8.12

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Q 2.8.13

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Q 2.8.14

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table 14.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

S 2.8.14

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that $23.32 + 12.95 = 36.27$ is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

Q 2.8.15

Table gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

This page titled 3.E: Descriptive Statistics (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.E: Measures of Position (Optional Exercises)

Exercise 1.E.10

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40th percentile.
b. Find the 78th percentile.

Answer

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

Exercise 1.E.11

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile of 37.
b. Find the percentile of 72.

Exercise 1.E.12

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Answer

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{143 + 0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercise 1.E.13

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 1.E.14

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Answer

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.

b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise 1.E.15

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise 1.E.16

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Answer

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise 1.E.17

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise 1.E.18

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Answer

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise 1.E.19

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise 1.E.20

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Answer

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Use Exercise to calculate the following values:

Exercise 1.E.21

First quartile = _____

Exercise 1.E.22

Second quartile = median = 50th percentile = _____

Answer

4

Exercise 1.E.23

Third quartile = _____

Exercise 1.E.24

Interquartile range (*IQR*) = _____ – _____ = _____

Answer

$6 - 4 = 2$

Exercise 1.E.25

10th percentile = _____

Exercise 1.E.26

70th percentile = _____

Answer

6

1.E: Measures of Position (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

4: Probability and Counting

Probability theory is concerned with probability, the analysis of random phenomena. The central objects of probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

4.1: Sample Spaces and Probability

4.1.1: Introduction to Probability

4.1.2: Terminology

4.2: Independent and Mutually Exclusive Events

4.3: The Addition and Multiplication Rules of Probability

4.3.1: Contingency Tables

4.3.2: Tree and Venn Diagrams

4.4: Counting Rules

4.4.1: Permutations

4.4.2: Permutations with Similar Elements

4.4.3: Combinations

4.5: Probability And Counting Rules

4.E: Probability Topics (Optional Exercises)

1.E: Combinations (Optional Exercises)

1.E: Permutations (Optional Exercises)

1.E: Permutations with Similar Elements (Optional Exercises)

1.E: Probability Using Tree Diagrams and Combinations (Optional Exercises)

1.E: Tree Diagrams and the Multiplication Axiom (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 4: Probability and Counting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Sample Spaces and Probability

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Glossary

You will learn more about probability in the following sections.

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 4.1: Sample Spaces and Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1.1: Introduction to Probability

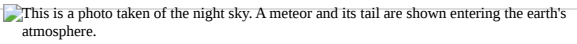


Figure **4.1.1.1**. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Collaborative Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

This page titled 4.1.1: Introduction to Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.1.2: Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is often used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where $H =$ heads and $T =$ tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

Definition: probability

The *probability* of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero and one and all numbers between these values). Also, if we add all the probabilities for each event in the sample space, the sum equals 1.

- $P(A) = 0$ means the event A can never happen.
- $P(A) = 1$ means the event A always happens.
- $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads)
- $\sum P(A) = 1$, where the Greek letter Σ represents "sum".

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where $T =$ tails and $H =$ heads. The sample space has four outcomes. $A =$ getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event $E =$ rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

Definition: law of large numbers

This important characteristic of probability experiments is known as the law of large numbers which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a

game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

The "OR" Event

An outcome is in the event **A OR B** if the outcome is in **A** or is in **B** or is in both **A** and **B**. For example, let $\mathbf{A} = \{1, 2, 3, 4, 5\}$ and $\mathbf{B} = \{4, 5, 6, 7, 8\}$. **A OR B** $= \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event **A AND B** if the outcome is in both **A** and **B** at the same time. For example, let **A** and **B** be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then **A AND B** $= 4, 5$.

The **complement** of event **A** is denoted **A'** (read "A prime"). **A'** consists of all outcomes that are **NOT** in **A**. Notice that $P(\mathrm{A}) + P(\mathrm{A}') = 1$. For example, let $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$ and let $\mathrm{A} = 1, 2, 3, 4$. Then, $\mathrm{A}' = 5, 6$. $P(A) = \frac{4}{6}$, $P(\mathrm{A}') = \frac{2}{6}$, and $P(\mathrm{A}) + P(\mathrm{A}') = \frac{4}{6} + \frac{2}{6} = 1$

The conditional probability of **A** given **B** is written $P(\mathrm{A}|\mathrm{B})$. $P(\mathrm{A}|\mathrm{B})$ is the probability that event **A** will occur given that the event **B** has already occurred. **A conditional reduces the sample space**. We calculate the probability of **A** from the reduced sample space **B**. The formula to calculate $P(\mathrm{A}|\mathrm{B})$ is $P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})}$ where $P(\mathrm{B})$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$. Let **A** = face is 2 or 3 and **B** = face is even (2, 4, 6). To calculate $P(\mathrm{A}|\mathrm{B})$, we count the number of outcomes 2 or 3 in the sample space $\mathrm{B} = \{2, 4, 6\}$. Then we divide that by the number of outcomes **B** (rather than **S**).

We get the same result by using the formula. Remember that **S** has six outcomes.

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} \tag{4.1.2.1}$$
$$= \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in S})}{6}}{\frac{(\text{the number of outcomes that are even in S})}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 4.1.2.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____

Let event $A =$ the even numbers and event $B =$ numbers greater than 13.

b. $A =$ _____, $B =$ _____
c. $P(\mathrm{A}) =$ _____, $P(\mathrm{B}) =$ _____
d. **A AND B** = _____, **A OR B** = _____
e. $P(\mathrm{A\ AND\ B}) =$ _____, $P(\mathrm{A\ OR\ B}) =$ _____
f. $\mathrm{A}' =$ _____, $P(\mathrm{A}') =$ _____

g. $P(\text{A}) + P(\text{A}') =$ _____

h. $P(\text{A}|\text{B}) =$ _____, $P(\text{B}|\text{A}) =$ _____; are the probabilities equal?

Answer

a. $\text{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$

b. $\text{A} = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, \text{B} = \{14, 15, 16, 17, 18, 19\}$

c. $P(\text{A}) = \frac{9}{19}$, $P(\text{B}) = \frac{6}{19}$

d. $\text{A AND B} = \{14, 16, 18\}$, $\text{A OR B} = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$

e. $P(\text{A AND B}) = \frac{3}{19}$, $P(\text{A OR B}) = \frac{12}{19}$

f. $\text{A}' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(\text{A}') = \frac{10}{19}$

g. $P(\text{A}) + P(\text{A}') = 1 \left(\left(\frac{9}{19} + \frac{10}{19} = 1\right)\right)$

h. $P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{\text{P(B)}} = \frac{3}{6}, P(\text{B}|\text{A}) = \frac{\text{P(A AND B)}}{\text{P(A)}} = \frac{3}{9}$, No

Exercise 4.1.2.1

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ _____

Let event $A =$ the sum is even and event $B =$ the first number is prime.

b. $A =$ _____, $B =$ _____

c. $P(\text{A}) =$ _____, $P(\text{B}) =$ _____

d. $\text{A AND B} =$ _____, $\text{A OR B} =$ _____

e. $P(\text{A AND B}) =$ _____, $P(\text{A OR B}) =$ _____

f. $\text{B}' =$ _____, $P(\text{B}') =$ _____

g. $P(\text{A}) + P(\text{A}') =$ _____

h. $P(\text{A}|\text{B}) =$ _____, $P(\text{B}|\text{A}) =$ _____; are the probabilities equal?

Answer

a. $\text{S} = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$

b. $\text{A} = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$

$\text{B} = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$

c. $P(\text{A}) = \frac{1}{2}$, $P(\text{B}) = \frac{2}{3}$

d. $\text{A AND B} = \{(2,2), (2,4), (3,1), (3,3)\}$

$\text{A OR B} = \{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$

e. $P(\text{A AND B}) = \frac{1}{3}, P(\text{A OR B}) = \frac{5}{6}$

f. $\text{B}' = \{(1,1), (1,2), (1,3), (1,4)\}, P(\text{B}') = \frac{1}{3}$

g. $P(\text{B}) + P(\text{B}') = 1$

h. $P(\text{A}|\text{B}) = \frac{P(\text{A AND B})}{P(\text{B})} = \frac{1}{2}, P(\text{B}|\text{A}) = \frac{P(\text{A AND B})}{P(\text{B})} = \frac{2}{3}$, No.

Example 4.1.2.2A

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event $\text{T} =$ the outcome is two.

b. Event $\text{A} =$ the outcome is an even number.

c. Event $\text{B} =$ the outcome is less than four.

d. The complement of A.

e. A GIVEN B

f. B GIVEN A

g. A AND B

h. A OR B

i. A OR B′
j. Event N = the outcome is a prime number.
k. Event I = the outcome is seven.

Solution

a. $\mathrm{T} = \{2\}$, $P(\mathrm{T}) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(\mathrm{A}) = \frac{1}{2}$
c. $\mathrm{B} = \{1, 2, 3\}$, $P(\mathrm{B}) = \frac{1}{2}$
d. $\mathrm{A}' = \{1, 3, 5\}, P(\mathrm{A}') = \frac{1}{2}$
e. $\mathrm{A|B} = \{2\}$, $P(\mathrm{A|B}) = \frac{1}{3}$
f. $\mathrm{B|A} = \{2\}$, $P(\mathrm{B|A}) = \frac{1}{3}$
g. A AND B $= 2, P(\mathrm{A\ AND\ B}) = \frac{1}{6}$
h. A OR B $= \{1, 2, 3, 4, 6\}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
i. A OR B′ $= \{2, 4, 5, 6\}$, $P(\mathrm{A\ OR\ B}') = \frac{2}{3}$
j. $\mathrm{N} = \{2, 3, 5\}$, $P(\mathrm{N}) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example 4.1.2.2B

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(\mathrm{M})$
b. $P(\mathrm{F})$
c. $P(\mathrm{R})$
d. $P(\mathrm{L})$
e. $P(\mathrm{M\ AND\ R})$
f. $P(\mathrm{F\ AND\ L})$
g. $P(\mathrm{M\ OR\ F})$
h. $P(\mathrm{M\ OR\ R})$
i. $P(\mathrm{F\ OR\ L})$
j. $P(\mathrm{M'})$
k. $P(\mathrm{R|M})$
l. $P(\mathrm{F|L})$
m. $P(\mathrm{L|F})$

Answer

a. $P(\mathrm{M}) = 0.52$
b. $P(\mathrm{F}) = 0.48$
c. $P(\mathrm{R}) = 0.87$
d. $P(\mathrm{L}) = 0.13$
e. $P(\mathrm{M\ AND\ R}) = 0.43$
f. $P(\mathrm{F\ AND\ L}) = 0.04$
g. $P(\mathrm{M\ OR\ F}) = 1$

h. $P(\text{M OR R}) = 0.96$
i. $P(\text{F OR L}) = 0.57$
j. $P(\text{M'}) = 0.48$
k. $P(\text{R|M}) = 0.8269$ (rounded to four decimal places)
l. $P(\text{F|L}) = 0.3077$ (rounded to four decimal places)
m. $P(\text{L|F}) = 0.0833$

References

1. "Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(\text{S}) = 1$ where S is the sample space

$0 \leq P(\text{A}) \leq 1$

$P(\text{A|B}) = \frac{\text{P(A AND B)}}{\text{P(B)}}$

Glossary

Conditional Probability
the likelihood that an event will occur given that another event has already occurred

Equally Likely
Each outcome of an experiment has the same probability.

Event
a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment
a planned activity carried out under controlled conditions

Outcome
a particular result of an experiment

Probability
a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

- $0 \leq P(\text{A}) \leq 1$
- If A and B are any two mutually exclusive events, then $\text{P(A OR B)} = P(\text{A}) + P(\text{B})$.
- $P(\text{S}) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event **A AND B** if the outcome is in both **A AND B** at the same time.

The Complement Event

The complement of event **A** consists of all outcomes that are NOT in **A**.

The Conditional Probability of *A* GIVEN *B*

$P(A|B)$ is the probability that event **A** will occur given that the event **B** has already occurred.

The Or Event

An outcome is in the event **A OR B** if the outcome is in **A** or is in **B** or is in both **A** and **B**.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

Exercise 3.2.2

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let **F** be the event that a student is female.
- Let **M** be the event that a student is male.
- Let **S** be the event that a student has short hair.
- Let **L** be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Answer

a. $P(\text{L}') = P(\text{S})$
b. $P(\text{M OR S})$
c. $P(\text{F AND L})$
d. $P(\text{M}|\text{L})$
e. $P(\text{L}|\text{M})$
f. $P(\text{S}|\text{F})$
g. $P(\text{F}|\text{L})$
h. $P(\text{F OR L})$
i. $P(\text{M AND S})$
j. $P(\text{F})$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let $H =$ the event of getting a hat.

Let $N =$ the event of getting a noisemaker.

Let $F =$ the event of getting a finger trap.

Let $C =$ the event of getting a bag of confetti.

Exercise 3.2.3

Find $P(\text{H})$.

Exercise 3.2.4

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise 3.2.5

Find $P(\text{F})$.

Exercise 3.2.6

Find $P(\text{C})$.

Answer

$P(\text{C}) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let $B =$ the event of getting a blue jelly bean

Let $G =$ the event of getting a green jelly bean.

Let $O =$ the event of getting an orange jelly bean.

Let $P =$ the event of getting a purple jelly bean.

Let $R =$ the event of getting a red jelly bean.

Let $Y =$ the event of getting a yellow jelly bean.

Exercise 3.2.7

Find $P(\text{B})$.

Exercise 3.2.8

Find $P(\text{G})$.

Answer

$P(\text{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise 3.2.9

Find $P(\mathrm{P})$.

Exercise 3.2.10

Find $P(\mathrm{R})$.

Answer

$P(\mathrm{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise 3.2.11

Find $P(\mathrm{Y})$.

Exercise 3.2.12

Find $P(\mathrm{O})$.

Answer

$P(textO) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let A = the event that a country is in Asia.

Let E = the event that a country is in Europe.

Let F = the event that a country is in Africa.

Let N = the event that a country is in North America.

Let O = the event that a country is in Oceania.

Let S = the event that a country is in South America.

Exercise 3.2.13

Find $P(\mathrm{A})$.

Exercise 3.2.14

Find $P(\mathrm{E})$.

Answer

$P(\mathrm{E}) = \frac{47}{194} = 0.24$

Exercise 3.2.15

Find $P(\mathrm{F})$.

Exercise 3.2.16

Find $P(\mathrm{N})$.

Answer

$P(\mathrm{N}) = \frac{23}{194} = 0.12$

Exercise 3.2.17

Find $P(O)$.

Exercise 3.2.18

Find $P(S)$.

Answer

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise 3.2.19

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise 3.2.20

What is the probability of drawing a club in a standard deck of 52 cards?

Answer

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise 3.2.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise 3.2.22

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Answer

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

Figure 3.2.1.

Let $B =$ the event of landing on blue.

Let $R =$ the event of landing on red.

Let $G =$ the event of landing on green.

Let $Y =$ the event of landing on yellow.

Exercise 3.2.23

If you land on Y, you get the biggest prize. Find $P(Y)$.

Exercise 3.2.24

If you land on red, you don't get a prize. What is $P(R)$?

Answer

$P(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let $I =$ the event that a player in an infielder.

Let $O =$ the event that a player is an outfielder.

Let $H =$ the event that a player is a great hitter.

Let $N =$ the event that a player is not a great hitter.

Exercise 3.2.25

Write the symbols for the probability that a player is not an outfielder.

Exercise 3.2.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(O \text{ OR } H)$

Exercise 3.2.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise 3.2.28

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Answer

$P(H|I)$

Exercise 3.2.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise 3.2.30

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Answer

$P(N|O)$

Exercise 3.2.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.2.32

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Answer

$P(I \text{ OR } N)$

Exercise 3.2.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.2.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\mathrm{I})$

Exercise 3.2.35

What is the word for the set of all possible outcomes?

Exercise 3.2.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.2.37

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let $\mathrm{F} =$ event that book is fiction

Let $\mathrm{N} =$ event that book is nonfiction

What is the sample space?

Exercise 3.2.38

What is the sum of the probabilities of an event and its complement?

Answer

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let $\mathrm{E} =$ the event that it lands on an even number. Let $\mathrm{M} =$ the event that it lands on a multiple of three.

Exercise 3.2.39

What does $P(\mathrm{E}|\mathrm{M})$ mean in words?

Exercise 3.2.40

What does $P(\mathrm{E}\ \mathrm{OR}\ \mathrm{M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

This page titled 4.1.2: Terminology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.2: Independent and Mutually Exclusive Events

Independent and *mutually exclusive* do not mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$
- $P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})$
- $P(\mathrm{A\ AND\ B}) = P(\mathrm{A})P(\mathrm{B})$

Two events **A** and **B** are independent if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show only one of the above conditions. If two events are NOT independent, then we say that they are dependent.

Sampling a population

Sampling may be done with replacement or without replacement (Figure **4.2.1**):

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be *independent*, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be *dependent* or *not independent*.

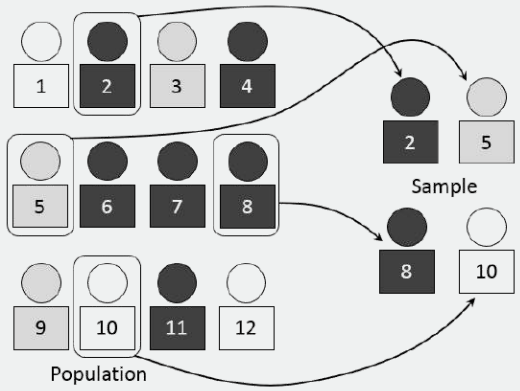


Figure **4.2.1**: A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". (CC BY-SA 4.0; Dan Kernler).

If it is not known whether **A** and **B** are independent or dependent, assume they are dependent until you can show otherwise.

Example **4.2.1**: Sampling with and without replacement

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, **J** (jack), **Q** (queen), **K** (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the **Q** of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the **Q** of spades again. Your picks are {**Q** of spades, ten of clubs, **Q** of spades}. You have picked the **Q** of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Exercise 4.2.1

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Answer a

With replacement

Answer b

No

Example 4.2.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are QS, 1D, 1C, QD.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are KH, 7D, 6D, KH.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Answer a

Without replacement

Answer b

With replacement

Exercise 4.2.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. QS, 1D, 1C, QD
b. KH, 7D, 6D, KH
c. QS, 7D, 6D, KS

Answer - without replacement

a. Possible; b. Impossible, c. Possible

Answer - with replacement

a. Possible; c. Possible, c. Possible

Mutually Exclusive Events

A and B are mutually exclusive events if they **cannot** occur at the same time. This means that A and B do not share any outcomes and $P(\text{A AND B}) = 0$.

For example, suppose the sample space

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Let $\text{A} = \{1, 2, 3, 4, 5\}, \text{B} = \{4, 5, 6, 7, 8\}$, and $\text{C} = \{7, 9\}$. A AND B $= \{4, 5\}$.

$$P(\text{A AND B}) = \frac{2}{10}$$

and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(\text{A AND C}) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example 4.2.3

Flip two fair coins.

The sample space is $\{HH, HT, TH, TT\}$ where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as $\{HH, HT, TH\}$. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A, so $\text{B} = \text{A}'$. Also, $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$.
- The probabilities for A and for B are $P(\text{A}) = \frac{3}{4}$ and $P(\text{B}) = \frac{1}{4}$.
- Let C = the event of getting all heads. $\text{C} = \{HH\}$. Since $\text{B} = \{TT\}$, $P(\text{B AND C}) = 0$. B and Care mutually exclusive. B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $\text{D} = \{TT\}$. $P(\text{D}) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $\text{E} = \{HT, HH\}$. $P(\text{E}) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. $\text{F} = \{HT, TH, TT\}$. $P(\text{F}) = \frac{3}{4}$

Exercise 4.2.3

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event A = Getting at least one black card $= \{BB, BR, RB\}$

$P(\text{A}) = \frac{3}{4} = 0.75$

Example 4.2.4

Flip two fair coins. Find the probabilities of the events.

a. Let $\mathbf{F}$ = the event of getting at most one tail (zero or one tail).
b. Let $\mathbf{G}$ = the event of getting two faces that are the same.
c. Let $\mathbf{H}$ = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are $\mathbf{F}$ and $\mathbf{G}$ mutually exclusive?
e. Let $\mathbf{J}$ = the event of getting all tails. Are $\mathbf{J}$ and $\mathbf{H}$ mutually exclusive?

Solution

Look at the sample space in Example 4.2.3.

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(\mathbf{F}) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(\mathbf{G}) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(\mathbf{H}) = \frac{2}{4}$
4. $\mathbf{F}$ and $\mathbf{G}$ share HH so $P(\mathbf{F}\text{ AND }\mathbf{G})$ is not equal to zero (0). $\mathbf{F}$ and $\mathbf{G}$ are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (TT). $\mathbf{H}$'s outcomes are HH and HT.

$\mathbf{J}$ and $\mathbf{H}$ have nothing in common so $P(\mathbf{J}\text{ AND }\mathbf{H}) = 0$. $\mathbf{J}$ and $\mathbf{H}$ are mutually exclusive.

Exercise 4.2.4

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let $\mathbf{F}$ = the event of getting the white ball twice.
b. Let $\mathbf{G}$ = the event of getting two balls of different colors.
c. Let $\mathbf{H}$ = the event of getting white on the first pick.
d. Are $\mathbf{F}$ and $\mathbf{G}$ mutually exclusive?
e. Are $\mathbf{G}$ and $\mathbf{H}$ mutually exclusive?

Answer

a. $P(\mathbf{F}) = \frac{1}{4}$
b. $P(\mathbf{G}) = \frac{1}{2}$
c. $P(\mathbf{H}) = \frac{1}{2}$
d. Yes
e. No

Example 4.2.5

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event $\mathbf{A}$ = a face is odd. Then $\mathbf{A} = \{1, 3, 5\}$. Let event $\mathbf{B}$ = a face is even. Then $\mathbf{B} = \{2, 4, 6\}$.

- Find the complement of $\mathbf{A}$, $\mathbf{A}'$. The complement of $\mathbf{A}$, $\mathbf{A}'$, is $\mathbf{B}$ because $\mathbf{A}$ and $\mathbf{B}$ together make up the sample space. $P(\mathbf{A}) + P(\mathbf{B}) = P(\mathbf{A}) + P(\mathbf{A}') = 1$. Also, $P(\mathbf{A}) = \frac{3}{6}$ and $P(\mathbf{B}) = \frac{3}{6}$.
- Let event $\mathbf{C}$ = odd faces larger than two. Then $\mathbf{C} = \{3, 5\}$. Let event $\mathbf{D}$ = all even faces smaller than five. Then $\mathbf{D} = \{2, 4\}$. $P(\mathbf{C}\text{ AND }\mathbf{D}) = 0$ because you cannot have an odd and even face at the same time. Therefore, $\mathbf{C}$ and $\mathbf{D}$ are mutually exclusive events.
- Let event $\mathbf{E}$ = all faces less than five. $\mathbf{E} = \{1, 2, 3, 4\}$.

Are $\mathbf{C}$ and $\mathbf{E}$ mutually exclusive events? (Answer yes or no.) Why or why not?

Answer

No. $C = \{3, 5\}$ and $E = \{1, 2, 3, 4\}$. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

Exercise 4.2.5

Let event A = learning Spanish. Let event B = learning German. Then $A \text{ AND } B$ = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A)$
- $P(A \text{ AND } B) = P(A)P(B)$

Answer

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A) \quad (4.2.1)$$

The events are independent because $P(A|B) = P(A)$.

Example 4.2.6

Let event G = taking a math class. Let event H = taking a science class. Then, $G \text{ AND } H$ = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. a. Show that $P(G|H) = P(G)$.
b. b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution

a. $P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$
b. $P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Exercise 4.2.6

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(\text{G AND O})$?

Answer

Event G and O $= \{G1, G3\}$

$P(\text{G and O}) = \frac{2}{10} = 0.2$

Example 4.2.7

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(\text{C}) = 0.75$, $P(\text{D}) = 0.3$, $P(\text{C|D}) = 0.75$ and $P(\text{C AND D}) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(\text{D|C})$?

Solution

a. Yes, because $P(\text{C|D}) = P(\text{C})$.
b. No, because $P(\text{C AND D})$ is not equal to zero.
c. $P(\text{D|C}) = \frac{P(\text{C AND D})}{P(\text{C})} = \frac{0.225}{0.75} = 0.3$

Exercise 4.2.7

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{B AND D}) = 0.20$.

a. Find $P(\text{B|D})$.
b. Find $P(\text{D|B})$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Answer

a. $P(\text{B|D}) = 0.6667$
b. $P(\text{D|B}) = 0.5$
c. No
d. No

Example 4.2.8

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let

- R = red card is drawn,
- B = blue card is drawn,
- E = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$.

S has eight outcomes.

- $P(\text{R}) = \frac{3}{8}$. $P(\text{B}) = \frac{5}{8}$. $P(\text{R AND B}) = 0$. (You cannot draw one card that is both red and blue.)

- $P(\text{E}) = \frac{3}{8}$. (There are three even-numbered cards, $R2, B2$, and $B4$.)
- $P(\text{E}|\text{B}) = \frac{2}{5}$. (There are five blue cards: $B1, B2, B3, B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(\text{B}|\text{E}) = \frac{2}{3}$. (There are three even-numbered cards: $R2, B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(\text{R AND B}) = 0$.
- Let G = card with a number greater than 3. $\text{G} = \{B4, B5\}$. $P(\text{G}) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $\text{H} = \{B1, B2, B3, B4\}$. $P(\text{G}|\text{H}) = \frac{1}{4}$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(\text{G}) = P(\text{G}|\text{H})$, which means that G and H are independent.

? Exercise 4.2.8

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Answer

- $P(\text{B}|\text{A}) = 0.67$
- $P(\text{B}) = 0.25$

So $P(\text{B})$ does not equal $P(\text{B}|\text{A})$ which means that $\text{B} and \text{A}$ are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND A}) = 0.20$, not 0.

✓ Example 4.2.9

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(\text{F}) = 0.60$; $P(\text{L}) = 0.50$
- $P(\text{F AND L}) = 0.45$
- $P(\text{L}|\text{F}) = 0.75$

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(\text{F}|\text{L})$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(\text{F AND L}) = P(\text{F})P(\text{L})$. We are given that $P(\text{F AND L}) = 0.45$, but $P(\text{F})P(\text{L}) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(\text{F AND L})$ does not equal $P(\text{F})P(\text{L})$.

Solution 2

Check whether $P(\text{L}|\text{F})$ equals $P(\text{L})$. We are given that $P(\text{L}|\text{F}) = 0.75$, but $P(\text{L}) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

? Exercise 4.2.9

Mark is deciding which route to take to work. His choices are $\mathbf{I = \text{the Interstate}}$ and $\mathbf{F = \text{Fifth Street}}$

- $P(\mathbf{I}) = 0.44$ and $P(\mathbf{F}) = 0.55$
- $P(\mathbf{I \text{ AND } F}) = 0$ because Mark will take only one route to work.

What is the probability of $P(\mathbf{I \text{ OR } F})$?

Answer

Because $P(\mathbf{I \text{ AND } F}) = 0$,

$P(\mathbf{I \text{ OR } F}) = P(\mathbf{I}) + P(\mathbf{F}) - P(\mathbf{I \text{ AND } F}) = 0.44 + 0.56 - 0 = 1$

✓ Example 4.2.10

a. Toss one fair coin (the coin has two sides, **H** and **T**). The outcomes are _____. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is _____.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are $H1$ and $T6$.)
e. Event $\mathbf{A}$ = heads ($\mathbf{H}$) on the coin followed by an even number (2, 4, 6) on the die.
$\mathbf{A}$ = {_____}. Find $P(\mathbf{A})$.
f. Event $\mathbf{B}$ = heads on the coin followed by a three on the die. $\mathbf{B}$ = {_____}. Find $P(\mathbf{B})$.
g. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive? (Hint: What is $P(\mathbf{A \text{ AND } B})$? If $P(\mathbf{A \text{ AND } B}) = 0$, then $\mathbf{A}$ and $\mathbf{B}$ are mutually exclusive.)
h. Are $\mathbf{A}$ and $\mathbf{B}$ independent? (Hint: Is $P(\mathbf{A \text{ AND } B}) = P(\mathbf{A})P(\mathbf{B})$? If $P(\mathbf{A \text{ AND } B}) = P(\mathbf{A})P(\mathbf{B})$, then $\mathbf{A}$ and $\mathbf{B}$ are independent. If not, then they are dependent).

Solution

a. **H** and **T**; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. $T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6$
e. $\mathbf{A} = \{H2, H4, H6\}$; $P(\mathbf{A}) = \frac{3}{12}$
f. $\mathbf{B} = \{H3\}$; $P(\mathbf{B}) = \frac{1}{12}$
g. Yes, because $P(\mathbf{A \text{ AND } B}) = 0$
h. $P(\mathbf{A \text{ AND } B}) = 0$. $P(\mathbf{A})P(\mathbf{B}) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(\mathbf{A \text{ AND } B})$ does not equal $P(\mathbf{A})P(\mathbf{B})$, so $\mathbf{A}$ and $\mathbf{B}$ are dependent.

? Exercise 4.2.10

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let **T** be the event of getting the white ball twice, **F** the event of picking the white ball first, **S** the event of picking the white ball in the second drawing.

a. Compute $P(\mathbf{T})$.
b. Compute $P(\mathbf{T}|\mathbf{F})$.
c. Are **T** and **F** independent?.
d. Are **F** and **S** mutually exclusive?

e. Are F and S independent?

Answer

a. $P(\text{T}) = \frac{1}{4}$
b. $P(\text{T}|\text{F}) = \frac{1}{2}$
c. No
d. No
e. Yes

References

1. Lopez, Shane, Preety Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).
2. Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

- If A and B are independent, $P(\text{A AND B}) = P(\text{A})P(\text{B})$, $P(\text{A}|\text{B}) = P(\text{A})$ and $P(\text{B}|\text{A}) = P(\text{B})$.
- If A and B are mutually exclusive, $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) \text{and} P(\text{A AND B}) = 0$.

? Exercise 4.2.11

E and F are mutually exclusive events. $P(\text{E}) = 0.4$; $P(\text{F}) = 0.5$. Find $P(\text{E}|\text{F})$.

? Exercise 4.2.12

J and K are independent events. $P(\text{J}|\text{K}) = 0.3$. Find $P(\text{J})$.

Answer

$P(\text{J}) = 0.3$

? Exercise 4.2.13

U and V are mutually exclusive events. $P(\text{U}) = 0.26$; $P(\text{V}) = 0.37$. Find:

a. $P(\text{U AND V}) =$
b. $P(\text{U}|\text{V}) =$
c. $P(\text{U OR V}) =$

? Exercise 4.2.14

Q and R are independent events. $P(\text{Q}) = 0.4$ and $P(\text{Q AND R}) = 0.1$. Find $P(\text{R})$.

Answer

$P(\text{Q AND R}) = P(\text{Q})P(\text{R})$

$0.1 = (0.4)P(\text{R})$

$P(\text{R}) = 0.25$

Bringing It Together

Exercise 4.2.16

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into Table.

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\#1–33} | \leq 210 \text{ pounds})$?

Exercise 4.2.17

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(\text{C}) =$ _____
b. $P(\text{P}|\text{C}) =$ _____
c. $P(\text{P}|\text{C'}) =$ _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Answer

a. $P(\text{C}) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise 4.2.18

Given events G and H : $P(\text{G}) = 0.43$; $P(\text{H}) = 0.26$; $P(\text{H AND G}) = 0.14$

a. Find $P(\text{H OR G})$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise 4.2.19

Given events J and K : $P(\text{J}) = 0.18$; $P(\text{K}) = 0.37$; $P(\text{J OR K}) = 0.45$

a. Find $P(\text{J AND K})$.
b. Find the probability of the complement of event (J AND K).

c. Find the probability of the complement of event (J AND K).

Answer

a. $P(\text{J OR K}) = P(\text{J}) + P(\text{K}) - P(\text{J AND K}); 0.45 = 0.18 + 0.37 - P(\text{J AND K})$; solve to find $P(\text{J AND K}) = 0.10$
b. $P(\text{NOT (J AND K)}) = 1 - P(\text{J AND K}) = 1 - 0.10 = 0.90$
c. $P(\text{NOT (J OR K)}) = 1 - P(\text{J OR K}) = 1 - 0.45 = 0.55$

Glossary

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

The Conditional Probability of One Event Given Another Event

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The OR of Two Events

An outcome is in the event A OR B if the outcome is in A, is in B, or is in both A and B.

This page titled 4.2: Independent and Mutually Exclusive Events is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.3: The Addition and Multiplication Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a sample space, then:

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{4.3.1}$$

This rule may also be written as:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are *independent*, then

$$P(A|B) = P(A).$$

and Equation 4.3.1 becomes

$$P(A \text{ AND } B) = P(A)P(B).$$

The Addition Rule

If A and B are defined on a sample space, then:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) \tag{4.3.2}$$

If A and B are **mutually exclusive**, then

$$P(A \text{ AND } B) = 0.$$

and Equation 4.3.2 becomes

$$P(A \text{ OR } B) = P(A) + P(B).$$

Example 4.3.1

Klaus is trying to choose where to go on vacation. His two choices are: $\mathbf{A} = \textbf{New Zealand}$ and $\mathbf{B} = \textbf{Alaska}$.

- Klaus can only afford one vacation. The probability that he chooses $\mathbf{A}$ is $P(\mathbf{A}) = 0.6$ and the probability that he chooses $\mathbf{B}$ is $P(\mathbf{B}) = 0.35$.
- $P(\mathbf{A} \text{ AND } \mathbf{B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\mathbf{A} \text{ OR } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. $\mathbf{A}$ = the event Carlos is successful on his first attempt. $P(\mathbf{A}) = 0.65$. $\mathbf{B}$ = the event Carlos is successful on his second attempt. $P(\mathbf{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are $\mathbf{A}$ and $\mathbf{B}$ independent?
d. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive?

Solutions

a. The problem is asking you to find $P(\text{A AND B}) = P(\text{B AND A})$. Since $P(\text{B}|\text{A}) = 0.90 : P(\text{B AND A}) = P(\text{B}|\text{A})P(\text{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\text{A OR B})$.

$$P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B}) = 0.65 + 0.65 - 0.585 = 0.715 \tag{4.3.3}$$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\text{B AND A}) = 0.585$.

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \tag{4.3.4}$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \tag{4.3.5}$$

So, $P(\text{B AND A})$ is **not** equal to $P(\text{B})P(\text{A})$.

d. No, they are not because $P(\text{A and B}) = 0.585$.

To be mutually exclusive, $P(\text{A AND B})$ must equal zero.

Exercise 4.3.1

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. $\text{C} =$ the event that Helen makes the first shot. $P(\text{C}) = 0.75$. $\text{D} =$ the event Helen makes the second shot. $P(\text{D}) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Answer

$$P(\text{D}|\text{C}) = 0.85 \tag{4.3.6}$$

$$P(\text{C AND D}) = P(\text{D AND C}) \tag{4.3.7}$$

$$P(\text{D AND C}) = P(\text{D}|\text{C})P(\text{C}) = (0.85)(0.75) = 0.6375 \tag{4.3.8}$$

Helen makes the first and second free throws with probability 0.6375.

Example 4.3.2

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Answer

a. $\frac{28}{150}$
b. $\frac{80}{150}$
c. $\frac{40}{150}$
d. $P(\text{advanced AND intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

e. No, these are not independent events.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{4.3.9}$$

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{4.3.10}$$

$$0.0667 \neq 0.0996 \tag{4.3.11}$$

Exercise 4.3.2

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Answer

$$P = \frac{200 - 140 - 40}{200} = \frac{20}{200} = 0.1 \tag{4.3.12}$$

Example 4.3.3

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, $\mathrm{M|S}$ = math given speech

a. What is the probability that Felicity enrolls in math and speech?
   Find $P(\mathrm{M\ AND\ S}) = P(\mathrm{M|S})P(\mathrm{S})$.
b. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(\mathrm{M\ OR\ S}) = P(\mathrm{M}) + P(\mathrm{S}) - P(\mathrm{M\ AND\ S})$.
c. Are M and S independent? Is $P(\mathrm{M|S}) = P(\mathrm{M})$?
d. Are M and S mutually exclusive? Is $P(\mathrm{M\ AND\ S}) = 0$?

Answer

a. 0.1625, b. 0.6875, c. No, d. No

Exercise 4.3.3

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(\mathrm{B}) = 0.40, P(\mathrm{D}) = 0.30$ and $P(\mathrm{D|B}) = 0.5$.

a. Find $P(\mathrm{B\ AND\ D})$.
b. Find $P(\mathrm{B\ OR\ D})$.

Answer

a. $P(\mathrm{B\ AND\ D}) = P(\mathrm{D|B})P(\mathrm{B}) = (0.5)(0.4) = 0.20$.
b. $P(\mathrm{B\ OR\ D}) = P(\mathrm{B}) + P(\mathrm{D}) - P(\mathrm{B\ AND\ D}) = 0.40 + 0.30 - 0.20 = 0.50$

Example 4.3.4

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative?
d. What is the probability that the woman has breast cancer or tests negative?
e. Are having breast cancer and testing negative independent events?
f. Are having breast cancer and testing negative mutually exclusive?

Answers

a. $P(\text{B}) = 0.143; P(\text{N}) = 0.85$
b. $P(\text{N}|\text{B}) = 0.02$
c. $P(\text{B AND N}) = P(\text{B})P(\text{N}|\text{B}) = (0.143)(0.02) = 0.0029$
d. $P(\text{B OR N}) = P(\text{B}) + P(\text{N}) - P(\text{B AND N}) = 0.143 + 0.85 - 0.0029 = 0.9901$
e. No. $P(\text{N}) = 0.85; P(\text{N}|\text{B}) = 0.02$. So, $P(\text{N}|\text{B})$ does not equal $P(\text{N})$.
f. No. $P(\text{B AND N}) = 0.0029$. For B and N to be mutually exclusive, $P(\text{B AND N})$ must be zero

Exercise 4.3.4

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Answer

Let $\text{A} =$ student is a senior going to college.

Let $\text{B} =$ student plays sports.

$$P(\text{B}) = \frac{140}{200}$$

$$P(\text{B}|\text{A}) = \frac{50}{140}$$

$$P(\text{A AND B}) = P(\text{B}|\text{A})P(\text{A})$$

$$P(\text{A AND B}) = (\frac{140}{200})(\frac{50}{140}) = \frac{1}{4}$$

Example 4.3.5

Refer to the information in Example 4.3.4. $\text{P} =$ tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(\text{P}|\text{B}) = 1 - P(\text{N}|\text{B})$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(\text{B AND P}) = P(\text{P}|\text{B})P(\text{B})$.
c. What is the probability that a woman does not develop breast cancer. Find $P(\text{B}') = 1 - P(\text{B})$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(\text{P}) = 1 - P(\text{N})$.

Answer

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Exercise 4.3.5

A student goes to the library. Let events $\text{B} =$ the student checks out a book and $\text{D} =$ the student checks out a DVD. Suppose that $P(\text{B}) = 0.40, P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B}')$.
b. Find $P(\text{D AND B})$.
c. Find $P(\text{B}|\text{D})$.
d. Find $P(\text{D AND B}')$.
e. Find $P(\text{D}|\text{B}')$.

Answer

a. $P(\text{B}') = 0.60$
b. $P(\text{D AND B}) = P(\text{D}|\text{B})P(\text{B}) = 0.20$
c. $P(\text{B}|\text{D}) = \frac{P(\text{B AND D})}{P(\text{D})} = \frac{(0.20)}{(0.30)} = 0.66$
d. $P(\text{D AND B}') = P(\text{D}) - P(\text{D AND B}) = 0.30 - 0.20 = 0.10$
e. $P(\text{D}|\text{B}') = P(\text{D AND B}')P(\text{B}') = (P(\text{D}) - P(\text{D AND B}))(0.60) = (0.10)(0.60) = 0.06$

References

1. DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).
2. Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).
3. “Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).
4. “Roulette.” Wikipedia. Available online at http://en.Wikipedia.org/wiki/Roulette (accessed May 2, 2013).
5. Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at www.census.gov/hhes/socdemo/l...acs/ACS-12.pdf (accessed May 2, 2013).
6. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
7. Data from U.S. Census Bureau.
8. Data from the Wall Street Journal.
9. Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at www.ropercenter.uconn.edu/ (accessed May 2, 2013).
10. Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(\text{A AND B}) = P(\text{A}|\text{B})P(\text{B})$

The addition rule: $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

Exercise 4.3.5

Find $P(C)$.

Exercise 4.3.6

Find $P(L)$.

Answer

0.376

Exercise 4.3.7

Find $P(C|L)$.

Exercise 4.3.8

In words, what is $C|L$?

Answer

$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise 4.3.9

Find $P(L \text{ AND } C)$

Exercise 4.3.10

In words, what is $L \text{ AND } C$?

Answer

$L \text{ AND } C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Exercise 4.3.11

Are L and C independent events? Show why or why not.

Exercise 4.3.12

Find $P(L \text{ OR } C)$.

Answer

0.6492

Exercise 4.3.13

In words, what is $L \text{ OR } C$?

? Exercise 4.3.14

Are $\mathbf{L}$ and $\mathbf{C}$ mutually exclusive events? Show why or why not.

Answer

No, because $P(\mathbf{L \text{ AND } C})$ does not equal 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events $\mathbf{A}$ and $\mathbf{B}$ are independent if one of the following is true:

1. $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A})$
2. $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$
3. $P(\mathbf{A \text{ AND } B}) = P(\mathbf{A})P(\mathbf{B})$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events $\mathbf{A}$ and $\mathbf{B}$ are mutually exclusive, then $P(\mathbf{A \text{ AND } B}) = 0$.

This page titled 4.3: The Addition and Multiplication Rules of Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **3.4: Two Basic Rules of Probability** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

4.3.1: Contingency Tables

A *contingency table* provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 4.3.1.1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find $P(\text{Person is a cell phone user})$.
b. Find $P(\text{person had no violation in the last year})$.
c. Find $P(\text{Person had no violation in the last year AND was a cell phone user})$.
d. Find $P(\text{Person is a cell phone user OR person had no violation in the last year})$.
e. Find $P(\text{Person is a cell phone user GIVEN person had a violation in the last year})$.
f. Find $P(\text{Person had no violation last year GIVEN person was not a cell phone user})$

Answer

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$
b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$
c. $\frac{280}{755}$
d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$
e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)
f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not cell phone users.)

Exercise 4.3.1.1

Table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising}|\text{no injury in the last year})$?

Answer

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$

b. $P(\text{athlete stretches before exercising}|\text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example 4.3.1.2

Table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

a. Complete the table.

b. Are the events "being female" and "preferring the coastline" independent events? Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$
3. Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let $\mathbf{M} =$ being male, and let $\mathbf{L} =$ prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(\_\_\_|\_\_\_) = \_\_\_$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let $\mathbf{F} =$ being female, and let $\mathbf{P} =$ prefers mountain peaks.

1. Find $P(\mathbf{F})$.
2. Find $P(\mathbf{P})$.
3. Find $P(\mathbf{F} \text{ AND } \mathbf{P})$.
4. Find $P(\mathbf{F} \text{ OR } \mathbf{P})$.

Answers

a.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

b.

$$P(\mathbf{F} \text{ AND } \mathbf{C}) = \frac{18}{100} = 0.18$$

$$P(\text{F})P(\text{C}) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(\text{F AND C}) \neq P(\text{F})P(\text{C})$, so the events F and C are not independent.

c.

1. The word 'given' tells you that this is a conditional.
2. $P(\text{M}|\text{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

a. Find $P(\text{F})$.
b. Find $P(\text{P})$.
c. Find $P(\text{F AND P})$.
d. Find $P(\text{F OR P})$.

d.

1. $P(\text{F}) = \frac{45}{100}$
2. $P(\text{P}) = \frac{25}{100}$
3. $P(\text{F AND P}) = \frac{11}{100}$
4. $P(\text{F OR P}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 4.3.1.2

Table shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Answer

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$
$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$
$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example 4.3.1.3

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.
b. What is the probability that Alissa does not catch Muddy?
c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. $\frac{41}{60}$
c. $\frac{9}{19}$

Example 4.3.1.4

Table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

a. Find $P(\text{2009 AND Robbery})$.
b. Find $P(\text{2010 AND Burglary})$.
c. Find $P(\text{2010 OR Burglary})$.
d. Find $P(\text{2011}|\text{Rape})$.
e. Find $P(\text{Vehicle}|\text{2008})$.

Answer

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Exercise 4.3.1.3

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

1. “Blood Types.” American Red Cross, 2013. Available online at www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).
2. Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.
3. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
4. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

5. “Human Blood Types.” Unite Blood Services, 2011. Available online at www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).
6. Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).
7. “United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. Table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

? Exercise 3.5.4

Find P(musician is a female).

? Exercise 3.5.5

Find $P(\text{musician is a male AND had private instruction})$.

Answer

$$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$$

? Exercise 3.5.6

Find P(musician is a female OR is self taught).

? Exercise 3.5.7

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Answer

The events are not mutually exclusive. It is possible to be a female musician who learned music in school.

Bringing it Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Exercise 3.5.8

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Exercise 3.5.9

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Answer

$\frac{35,065}{100,450}$

Exercise 3.5.10

Find the probability that the person was Latino.

Exercise 3.5.11

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

Exercise 3.5.12

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

Exercise 3.5.13

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4,715}{15,273}$.

? Exercise 3.5.14

Prove that smoking level/day and ethnicity are dependent events.

Glossary

contingency table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

This page titled 4.3.1: Contingency Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.3.2: Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A *tree diagram* is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 4.3.2.1: Probabilities from Sampling with replacement

In an urn, there are 11 balls. Three balls are red (**R**) and eight balls are blue (**B**). Draw two balls, one at a time, with replacement (remember that "with replacement" means that you put the first ball back in the urn before you select the second ball). The tree diagram using frequencies that show all the possible outcomes follows.

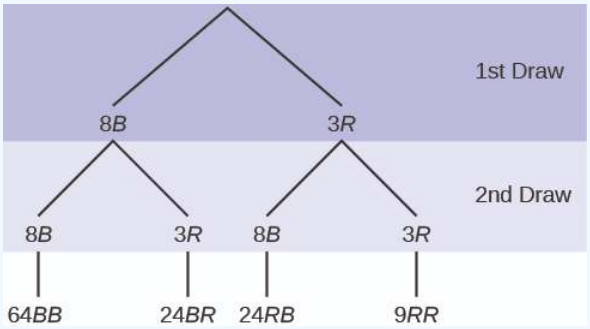


Figure 4.3.2.1: Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as $R1$, $R2$, and $R3$ and each blue ball as $B1$, $B2$, $B3$, $B4$, $B5$, $B6$, $B7$, and $B8$. Then the nine RR outcomes can be written as:

$R1R1$ $R1R2$ $R1R3$ $R2R1$ $R2R2$ $R2R3$ $R3R1$ $R3R2$ $R3R3$

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

Exercise 4.3.2.1

In a standard deck, there are 52 cards. 12 cards are face cards (event **F**) and 40 cards are not face cards (event **N**). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(\mathbf{FF})$.

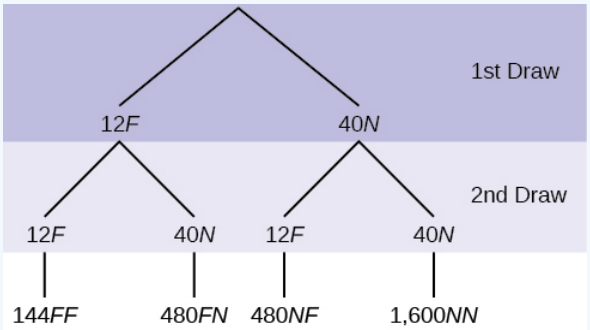


Figure 4.3.2.2:

Answer

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(\text{FF}) = \frac{144}{144+480+480+1,600} = \frac{144}{2,704} = \frac{9}{169} \tag{4.3.2.1}$$

1. a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...
2. b. Using the tree diagram, calculate *P*(*RR*).
3. c. Using the tree diagram, calculate $P(\text{RB OR BR})$.
4. d. Using the tree diagram, calculate $P(\text{R on 1st draw AND B on 2nd draw})$.
5. e. Using the tree diagram, calculate *P*(*R* on 2nd draw GIVEN *B* on 1st draw).
6. Using the tree diagram, calculate $P(\text{BB})$.
7. g. Using the tree diagram, calculate $P(\text{B on the 2nd draw given R on the first draw})$.

Solution

a. *B*1*R*1; *B*1*R*2; *B*1*R*3; *B*2*R*1; *B*2*R*2; *B*2*R*3; *B*3*R*1; *B*3*R*2; *B*3*R*3; *B*4*R*1; *B*4*R*2; *B*4*R*3; *B*5*R*1; *B*5*R*2; *B*5*R*3; *B*6*R*1; *B*6*R*2; *B*6*R*3 *B*7*R*1; *B*7*R*2; *B*7*R*3; *B*8*R*1; *B*8*R*2; *B*8*R*3

b. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. $P(\text{R on 1st draw AND B on 2nd draw}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. *P*(*R* on 2nd draw GIVEN *B* on 1st draw) = *P*(*R* on 2nd|*B* on 1st) = $\frac{24}{88} = \frac{3}{11}$ This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88} = \frac{3}{11}$

f. $P(\text{BB}) = \frac{64}{121}$

g. $P(\text{B on 2nd draw}|\text{R on 1st draw}) = \frac{8}{11}$. There are 9 + 24 outcomes that have R on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Example 4.3.2.2: Probabilities from Sampling without replacement

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. (remember that "without replacement" means that you do not put the first ball back before you select the second marble). Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

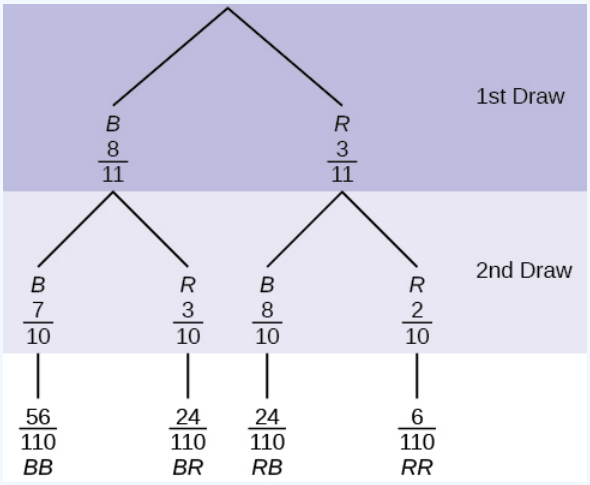


Figure 4.3.2.3: Total $= \frac{56+24+24+6}{110} = \frac{110}{110} = 1$

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw ***without replacement****, so that on the second draw there are ten marbles left in the urn.*

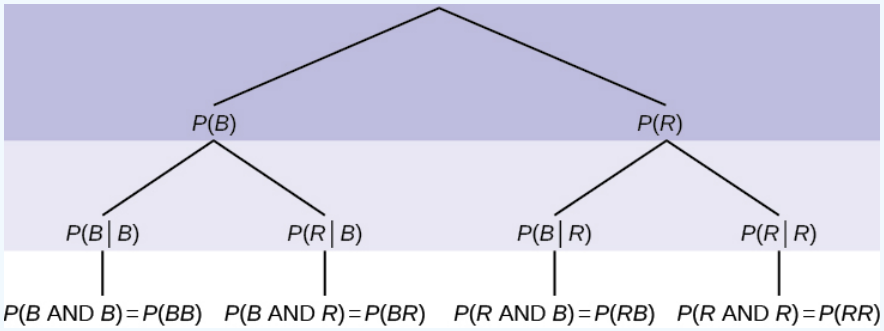
Calculate the following probabilities using the tree diagram.

a. $P(\text{RR}) =$ _____
b. Fill in the blanks: $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\_\_\_)(\_\_\_) = \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) =$
d. Fill in the blanks: $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = (\_\_\_)(\_\_\_) = \frac{24}{110}$
e. Find $P(\text{BB})$.
f. Find $P(\text{B on 2nd}|\text{R on 1st})$.

Answers

a. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$
b. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) = \frac{3}{10}$
d. $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$
e. $P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$
f. Using the tree diagram, $P(\text{B on 2nd}|\text{R on 1st}) = P(\text{R}|\text{B}) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



- $P(\text{R}|\text{R})$ here means $P(\text{R on 2nd}|\text{R on 1st})$
- $P(\text{B}|\text{R})$ here means $P(\text{B on 2nd}|\text{R on 1st})$
- $P(\text{R}|\text{B})$ here means $P(\text{R on 2nd}|\text{B on 1st})$
- $P(\text{B}|\text{B})$ here means $P(\text{B on 2nd}|\text{B on 1st})$

? Exercise 4.3.2.2

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

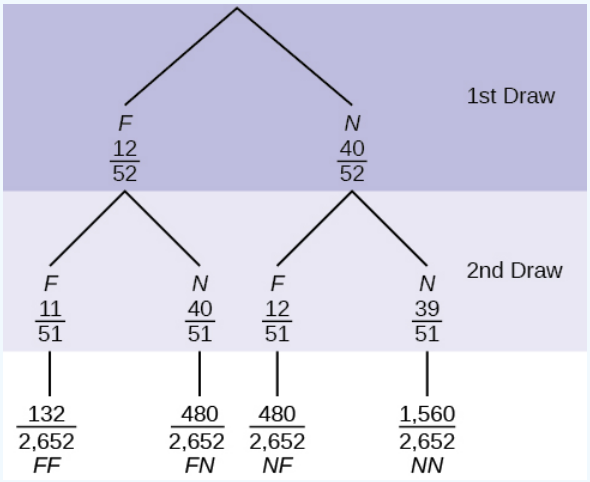


Figure 4.3.2.4:

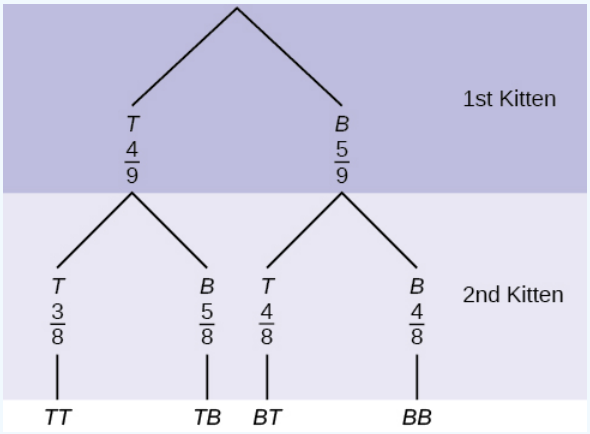
a. Find $P(\text{FN OR NF})$.
b. Find $P(\text{N}|\text{F})$.
c. Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
d. Find $P(\text{at least on face card})$.
Hint: "At least one face card" means one or two face cards.

Answer

a. $P(\text{FN OR NF}) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(\text{N}|\text{F}) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480+480+1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132+480+480)}{2,652} = \frac{1,092}{2,652}$

Example 4.3.2.3

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



a. What is the probability that both kittens are tabby?
a. $(\frac{1}{2})(\frac{1}{2})$ b. $(\frac{4}{9})(\frac{4}{9})$ c. $(\frac{4}{9})(\frac{3}{8})$ d. $(\frac{4}{9})(\frac{5}{9})$
b. What is the probability that one kitten of each coloring is selected?
a. $(\frac{4}{9})(\frac{5}{9})$ b. $(\frac{4}{9})(\frac{5}{8})$ c. $(\frac{4}{9})(\frac{5}{9}) + (\frac{5}{9})(\frac{4}{9})$ d. $(\frac{4}{9})(\frac{5}{8}) + (\frac{5}{9})(\frac{4}{8})$
c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Answer

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Exercise 4.3.2.3

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Answer

$(\frac{4}{7})(\frac{3}{6}) + (\frac{3}{7})(\frac{4}{6})$

Venn Diagram

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example 4.3.2.4

Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event **A** = {1, 2, 3, 4, 5, 6} and event **B** = {6, 7, 8, 9}. Then **A AND B** = {6} and **A OR B** = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:

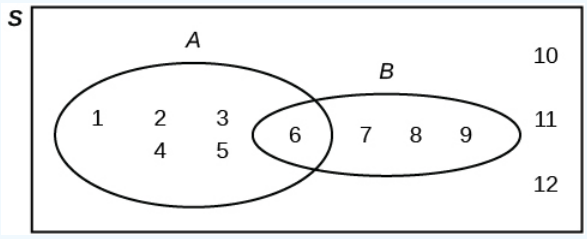


Figure **4.3.2.5**:

Exercise 4.3.2.4

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event **C** = {green, blue, purple} and event **P** = {red, yellow, blue}. Then **C AND P** = {blue} and **C OR P** = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

Answer

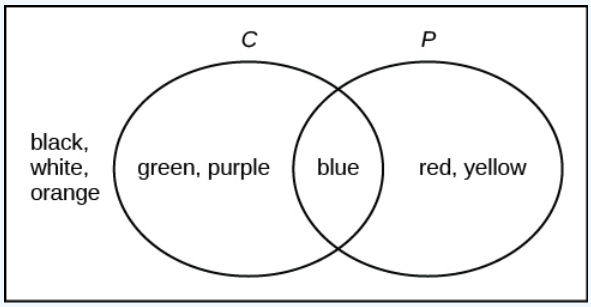


Figure **4.3.2.6**:

Example 4.3.2.5

Flip two fair coins. Let **A** = tails on the first coin. Let **B** = tails on the second coin. Then **A** = {*TT*,*TH*} and **B** = {*TT*, *HT*}. Therefore, **A AND B** = {*TT*}. **A OR B** = {*TH*, *TT*, *HT*}.

The sample space when you flip two fair coins is ***X*** = {*HH*, *HT*, *TH*, *TT*}. The outcome *HH* is in **NEITHER A NOR B**. The Venn diagram is as follows:

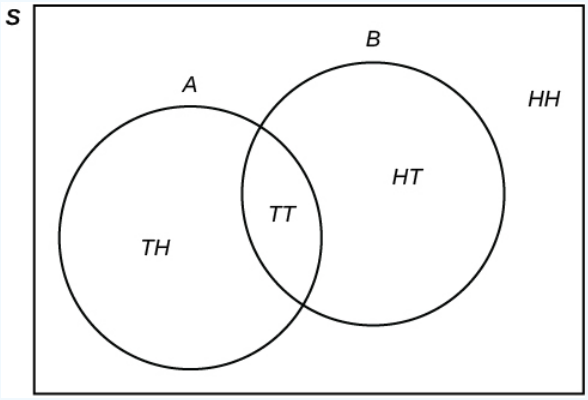
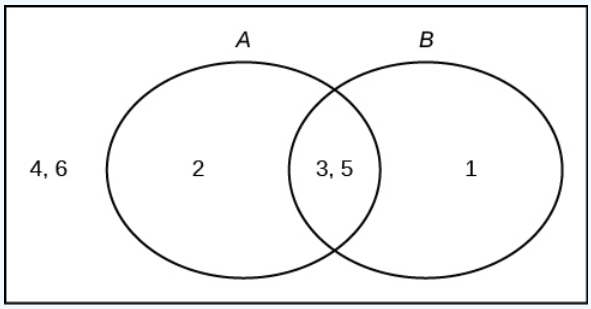


Figure **4.3.2.7**:

Exercise 4.3.2.5

Roll a fair, six-sided die. Let **A** = a prime number of dots is rolled. Let **B** = an odd number of dots is rolled. Then **A** = {2, 3, 5} and **B** = {1, 3, 5}. Therefore, **A AND B** = {3, 5}. **A OR B** = {1, 2, 3, 5}. The sample space for rolling a fair die is **S** = {1, 2, 3, 4, 5, 6}. Draw a Venn diagram representing this situation.

Answer



Example 4.3.2.6: Probability and Venn Diagrams

Forty percent of the students at a local college belong to a club and 50% work part time. Five percent of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let **C** = student belongs to a club and **PT** = student works part time.

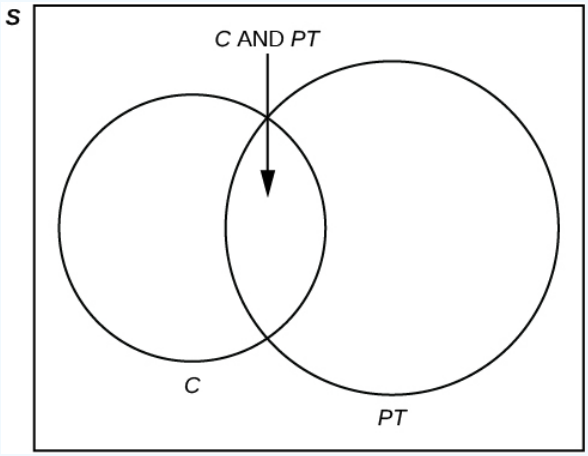


Figure 4.3.2.8:

If a student is selected at random, find

- the probability that the student belongs to a club. $P(\text{C}) = 0.40$
- the probability that the student works part time. $P(\text{PT}) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(\text{C AND PT}) = 0.05$
- the probability that the student belongs to a club **given** that the student works part time. $P(\text{C}|\text{PT}) = \frac{P(\text{C AND PT})}{P(\text{PT})} = \frac{0.05}{0.50} = 0.1$
- the probability that the student belongs to a club **OR** works part time. $P(\text{C OR PT}) = P(\text{C}) + P(\text{PT}) - P(\text{C AND PT})$ $= 0.40 + 0.50 - 0.05 = 0.85$

Exercise 4.3.2.6

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let **W** = works a second job and **S** = spouse also works.

Answer

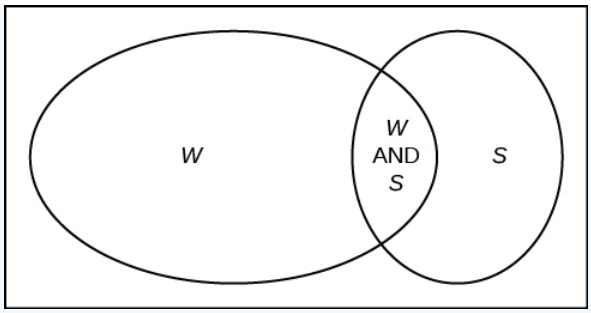


Figure **4.3.2.9**:

Example 4.3.2.7

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.

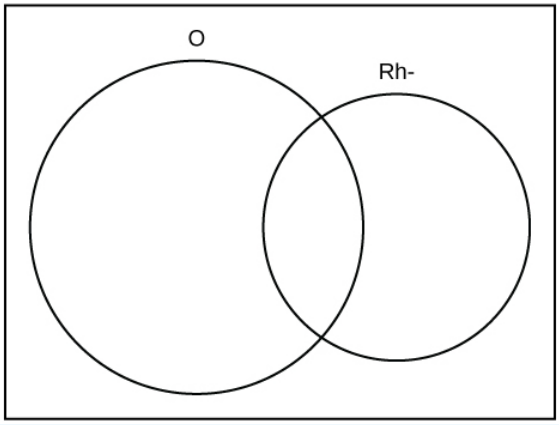


Figure **4.3.2.10**:

The "O" circle represents the African Americans with type O blood. The "Rh-" oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let $\mathbf{O} =$ African American with Type O blood and $\mathbf{R} =$ African American with Rh- factor.

a. $P(\mathbf{O}) =$ _____
b. $P(\mathbf{R}) =$ _____
c. $P(\mathbf{O} \text{ AND } \mathbf{R}) =$ _____
d. $P(\mathbf{O} \text{ OR } \mathbf{R}) =$ _____
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Answer

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Exercise 4.3.2.7

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.

c. In the Venn diagram, describe the overlapping area using a complete sentence.

d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Answer

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.

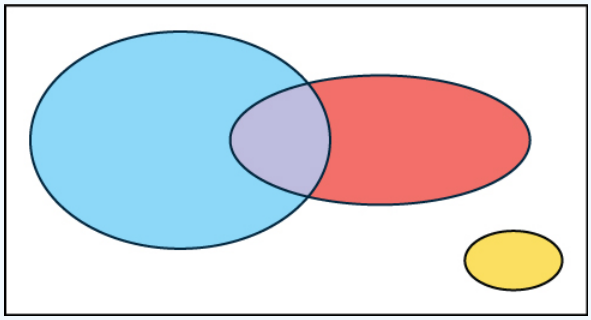


Figure **4.3.2.11**:

b. $P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8$.

c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

1. Data from Clara County Public H.D.
2. Data from the American Cancer Society.
3. Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).
4. Data from the Federal Highway Administration, part of the United States Department of Transportation.
5. Data from the United States Census Bureau, part of the United States Department of Commerce.
6. Data from USA Today.
7. “Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).
8. “Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Review

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize. A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Glossary

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

This page titled 4.3.2: Tree and Venn Diagrams is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.4: Counting Rules

Learning Objectives

In this section you will learn to

1. Use trees to count possible outcomes in a multi-step process
2. Use the multiplication axiom to count possible outcomes in a multi-step process.

In this chapter, we are trying to develop counting techniques that will be used in the next chapter to study probability. One of the most fundamental of such techniques is called the Multiplication Axiom. Before we introduce the multiplication axiom, we first look at some examples.

Example 4.4.1

If a woman has two blouses and three skirts, how many different outfits consisting of a blouse and a skirt can she wear?

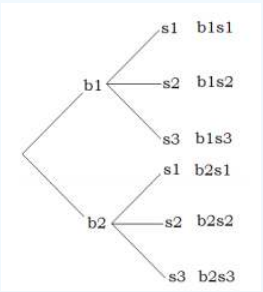
Solution

Suppose we call the blouses b_1 and b_2, and skirts s_1, s_2, and s_3.

We can have the following six outfits.

$$b_1s_1, b_1s_2, b_1s_3, b_2s_1, b_2s_2, b_2s_3$$

Alternatively, we can draw a tree diagram:



The tree diagram gives us all six possibilities. The method involves two steps. First the woman chooses a blouse. She has two choices: blouse one or blouse two. If she chooses blouse one, she has three skirts to match it with; skirt one, skirt two, or skirt three. Similarly if she chooses blouse two, she can match it with each of the three skirts, again. The tree diagram helps us visualize these possibilities.

The reader should note that the process involves two steps. For the first step of choosing a blouse, there are two choices, and for each choice of a blouse, there are three choices of choosing a skirt. So altogether there are $2 \cdot 3 = 6$ possibilities.

If, in the previous example, we add the shoes to the outfit, we have the following problem.

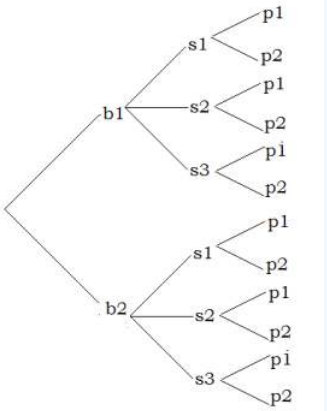
Example 4.4.2

If a woman has two blouses, three skirts, and two pumps, how many different outfits consisting of a blouse, a skirt, and a pair of pumps can she wear?

Solution

Suppose we call the blouses b_1 and b_2, the skirts s_1, s_2, and s_3, and the pumps p_1, and p_2.

The following tree diagram results.



We count the number of branches in the tree, and see that there are 12 different possibilities.

This time the method involves three steps. First, the woman chooses a blouse. She has two choices: blouse one or blouse two. Now suppose she chooses blouse one. This takes us to step two of the process which consists of choosing a skirt. She has three choices for a skirt, and let us suppose she chooses skirt two. Now that she has chosen a blouse and a skirt, we have moved to the third step of choosing a pair of pumps. Since she has two pairs of pumps, she has two choices for the last step. Let us suppose she chooses pumps two. She has chosen the outfit consisting of blouse one, skirt two, and pumps two, or $b_1 s_2 p_2$. By looking at the different branches on the tree, one can easily see the other possibilities.

The important thing to observe here, again, is that this is a three step process. There are two choices for the first step of choosing a blouse. For each choice of a blouse, there are three choices of choosing a skirt, and for each combination of a blouse and a skirt, there are two choices of selecting a pair of pumps.

All in all, we have $2 \cdot 3 \cdot 2 = 12$ different possibilities.

Tree diagrams help us visualize the different possibilities, but they are not practical when the possibilities are numerous. Besides, we are mostly interested in finding the number of elements in the set and not the actual list of all possibilities; once the problem is envisioned, we can solve it without a tree diagram. The two examples we just solved may have given us a clue to do just that.

Let us now try to solve Example **4.4.2** without a tree diagram. The problem involves three steps: choosing a blouse, choosing a skirt, and choosing a pair of pumps. The number of ways of choosing each are listed below. By multiplying these three numbers we get 12, which is what we got when we did the problem using a tree diagram.

The number of ways of choosing a blouse	The number of ways of choosing a skirt	The number of ways of choosing pumps
2	3	2

The procedure we just employed is called the multiplication axiom.

The Multiplication Axiom

If a task can be done in m ways, and a second task can be done in n ways, then the operation involving the first task followed by the second can be performed in $m \cdot n$ ways.

The general multiplication axiom is not limited to just two tasks and can be used for any number of tasks.

Example 4.4.3

A truck license plate consists of a letter followed by four digits. How many such license plates are possible?

Solution

Since there are 26 letters and 10 digits, we have the following choices for each.

Letter	Digit	Digit	Digit	Digit
26	10	10	10	10

Therefore, the number of possible license plates is $26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 260,000$.

Example 4.4.4

In how many different ways can a 3-question true-false test be answered?

Solution

Since there are two choices for each question, we have

Question 1	Question 2	Question 3
2	2	2

Applying the multiplication axiom, we get $2 \cdot 2 \cdot 2 = 8$ different ways.

We list all eight possibilities: TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF

The reader should note that the first letter in each possibility is the answer corresponding to the first question, the second letter corresponds to the answer to the second question, and so on. For example, TFF, says that the answer to the first question is given as true, and the answers to the second and third questions false.

Example 4.4.5

In how many different ways can four people be seated in a row?

Solution

Suppose we put four chairs in a row, and proceed to put four people in these seats.

There are four choices for the first chair we choose. Once a person sits down in that chair, there are only three choices for the second chair, and so on. We list as shown below.

4	3	2	1

So there are altogether $4 \cdot 3 \cdot 2 \cdot 1 = 24$ different ways.

Example 4.4.6

How many three-letter word sequences can be formed using the letters { A, B, C } if no letter is to be repeated?

Solution

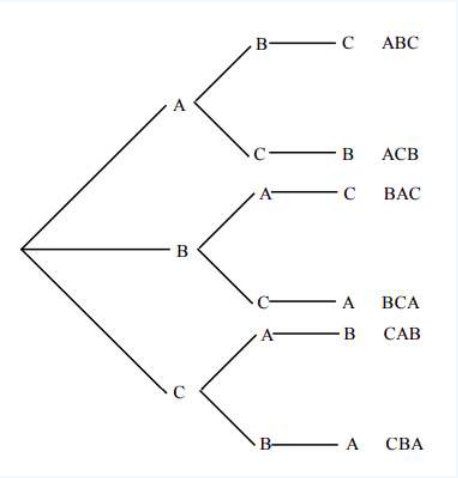
The problem is very similar to the previous example.

Imagine a child having three building blocks labeled A, B, and C. Suppose he puts these blocks on top of each other to make word sequences. For the first letter he has three choices, namely A, B, or C. Let us suppose he chooses the first letter to be a B, then for the second block which must go on top of the first, he has only two choices: A or C. And for the last letter he has only one choice. We list the choices below.

3	2	1

Therefore, 6 different word sequences can be formed.

Finally, we'd like to illustrate this with a tree diagram showing all six possibilities.



This page titled 4.4: Counting Rules is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.2: Tree Diagrams and the Multiplication Axiom** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

4.4.1: Permutations

Learning Objectives

In this section you will learn to

1. Count the number of possible permutations (ordered arrangement) of n items taken r at a time
2. Count the number of possible permutations when there are conditions imposed on the arrangements
3. Perform calculations using factorials

In the previous section, we were asked to find the word sequences formed by using the letters { A, B, C } if no letter is to be repeated. The tree diagram gave us the following six arrangements.

ABC, ACB, BAC, BCA, CAB, and CBA.

Arrangements like these, where order is important and no element is repeated, are called permutations.

Definition: Permutations

A permutation of a set of elements is an ordered arrangement where each element is used once.

Example 4.4.1.1

How many three-letter word sequences can be formed using the letters { A, B, C, D }?

Solution

There are four choices for the first letter of our word, three choices for the second letter, and two choices for the third.

4	3	2

Applying the multiplication axiom, we get $4 \cdot 3 \cdot 2 = 24$ different arrangements.

Example 4.4.1.2

How many permutations of the letters of the word ARTICLE have consonants in the first and last positions?

Solution

In the word ARTICLE, there are 4 consonants.

Since the first letter must be a consonant, we have four choices for the first position, and once we use up a consonant, there are only three consonants left for the last spot. We show as follows:

4						3

Since there are no more restrictions, we can go ahead and make the choices for the rest of the positions.

So far we have used up 2 letters, therefore, five remain. So for the next position there are five choices, for the position after that there are four choices, and so on. We get

4	5	4	3	2	1	3

So the total permutations are $4 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 3 = 1440$.

Example 4.4.1.3

Given five letters { A, B, C, D, E }. Find the following:

a. The number of four-letter word sequences.
b. The number of three-letter word sequences.
c. The number of two-letter word sequences.

Solution

The problem is easily solved by the multiplication axiom, and answers are as follows:

a. The number of four-letter word sequences is $5 \cdot 4 \cdot 3 \cdot 2 = 120$.
b. The number of three-letter word sequences is $5 \cdot 4 \cdot 3 = 60$.
c. The number of two-letter word sequences is $5 \cdot 4 = 20$.

We often encounter situations where we have a set of n objects and we are selecting r objects to form permutations. We refer to this as **permutations of n objects taken r at a time**, and we write it as **nPr**.

Therefore, the above example can also be answered as listed below.

a. The number of four-letter word sequences is 5P4 = 120.
b. The number of three-letter word sequences is 5P3 = 60.
c. The number of two-letter word sequences is 5P2 = 20.

Before we give a formula for nPr, we'd like to introduce a symbol that we will use a great deal in this as well as in the next chapter.

Definition: Factorial

$$n! = n(n-1)(n-2)(n-3)\cdots 3 \cdot 2 \cdot 1 \tag{4.4.1.1}$$

where n is a natural number.

$$0! = 1 \tag{4.4.1.2}$$

Now we define nPr.

Definition: nPr

The Number of Permutations of n Objects Taken r at a Time

$$nPr = n(n-1)(n-2)(n-3)\cdots(n-r+1) \tag{4.4.1.3}$$

or

$$nPr = \frac{n!}{(n-r)!} \tag{4.4.1.4}$$

where n and r are natural numbers.

The reader should become familiar with both formulas and should feel comfortable in applying either.

Example 4.4.1.4

Compute the following using both formulas.

a. 6P3
b. 7P2

Solution

We will identify n and r in each case and solve using the formulas provided.

a. 6P3 = $6 \cdot 5 \cdot 4 = 120$, alternately

$$6P3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 120 \tag{4.4.1.5}$$

b. 7P2 = $7 \cdot 6 = 42$, or

$$7P2 = \frac{7!}{5!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 42 \tag{4.4.1.6}$$

Next we consider some more permutation problems to get further insight into these concepts.

Example 4.4.1.5

In how many different ways can 4 people be seated in a straight line if two of them insist on sitting next to each other?

Solution

Let us suppose we have four people A, B, C, and D. Further suppose that A and B want to sit together. For the sake of argument, we tie A and B together and treat them as one person.

The four people are $\boxed{AB}$ CD. Since $\boxed{AB}$ is treated as one person, we have the following possible arrangements.

$$\boxed{AB}CD, \boxed{AB}DC, C\boxed{AB}D, D\boxed{AB}C, CD\boxed{AB}, DC\boxed{AB}$$

Note that there are six more such permutations because A and B could also be tied in the order BA. And they are

$$\boxed{BA}CD, \boxed{BA}DC, C\boxed{BA}D, D\boxed{BA}C, CD\boxed{BA}, DC\boxed{BA}$$

So altogether there are 12 different permutations.

Let us now do the problem using the multiplication axiom.

After we tie two of the people together and treat them as one person, we can say we have only three people. The multiplication axiom tells us that three people can be seated in 3! ways. Since two people can be tied together 2! ways, there are 3! 2! = 12 different arrangements

Example 4.4.1.6

You have 4 math books and 5 history books to put on a shelf that has 5 slots. In how many ways can the books be shelved if the first three slots are filled with math books and the next two slots are filled with history books?

Solution

We first do the problem using the multiplication axiom.

Since the math books go in the first three slots, there are 4 choices for the first slot,
3 choices for the second and 2 choices for the third.

The fourth slot requires a history book, and has five choices. Once that choice is made, there are 4 history books left, and therefore, 4 choices for the last slot. The choices are shown below.

4	3	2	5	4

Therefore, the number of permutations are $4 \cdot 3 \cdot 2 \cdot 5 \cdot 4 = 480$.

Alternately, we can see that $4 \cdot 3 \cdot 2$ is really same as 4P3, and $5 \cdot 4$ is 5P2.

So the answer can be written as (4P3) (5P2) = 480.

Clearly, this makes sense. For every permutation of three math books placed in the first three slots, there are 5P2 permutations of history books that can be placed in the last two slots. Hence the multiplication axiom applies, and we have the answer (4P3) (5P2).

We summarize the concepts of this section:

Note

1. Permutations

A permutation of a set of elements is an ordered arrangement where each element is used once.

2. Factorial

$$n! = n(n-1)(n-2)(n-3)\cdots 3\cdot 2\cdot 1$$

Where n is a natural number.

$$0! = 1$$

3. Permutations of n Objects Taken r at a Time

$$nPr = n(n-1)(n-2)(n-3)\cdots(n-r+1)$$

or

$$nPr = \frac{n!}{(n-r)!}$$

where n and r are natural numbers.

This page titled 4.4.1: Permutations is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

4.4.2: Permutations with Similar Elements

Learning Objectives

In this section you will learn to

1. Count the number of possible permutations when there are repeated items

In this section we will address the following problem.

1. In how many different ways can the letters of the word MISSISSIPPI be arranged?

This is an example of Permutations with Similar Elements.

Permutations with Similar Elements

Let us determine the number of distinguishable permutations of the letters ELEMENT.

Suppose we make all the letters different by labeling the letters as follows.

$$E_1LE_2ME_3NT$$

Since all the letters are now different, there are 7! different permutations.

Let us now look at one such permutation, say

$$LE_1ME_2NE_3T$$

Suppose we form new permutations from this arrangement by only moving the E's. Clearly, there are 3! or 6 such arrangements. We list them below.

$$\begin{array}{l} \mathrm{LE_1ME_2NE_3} \\ \mathrm{LE_1ME_3NE_2} \\ \mathrm{LE_2ME_1NE_3T} \\ \mathrm{LE_2ME_3NE_1T} \\ \mathrm{LE_3ME_2NE_1T} \\ \mathrm{LE_3ME_1NE_2T} \end{array} \tag{4.4.2.1}$$

Because the E's are not different, there is only one arrangement LEMENET and not six. This is true for every permutation.

Let us suppose there are n different permutations of the letters ELEMENT.

Then there are $n \cdot 3!$ permutations of the letters $E_1LE_2ME_3NT$.

But we know there are 7! permutations of the letters $E_1LE_2ME_3NT$.

Therefore, $n \cdot 3! = 7!$

Or $n = \frac{7!}{3!}$.

This gives us the method we are looking for.

Definition: Permutations with Similar Elements

The number of permutations of n elements taken n at a time, with r_1 elements of one kind, r_2 elements of another kind, and so on, is

$$\frac{n!}{r_1!r_2!\ldots r_k!} \tag{4.4.2.2}$$

Example 4.4.2.3

Find the number of different permutations of the letters of the word MISSISSIPPI.

Solution

The word MISSISSIPPI has 11 letters. If the letters were all different there would have been 11! different permutations. But MISSISSIPPI has 4 S's, 4 I's, and 2 P's that are alike.

So the answer is $\frac{11!}{4!4!2!} = 34,650$.

Example 4.4.2.4

If a coin is tossed six times, how many different outcomes consisting of 4 heads and 2 tails are there?

Solution

Again, we have permutations with similar elements.

We are looking for permutations for the letters HHHHTT.

The answer is $\frac{6!}{4!2!} = 15$.

Example 4.4.2.5

In how many different ways can 4 nickels, 3 dimes, and 2 quarters be arranged in a row?

Solution

Assuming that all nickels are similar, all dimes are similar, and all quarters are similar, we have permutations with similar elements. Therefore, the answer is

$$\frac{9!}{4!3!2!} = 1260$$

Example 4.4.2.6

A stock broker wants to assign 20 new clients equally to 4 of its salespeople. In how many different ways can this be done?

Solution

This means that each sales person gets 5 clients. The problem can be thought of as an ordered partitions problem. In that case, using the formula we get

$$\frac{20!}{5!5!5!5!} = 11,732,745,024$$

Example 4.4.2.7

A shopping mall has a straight row of 5 flagpoles at its main entrance plaza. It has 3 identical green flags and 2 identical yellow flags. How many distinct arrangements of flags on the flagpoles are possible?

Solution

The problem can be thought of as distinct permutations of the letters GGGYY; that is arrangements of 5 letters, where 3 letters are similar, and the remaining 2 letters are similar:

$$\frac{5!}{3!2!} = 10$$

Just to provide a little more insight into the solution, we list all 10 distinct permutations:

GGGYY, GGYGY, GGYYG, GYGGY, GYGYG, GYYGG, YGGGY, YGGYG, YGYGY, YYGGG

We summarize.

Summary

Permutations with Similar Elements

The number of permutations of n elements taken n at a time, with r_1 elements of one kind, r_2 elements of another kind, and so on, such that $\mathbf{n = r_1 + r_2 + ... + r_k}$ is

$$\frac{n!}{r_1!r_2!...r_k!}$$

This is also referred to as **ordered partitions**.

This page titled 4.4.2: Permutations with Similar Elements is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

4.4.3: Combinations

Learning Objectives

In this section you will learn to

1. Count the number of combinations of `r` out of `n` items (selections without regard to arrangement)
2. Use factorials to perform calculations involving combinations

Suppose we have a set of three letters { A, B, C }, and we are asked to make two-letter word sequences. We have the following six permutations.

AB BA BC CB AC CA

Now suppose we have a group of three people { A, B, C } as Al, Bob, and Chris, respectively, and we are asked to form committees of two people each. This time we have only three committees, namely,

AB BC AC

When forming committees, the order is not important, because the committee that has Al and Bob is no different than the committee that has Bob and Al. As a result, we have only three committees and not six.

Forming word sequences is an example of permutations, while forming committees is an example of **combinations** - the topic of this section.

Permutations are those arrangements where order is important, while combinations are those arrangements where order is not significant. From now on, this is how we will tell permutations and combinations apart.

In the above example, there were six permutations, but only three combinations.

Just as the symbol nPr represents the number of permutations of n objects taken r at a time, nCr represents the number of combinations of n objects taken r at a time.

So in the above example, 3P2 = 6, and 3C2 = 3.

Our next goal is to determine the relationship between the number of combinations and the number of permutations in a given situation.

In the above example, if we knew that there were three combinations, we could have found the number of permutations by multiplying this number by 2!. That is because each combination consists of two letters, and that makes 2! permutations.

Example 4.4.3.1

Given the set of letters { A, B, C, D }. Write the number of combinations of three letters, and then from these combinations determine the number of permutations.

Solution

We have the following four combinations.

ABC BCD CDA BDA

Since every combination has three letters, there are 3! permutations for every combination. We list them below.

ABC	BCD	CDA	BDA
ACB	BDC	CAD	BAD
BAC	CDB	DAC	DAB
BCA	CBD	DCA	DBA
CAB	DCB	ACD	ADB
CBA	DBC	ADC	ABD

The number of permutations are 3! times the number of combinations; that is

4P3 = 3! · 4C3

or

$$4C3 = \frac{4P3}{3!}$$

In general,

$$nCr = \frac{nPr}{r!}$$

Since

$$nPr = \frac{n!}{(n-r)!}$$

We have,

$$nCr = \frac{n!}{(n-r)!r!}$$

Summarizing,

Note

1. Combinations

A combination of a set of elements is an arrangement where each element is used once, and order is not important.

2. The Number of Combinations of n Objects Taken r at a Time

$$nCr = \frac{n!}{(n-r)!r!}$$

where n and r are natural numbers.

Example 4.4.3.2

Compute:

a. 5C3
b. 7C3

Solution

We use the above formula.

$$5C3 = \frac{5!}{(5-3)!3!} = \frac{5!}{2!3!} = 10$$

$$7C3 = \frac{7!}{(7-3)!3!} = \frac{7!}{4!3!} = 35$$

Example 4.4.3.3

In how many different ways can a student select to answer five questions from a test that has seven questions, if the order of the selection is not important?

Solution

Since the order is not important, it is a combination problem, and the answer is

7C5 = 21

Example 4.4.3.4

How many line segments can be drawn by connecting any two of the six points that lie on the circumference of a circle?

Solution

Since the line that goes from point A to point B is same as the one that goes from B to A, this is a combination problem.

It is a combination of 6 objects taken 2 at a time. Therefore, the answer is

$$6C2 = \frac{6!}{4!2!} = 15$$

Example 4.4.3.5

There are ten people at a party. If they all shake hands, how many hand-shakes are possible?

Solution

Note that between any two people there is only one hand shake. Therefore, we have

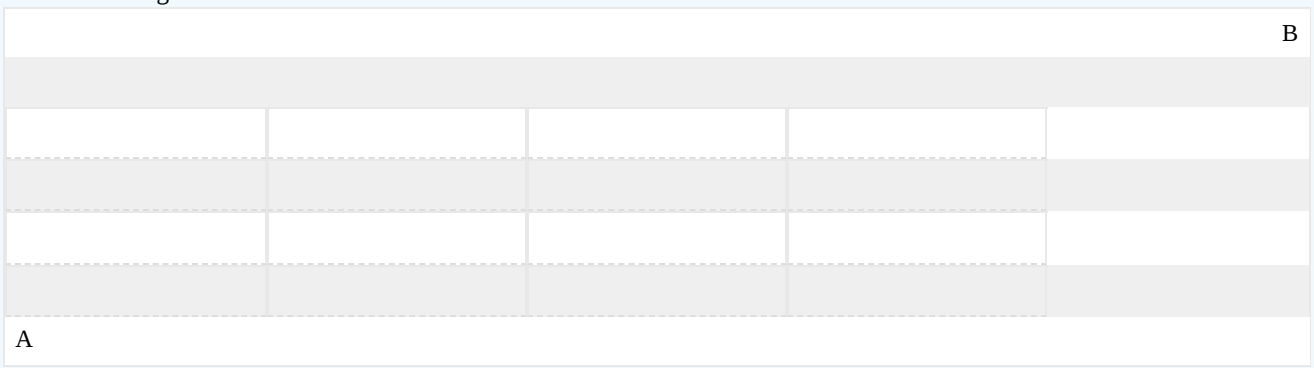
10C2 = 45 hand-shakes.

Example 4.4.3.6

The shopping area of a town is in the shape of square that is 5 blocks by 5 blocks. How many different routes can a taxi driver take to go from one corner of the shopping area to the opposite cater-corner?

Solution

Let us suppose the taxi driver drives from the point A, the lower left hand corner, to the point B, the upper right hand corner as shown in the figure below.



To reach his destination, he has to travel ten blocks; five horizontal, and five vertical. So if out of the ten blocks he chooses any five horizontal, the other five will have to be the vertical blocks, and vice versa.

Therefore, all he has to do is to choose 5 out of ten to be the horizontal blocks

The answer is 10C5, or 252.

Alternately, the problem can be solved by permutations with similar elements.

The taxi driver's route consists of five horizontal and five vertical blocks. If we call a horizontal block H, and a vertical block a V, then one possible route may be as follows.

HHHHHVVVVV

Clearly there are $\frac{10!}{5!5!} = 252$ permutations.

Further note that by definition 10C5 = $\frac{10!}{5!5!}$.

Example 4.4.3.7

If a coin is tossed six times, in how many ways can it fall four heads and two tails?

Solution

First we solve this problem using section 6.5 technique-permutations with similar elements.

We need 4 heads and 2 tails, that is

HHHHTT

There are $\frac{6!}{4!2!} = 15$ permutations.

Now we solve this problem using combinations.

Suppose we have six spots to put the coins on. If we choose any four spots for heads, the other two will automatically be tails. So the problem is simply

6C4 = 15.

Incidentally, we could have easily chosen the two tails, instead. In that case, we would have gotten

6C2 = 15.

Further observe that by definition

$$6C4 = \frac{6!}{2!4!}$$

and

$$6C2 = \frac{6!}{4!2!}$$

Which implies 6C4 = 6C2.

This page titled 4.4.3: Combinations is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.5: Combinations** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

4.5: Probability And Counting Rules

Learning Objectives

In this section, you will learn to:

1. Use probability tree diagrams to calculate probabilities
2. Use combinations to calculate probabilities

In this section, we will apply previously learnt counting techniques in calculating probabilities, and use tree diagrams to help us gain a better understanding of what is involved.

USING TREE DIAGRAMS TO CALCULATE PROBABILITIES

We already used tree diagrams to list events in a sample space. Tree diagrams can be helpful in organizing information in probability problems; they help provide a structure for understanding probability. In this section we expand our previous use of tree diagrams to situations in which the events in the sample space are not all equally likely.

We assign the appropriate probabilities to the events shown on the branches of the tree.
By multiplying probabilities along a path through the tree, we can find probabilities for “and” events, which are intersections of events.

We begin with an example.

Example 4.5.1

Suppose a jar contains 3 red and 4 white marbles. If two marbles are drawn with replacement, what is the probability that both marbles are red?

Solution

Let $\mathbf{E}$ be the event that the first marble drawn is red, and let $\mathbf{F}$ be the event that the second marble drawn is red.

We need to find $\mathbf{P(E \cap F)}$.

By the statement, "two marbles are drawn with replacement," we mean that the first marble is replaced before the second marble is drawn.

There are 7 choices for the first draw. And since the first marble is replaced before the second is drawn, there are, again, seven choices for the second draw. Using the multiplication axiom, we conclude that the sample space $\mathbf{S}$ consists of 49 ordered pairs. Of the 49 ordered pairs, there are $\mathbf{3 \times 3 = 9}$ ordered pairs that show red on the first draw and, also, red on the second draw. Therefore,

$$P(E \cap F) = \frac{9}{49}$$

Further note that in this particular case

$$P(E \cap F) = \frac{9}{49} = \frac{3}{7} \cdot \frac{3}{7}$$

giving us the result that in this example: $\mathbf{P(E \cap F) = P(E) \cdot P(F)}$

Example 4.5.2

If in Example **4.5.1**, the two marbles are drawn without replacement, then what is the probability that both marbles are red?

Solution

By the statement, "two marbles are drawn without replacement," we mean that the first marble is not replaced before the second marble is drawn.

Again, we need to find $\mathbf{P(E \cap F)}$.

There are, again, 7 choices for the first draw. And since the first marble is not replaced before the second is drawn, there are only six choices for the second draw. Using the multiplication axiom, we conclude that the sample space S consists of 42 ordered pairs. Of the 42 ordered pairs, there are $\mathbf{3 \times 2 = 6}$ ordered pairs that show red on the first draw and red on the second draw. Therefore,

$$P(E \cap F) = \frac{6}{42}$$

Note that we can break this calculation down as

$$P(E \cap F) = \frac{6}{42} = \frac{3}{7} \cdot \frac{2}{6}$$

.

Here 3/7 represents $\mathbf{P(E)}$, and 2/6 represents the probability of drawing a red on the second draw, given that the first draw resulted in a red.

We write the latter as $\mathbf{P}$(red on the second | red on first) or $\mathbf{P(F|E)}$. The "|" represents the word "given" or “if”. This leads to the result that:

$$\mathbf{P(E \cap F) = P(E) \cdot P(F|E)}$$

The is an important result, called the **Multiplication Rule,** which will appear again in later sections.

We now demonstrate the above results with a tree diagram.

Example 4.5.3

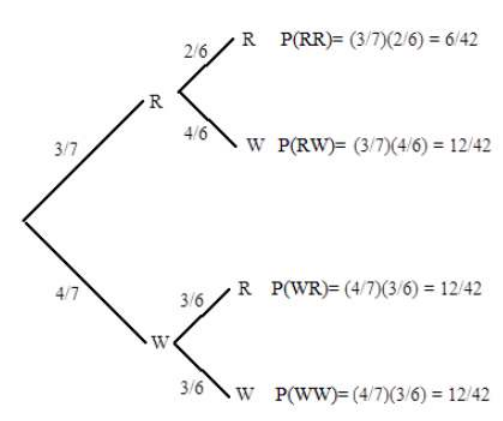
Suppose a jar contains 3 red and 4 white marbles. If two marbles are drawn without replacement, find the following probabilities using a tree diagram.

a. The probability that both marbles are red.
b. The probability that the first marble is red and the second white.
c. The probability that one marble is red and the other white.

Solution

Let $\mathbf{R}$ be the event that the marble drawn is red, and let W be the event that the marble drawn is white.

We draw the following tree diagram.



a. The probability that both marbles are red is $\mathbf{P(RR) = 6/42}$
b. The probability that the first marble is red and the second is white is $\mathbf{P(RW) = 12/42}$

c. For the probability that one marble is red and the other is white, we observe that this can be satisfied if the first is red and the second is white, **or** if the first is white and the second is red. The "or" tells us we'll be using the Addition Rule from Section 7.2.

Furthermore events **RW** and **WR** are mutually exclusive events, so we use the form of the Addition Rule that applies to mutually exclusive events.

Therefore

P(one marble is red and the other marble is white)

$$\begin{aligned} &= \mathrm{P(RW \text{ or } WR)} \\ &= \mathrm{P(RW)} + \mathrm{P(WR)} \\ &= 12/42 + 12/42 = 24/42 \end{aligned}$$

USING COMBINATIONS TO FIND PROBABILITIES

Although the tree diagrams give us better insight into a problem, they are not practical for problems where more than two or three things are chosen. In such cases, we use the concept of combinations that we learned in the last chapter. This method is best suited for problems where the order in which the objects are chosen is not important, and the objects are chosen without replacement.

Example 4.5.4

Suppose a jar contains 3 red, 2 white, and 3 blue marbles. If three marbles are drawn without replacement, find the following probabilities.

a. **P**(Two red and one white)
b. **P**(One of each color)
c. **P**(None blue)
d. **P**(At least one blue)

Solution

Let us suppose the marbles are labeled as $R_1, R_2, R_3, W_1, W_2, B_1, B_2, B_3$.

a. **P**(Two red and one white)

Since we are choosing 3 marbles from a total of 8, there are 8C3 = 56 possible combinations. Of these 56 combinations, there are $\mathrm{3C2 \times 2C1 = 6}$ combinations consisting of 2 red and one white. Therefore,

$$P(\text{ Two red and one white }) = \frac{\mathrm{3C2 \times 2C1}}{\mathrm{8C3}} = \frac{6}{56}.$$

b. **P**(One of each color)

Again, there are 8C3 = 56 possible combinations. Of these 56 combinations, there are $\mathrm{3C1 \times 2C1 \times 3C1 = 18}$ combinations consisting of one red, one white, and one blue. Therefore,

$$P(\text{ One of each color }) = \frac{\mathrm{3C1 \times 2C1 \times 3C1}}{\mathrm{8C3}} = \frac{18}{56}$$

c. **P**(None blue)

There are 5 non-blue marbles, therefore

$$\mathrm{P(\text{ None blue }) = \frac{5C3}{8C3} = \frac{10}{56} = \frac{5}{28}}$$

d. **P**(At least one blue)

By "at least one blue marble," we mean the following: one blue marble and two non-blue marbles, *OR* two blue marbles and one non-blue marble, *OR* all three blue marbles. So we have to find the sum of the probabilities of all three cases.

$$\mathrm{P(At \text{ least one blue }) = P(1 \text{ blue, } 2 \text{ non-blue}) + P(2 \text{ blue, } 1 \text{ non-blue}) + P(3 \text{ blue})}$$

$$P(\text{ At least one blue }) = \frac{3C1 \times 5C2}{8C3} + \frac{3C2 \times 5C1}{8C3} + \frac{3C3}{8C3}$$

$$P(\text{ At least one blue }) = 30/56 + 15/56 + 1/56 = 46/56 = 23/28$$

Alternately, we can use the fact that $\mathbf{P(E) = 1 - P(E^c)}$. If the event $\mathbf{E}$ = At least one blue, then $\boldsymbol{E^c}$ = None blue.

But from part c of this example, we have $\mathbf{(E^c) = 5/28}$, so $\mathbf{P(E) = 1 - 5/28 = 23/28}$.

Example 4.5.5

Five cards are drawn from a deck. Find the probability of obtaining two pairs, that is, two cards of one value, two of another value, and one other card.

Solution

Let us first do an easier problem-the probability of obtaining a pair of kings and queens.

Since there are four kings, and four queens in the deck, the probability of obtaining two kings, two queens and one other card is

$$\mathrm{P}(\text{A pair of kings and queens }) = \frac{4C2 \times 4C2 \times 44C1}{52C5}$$

To find the probability of obtaining two pairs, we have to consider all possible pairs.

Since there are altogether 13 values, that is, aces, deuces, and so on, there are 13C2 different combinations of pairs.

$$P(\text{ Two pairs }) = 13C2 \cdot \frac{4C2 \times 4C2 \times 44C1}{52C5} = .04754$$

Example 4.5.6

A cell phone store receives a shipment of 15 cell phones that contains 8 iPhones and 7 Android phones. Suppose that 6 cell phones are randomly selected from this shipment. Find the probability that a randomly selected set of 6 cell phones consists of 2 iPhones and 4 Android phones.

Solution

There are 8C2 ways of selecting 2 out of the 8 iPhones.

and 7C4 ways of selecting 4 out of the 7 Android phones

But altogether there are 15C6 ways of selecting 6 out of 15 cell phones.

Therefore we have

$$P(\text{2 iPhones and 4 Android phones }) = \frac{8C2 \times 7C4}{15C6} = \frac{(28)(35)}{5005} = \frac{980}{5005} = 0.1958$$

Example 4.5.7

One afternoon, a bagel store still has 53 bagels remaining: 20 plain, 15 poppyseed, and 18 sesame seed bagels. Suppose that the store owner packages up a bag of 9 bagels to bring home for tomorrow's breakfast, and selects the bagels randomly. Find the probability that the bag contains 4 plain, 3 poppyseed, and 2 sesame seed.

Solution

There are 20C4 ways of selecting 4 out of the 20 plain bagels,

and 15C3 ways of selecting 3 out of the 15 poppyseed bagels,

and 18C2 ways of selecting 2 out of the 18 sesame seed bagels.

But altogether there are 53C9 ways of selecting 9 out of the 53 bagels.

$$\begin{aligned} \mathbf{P}(\text{4 plain, 3 poppyseed, and 2 sesame seed}) &= \frac{{}_{20}C_4 \times {}_{15}C_3 \times {}_{18}C_2}{{}_{53}C_9} \\ &= \frac{(4845)(455)(153)}{4431613550} \\ &= 0.761 \end{aligned} \tag{4.5.1}$$

We end the section by solving a famous problem called the **Birthday Problem**.

Example 4.5.8: Birthday Problem

If there are 25 people in a room, what is the probability that at least two people have the same birthday?

Solution

Let event **E** represent that at least two people have the same birthday.

We first find the probability that no two people have the same birthday.

We analyze as follows.

Suppose there are 365 days to every year. According to the multiplication axiom, there are 365^{25} possible birthdays for 25 people. Therefore, the sample space has 365^{25} elements. We are interested in the probability that no two people have the same birthday. There are 365 possible choices for the first person and since the second person must have a different birthday, there are 364 choices for the second, 363 for the third, and so on. Therefore,

$$\mathbf{P}(\text{No two have the same birthday }) = \frac{365 \cdot 364 \cdot 363 \cdots 341}{365^{25}} = \frac{{}_{365}P_{25}}{365^{25}}$$

Since **P**(at least two people have the same birthday) = 1 - **P**(No two have the same birthday),

$$\mathbf{P}\text{ at least two people have the same birthday }) = 1 - \frac{{}_{365}P_{25}}{365^{25}} = .5687$$

This page titled 4.5: Probability And Counting Rules is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **8.3: Probability Using Tree Diagrams and Combinations** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

4.E: Probability Topics (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

3.1: Introduction

3.2: Terminology

Q 3.2.1

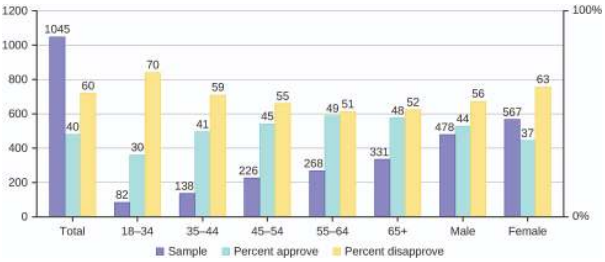


Figure 3.2.3.2.11.

The graph in Figure 3.2.1 displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Q 3.2.2

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

S 3.2.2

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

3.3: Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

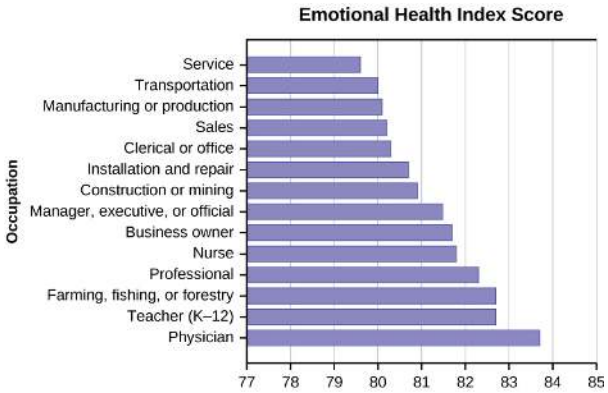


Figure 3.3.1.

Q 3.3.1

Find the probability that an Emotional Health Index Score is 82.7.

Q 3.3.2

Find the probability that an Emotional Health Index Score is 81.0.

S 3.3.2

0

Q 3.3.3

Find the probability that an Emotional Health Index Score is more than 81?

Q 3.3.4

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

S 3.3.4

0.3571

Q 3.3.5

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Q 3.3.6

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

S 3.3.6

0.2142

Q 3.3.7

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Q 3.3.8

What occupation has the highest emotional index score?

S 3.3.8

Physician (83.7)

Q 3.3.9

What occupation has the lowest emotional index score?

Q 3.3.10

What is the range of the data?

S 3.3.10

83.7 − 79.6 = 4.1

Q 3.3.11

Compute the average EHIS.

Q 3.3.12

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

S 3.3.12

$P(\text{Occupation} < 81.3) = 0.5$

3.4: Two Basic Rules of Probability

Q 3.4.1

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- $\mathbf{C}$ = California registered voters who support same-sex marriage.
- $\mathbf{B}$ = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- $\mathbf{A}$ = California registered voters who are 18 to 39 years old.

a. Find $P(\mathbf{C})$.
b. Find $P(\mathbf{B})$.
c. Find $P(\mathbf{C}|\mathbf{A})$.
d. Find $P(\mathbf{B}|\mathbf{C})$.
e. In words, what is $\mathbf{C}|\mathbf{A}$?
f. In words, what is $\mathbf{B}|\mathbf{C}$?
g. Find $P(\mathbf{C}\ \mathbf{AND}\ \mathbf{B})$.
h. In words, what is $\mathbf{C}\ \mathbf{AND}\ \mathbf{B}$?
i. Find $P(\mathbf{C}\ \mathbf{OR}\ \mathbf{B})$.
j. Are $\mathbf{C}$ and $\mathbf{B}$ mutually exclusive events? Show why or why not.

Q 3.4.2

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?

c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

S 3.4.2

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.

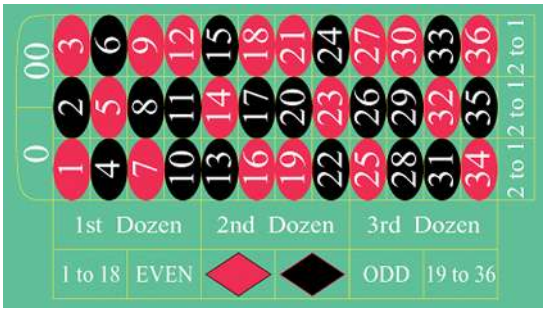


Figure 3.4.1

Q 3.4.3

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Q 3.4.4

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

S 3.4.4

1. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$
2. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
3. $P(\text{Betting on one number}) = \frac{1}{38}$
4. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$
5. $P(\text{Betting on two number that touch each other on the table}) = \frac{2}{38}$
6. $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$

https://stats.libretexts.org/@go/page/10897

7. $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

Q 3.4.5

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Q 3.4.6

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- $\mathbf{G}$ = card drawn is green
- $\mathbf{E}$ = card drawn is even-numbered

a. List the sample space.
b. $P(\mathbf{G}) =$ _____
c. $P(\mathbf{G}|\mathbf{E}) =$ _____
d. $P(\mathbf{G} \text{ AND } \mathbf{E}) =$ _____
e. $P(\mathbf{G} \text{ OR } \mathbf{E}) =$ _____
f. Are $\mathbf{G}$ and $\mathbf{E}$ mutually exclusive? Justify your answer numerically.

S 3.4.6

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{}{6}8$
6. No, because $P(\mathbf{G} \text{ AND } \mathbf{E})$ does not equal 0.

Q 3.4.7

Roll two fair dice. Each die has six faces.

a. List the sample space.
b. Let $\mathbf{A}$ be the event that either a three or four is rolled first, followed by an even number. Find $P(\mathbf{A})$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(\mathbf{B})$.
d. In words, explain what "$P(\mathbf{A}|\mathbf{B})$" represents. Find $P(\mathbf{A}|\mathbf{B})$.
e. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are $\mathbf{A}$ and $\mathbf{B}$ independent events? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.8

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let $\mathbf{A}$ be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(\mathbf{A})$.
c. Let $\mathbf{B}$ be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

S 3.4.9

The coin toss is independent of the card picked first.

a. $\{(G,H)(G,T)(B,H)(B,T)(R,H)(R,T)\}$
b. $P(\text{A}) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(\text{A AND B}) = 0$
d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(\text{A AND C}) = P(\text{A}) = 320$

Q 3.4.10

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(\text{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.11

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(\text{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

S 3.4.12

a. $S = (\text{HHH}), (\text{HHT}), (\text{HTH}), (\text{HTT}), (\text{THH}), (\text{THT}), (\text{TTH}), (\text{TTT})$
b. $\frac{4}{8}$
c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(\text{A AND B}) = 0$.

Q 3.4.13

Consider the following scenario:

Let $P(\text{C}) = 0.4$.

Let $P(\text{D}) = 0.5$.

Let $P(\text{C|D}) = 0.6$.

a. Find $P(CANDD)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(\text{C OR D})$.
e. Find $P(\text{D|C})$.

Q 3.4.14

Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(\text{Y OR Z}) = P(\text{Y}) + P(\text{Z}) - P(\text{Y AND Z})$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(\text{Z})$ if $P(\text{Y OR Z}) = 0.71$ and $P(\text{Y}) = 0.42$.

S 3.4.14

a. If Y and Z are independent, then $P(\mathrm{Y\ AND\ Z}) = P(\mathrm{Y})P(\mathrm{Z})$, so $P(\mathrm{Y\ OR\ Z}) = P(\mathrm{Y}) + P(\mathrm{Z}) - P(\mathrm{Y})P(\mathrm{Z})$.
b. 0.5

Q 3.4.15

G and H are mutually exclusive events. $P(\mathrm{G}) = 0.5 P(\mathrm{H}) = 0.3$

a. Explain why the following statement MUST be false: $P(\mathrm{H}|\mathrm{G}) = 0.4$.
b. Find $P(\mathrm{H\ OR\ G})$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Q 3.4.16

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E' = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(\mathrm{E}') =$	i. 0.8043
b. $P(\mathrm{E}) =$	ii. 0.623
c. $P(\mathrm{S\ and\ E}') =$	iii. 0.1957
d. $P(\mathrm{S}|\mathrm{E}') =$	iv. 0.1219

S 3.4.16

a. iii
b. i
c. iv
d. ii

Q 3.4.17

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Q 3.4.18

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.

c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

S 3.4.18

a. $P(\text{R}) = 0.44$
b. $P(\text{R}|\text{E}) = 0.56$
c. $P(\text{R}|\text{O}) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(\text{R}|\text{E}) \neq P(\text{R})$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(\text{R}|\text{E}) > P(\text{R})$.

Q 3.4.19

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) = P(\text{hit is a double})$

Q 3.4.29

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

S 3.4.30

a. $P(\text{type O OR Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O AND Rh-})$

$0.52 = 0.43 + 0.15 - P(\text{type O AND Rh-})$; solve to find $P(\text{type O AND Rh-}) = 0.06$

6% of people have type O, Rh- blood

b. $P(\text{NOT(type O AND Rh-)}) = 1 - P(\text{type O AND Rh-}) = 1 - 0.06 = 0.94$

94% of people do not have type O, Rh- blood

Q 3.4.31

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

Q 3.4.32

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

S 3.4.32

a. Let $C =$ be the event that the cookie contains chocolate. Let $N =$ the event that the cookie contains nuts.
b. $P(\text{C OR N}) = P(\text{C}) + P(\text{N}) - P(\text{C AND N}) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(\text{C OR N}) = 1 - 0.40 = 0.60$

Q 3.4.33

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let $\text{D} =$ event that a student takes a distance learning class and $\text{E} =$ event that a student is a part time student

a. Find $P(\text{D AND E})$.
b. Find $P(\text{E}|\text{D})$.
c. Find $P(\text{D OR E})$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.5: Contingency Tables

Use the information in the Table to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Q 3.5.1

What is the probability that a randomly selected senator has an “Other” affiliation?

S 3.5.1

0

Q 3.5.2

What is the probability that a randomly selected senator is up for reelection in November 2016?

Q 3.5.3

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

S 3.5.3

$\frac{10}{67}$

Q 3.5.4

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Q 3.5.5

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

S 3.5.5

$\frac{10}{34}$

Q 3.5.6

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

Q 3.5.7

The events “Republican” and “Up for reelection in 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

S 3.5.7

d

Q 3.5.8

The events “Other” and “Up for reelection in November 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Q 3.5.9

This table gives the number of participants in the recent National Health Interview Survey who had been treated for cancer in the previous 12 months. The results are sorted by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex.

Race and Sex	15-24	25-40	41-65	over 65	TOTALS
white, male	1,165	2,036	3,703		8,395
white, female	1,076	2,242	4,060		9,129
black, male	142	194	384		824
black, female	131	290	486		1,061
all others					
TOTALS	2,792	5,279	9,354		21,081

Do not include "all others" for parts f and g.

a. Fill in the column for cancer treatment for individuals over age 65.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was a black or white male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

S 3.5.9

a.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c. $\frac{22,050}{29,760}$
d. $\frac{330}{29,760}$
e. $\frac{2,000}{29,760}$
f. $\frac{23,720}{29,760}$
g. $\frac{5,010}{6,020}$

Use the following information to answer the next two exercises. The table of data obtained from*www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Q 3.5.10

Find $P(\text{hit was made by Babe Ruth})$.

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Q 3.5.11

Find $P(\text{hit was made by Ty Cobb}|\text{The hit was a Home Run})$.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

S 3.5.11

b

Q 3.5.12

Table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If **B** is the event of a child having brown hair, find the probability of the complement of **B**.
g. In words, what does the complement of **B** represent?

Q 3.5.13

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data were compiled into the following table.

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

S 3.5.13

a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

3.6: Tree and Venn Diagrams

Exercise 3.6.8

The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: $\mathrm{C}=$ a man develops cancer in his lifetime; $\mathrm{P}=$ man has at least one false positive. Construct a tree diagram of the situation.

Answer

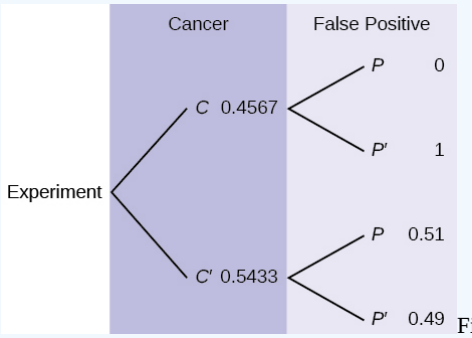


Figure 3.6.12

Bring It Together

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

Exercise 3.6.9

Suppose that you randomly draw two cards, one at a time, **with replacement**.

Let $\mathrm{G_1}=$ first card is green
Let $\mathrm{G_2}=$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(\mathrm{G_1 AND G_2})$.

c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_1 and G_2 independent events? Explain why or why not.

Answer

a.

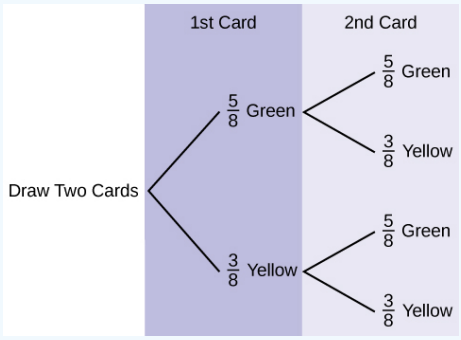


Figure 3.6.14

b. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
c. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
d. $P(G|G) = \frac{5}{8}$
e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercise 3.6.10

Suppose that you randomly draw two cards, one at a time, **without replacement**.

G_1 = first card is green
G_2 = second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{AND } G_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Exercise 3.6.11

Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find $P(\text{driver is female})$.
c. Find $P(\text{driver is age 65 or over}|\text{driver is female})$.
d. Find $P(\text{driver is age 65 or over AND female})$.
e. In words, explain the difference between the probabilities in part c and part d.
f. Find $P(\text{driver is age 65 or over})$.
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Answer

https://stats.libretexts.org/@go/page/10897

a.

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

b. $P(\text{F}) = 0.486$
c. $P(>64|\text{F}) = 0.1361$
d. $P(>64 \text{ and } \text{F}) = P(\text{F})P(>64|\text{F}) = (0.486)(0.1361) = 0.0661$
e. $P(>64|\text{F})$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } \text{F})$ is the percentage of drivers who are female and 65 or older.
f. $P(>64) = P(>64 \text{ and } \text{F}) + P(>64 \text{ and } \text{M}) = 0.1356$
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } \text{F}) = 0.0661$.

Exercise 3.6.12

Suppose that 10,000 U.S. licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercise 3.6.13

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Answer

a.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$.
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise 3.6.14

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(\mathrm{H})$ and $P(\mathrm{T})$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise 3.6.15

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person has a risk factor heterosexual contact})$.
c. Find $P(\text{Person is female OR has a risk factor of IV drug user})$.
d. Find $P(\text{Person is female AND has a risk factor of homosexual/bisexual})$.
e. Find $P(\text{Person is male AND has a risk factor of IV drug user})$.
f. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$.
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Answer

The completed contingency table is as follows:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

a. $\frac{255}{2059}$
b. $\frac{196}{3059}$
c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$

g.

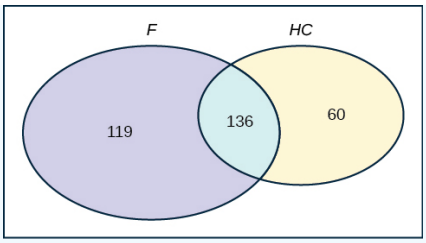


Figure 3.6.15

Exercise 3.6.16

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find $P(\textbf{Person is female})$.
b. Find $P(\textbf{Person obtained the disease through heterosexual contact})$.
c. Find $P(\textbf{Person is female GIVEN person got the disease from heterosexual contact})$
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (**R**), four yellow (**Y**) and five blue (**B**) beads. For the coin, $P(\textbf{H}) = \frac{2}{3}$ and $P(\textbf{T}) = \frac{1}{3}$ where **H** is heads and **T** is tails.

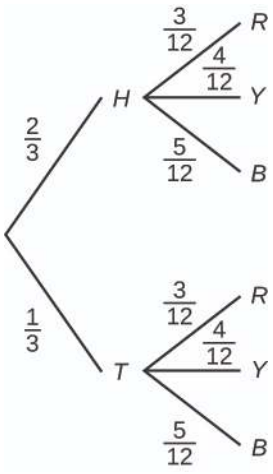


Figure 3.6.1.

Q 3.6.1

Find $P(\textbf{tossing a Head on the coin AND a Red bead})$

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Q 3.6.2

Find $P(\textbf{Blue bead})$.

a. $\frac{15}{36}$
b. $\frac{10}{36}$

https://stats.libretexts.org/@go/page/10897

c. $\frac{10}{12}$
d. $\frac{6}{36}$

S 3.6.2

a

Q 3.6.3

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let $\mathbf{S}$ be the event that both cookies selected were the same flavor. Find $P(\mathbf{S})$.
e. Let $\mathbf{T}$ be the event that the cookies selected were different flavors. Find $P(\mathbf{T})$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let $\mathbf{U}$ be the event that the second cookie selected is a butter cookie. Find $P(\mathbf{U})$.

3.7: Probability Topics

This page titled 4.E: Probability Topics (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **3.E: Probability Topics (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

1.E: Combinations (Optional Exercises)

Do the following problems using combinations.

1. How many different 3-people committees can be chosen from ten people?	2. How many different 5-player teams can be chosen from eight players?
3. In how many ways can a person chose to vote for three out of five candidates on a ballot for a school board election?	4. Compute the following: a. 9C2 b. 6C4 c. 8C3 d. 7C4
5. How many 5-card hands can be chosen from a deck of cards?	6. How many 13-card bridge hands can be chosen from a deck of cards?
7. There are twelve people at a party. If they all shake hands, how many different hand-shakes are there?	8. In how many ways can a student choose to do four questions out of five on a test?
9. Five points lie on a circle. How many chords can be drawn through them?	10. How many diagonals does a hexagon have?
11. There are five team in a league. How many games are played if every team plays each other twice?	12. A team plays 15 games a season. In how many ways can it have 8 wins and 7 losses?
13. In how many different ways can a 4-child family have 2 boys and 2 girls?	14. A coin is tossed five times. In how many ways can it fall three heads and two tails?
15. The shopping area of a town is a square that is six blocks by six blocks. How many different routes can a taxi driver take to go from one corner of the shopping area to the opposite cater-corner?	16. If the shopping area in the previous problem has a rectangular form of 5 blocks by 3 blocks, then how many different routes can a taxi driver take to drive from one end of the shopping area to the opposite kitty corner end?
17. A team of 7 workers is assigned to a project. In how many ways can 3 of the 7 workers be selected to make a presentation to the management about their progress on the project?	18. A real estate company has 12 houses listed for sale by their clients. In how many ways can 5 of the 12 houses be selected to be featured in an advertising brochures?
19. A frozen yogurt store has 9 toppings to choose from. In how many ways can 3 of the 9 toppings be selected ?	20. A kindergarten teacher has 14 books about a holiday. In how many ways can she select 4 of the books to read to her class in the week before the holiday?

This page titled 1.E: Combinations (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.5.1: Combinations (Exercises)** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

1.E: Permutations (Optional Exercises)

Do the following problems using permutations.

1. How many three-letter words can be made using the letters { a, b, c, d, e } if no repetitions are allowed?	2. A grocery store has five checkout counters, and seven clerks. How many different ways can the 7 clerks be assigned to the 5 counters?
3. A group of fifteen people who are members of an investment club wish to choose a president, and a secretary. How many different ways can this be done?	4. Compute the following. a. 9P2 b. 6P4 c. 8P3 d. 7P4
5. In how many ways can the letters of the word CUPERTINO be arranged if each letter is used only once in each arrangement?	6. How many permutations of the letters of the word PROBLEM end in a vowel?
7. How many permutations of the letters of the word SECURITY end in a consonant?	8. How many permutations of the letters PRODUCT have consonants in the second and third positions?
9. How many three-digit numbers are there?	10. How many three-digit odd numbers are there?
11. In how many different ways can five people be seated in a row if two of them insist on sitting next to each other?	12. In how many different ways can five people be seated in a row if two of them insist on not sitting next to each other?
13. In how many ways can 3 English, 3 history, and 2 math books be set on a shelf, if the English books are set on the left, history books in the middle, and math books on the right?	14. In how many ways can 3 English, 3 history, and 2 math books be set on a shelf, if they are grouped by subject?
15. You have 5 math books and 6 history books to put on a shelf with five slots. In how many ways can you put the books on the shelf if the first two slots are to be filled with math books and the next three with history books?	16. You have 5 math books and 6 history books to put on a shelf with five slots. In how many ways can you put the books on the shelf if the first two slots are to be filled with the books of one subject and the next three slots are to be filled with the books of the other subject?
17. A bakery has 9 different fancy cakes. In how many ways can 5 of the 9 fancy cakes be lined up in a row in the bakery display case?	18. A landscaper has 6 different flowering plants. She needs to plant 4 of them in a row in a garden. How many different ways can 4 of the 6 plants be arranged in a row?
19. At an auction of used construction vehicles, there are 7 different vehicles for sale. In how many orders could these 7 vehicles be listed in the auction program?	20. A landscaper has 6 different flowering plants and 4 different non-flowering bushes. She needs to plant a row of 6 plants in a garden. There must be a bush at each end, and four flowering plants in a row in between the bushes. How many different arrangements in a row are possible?
21. In how many ways can all 7 letters of the word QUIETLY be arranged if the letters Q and U must be next to each other in the order QU?	22. a. In how many ways can the letters ABCDEXY be arranged if the X and Y must be next to each other in either order XY or YX? b. In how many ways can the letters ABCDEXY be arranged if the X and Y can not be next to each other?

This page titled 1.E: Permutations (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.3.1: Permutations (Exercises)** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

1.E: Permutations with Similar Elements (Optional Exercises)

Do the following problems using the techniques learned in this section.

1. In how many different ways can five children hold hands to play "Ring Around the Rosy"?	2. In how many ways can three people be made to sit at a round table?
3. In how many different ways can six children ride a "Merry Go Around" with six horses?	4. In how many ways can three couples be seated at a round table, so that men and women sit alternately?
5. In how many ways can six trinkets be arranged on a chain?	6. In how many ways can five keys be put on a key ring?
7. Find the number of different permutations of the letters of the word MASSACHUSETTS.	8. Find the number of different permutations of the letters of the word MATHEMATICS.
9. Seven flags are to be flown on seven poles: 3 flags are red, 2 are white, and 2 are blue,. How many different arrangements are possible?	10. How many different ways can 3 pennies, 2 nickels and 5 dimes be arranged in a row?
11. How many four-digit numbers can be made using two 2's and two 3's?	12. How many five-digit numbers can be made using two 6's and three 7's?
13. If a coin is tossed 5 times, how many different outcomes of 3 heads and 2 tails are possible?	14. If a coin is tossed 10 times, how many different outcomes of 7 heads and 3 tails are possible?
15. If a team plays ten games, how many different outcomes of 6 wins and 4 losses are possible?	16. If a team plays ten games, how many different ways can the team have a winning season?

This page titled 1.E: Permutations with Similar Elements (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.4.1: Circular Permutations and Permutations with Similar Elements (Exercises)** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

1.E: Probability Using Tree Diagrams and Combinations (Optional Exercises)

SECTION 8.3 PROBLEM SET: PROBABILITIES USING TREE DIAGRAMS AND COMBINATIONS

Two apples are chosen from a basket containing five red and three yellow apples.
Draw a tree diagram below, and find the following probabilities.

1) P(both red)	2) P(one red, one yellow)
3) P(both yellow)	4) P(First red and second yellow)

A basket contains six red and four blue marbles. Three marbles are drawn at random.
Find the following probabilities using the method shown in Example 8.3.2. Do not use combinations.

5) P(All three red)	6) P(two red, one blue)
7) P(one red, two blue)	8) P(first red, second blue, third red)

Three marbles are drawn from a jar containing five red, four white, and three blue marbles.
Find the following probabilities using combinations.

9) P(all three red)	10) P(two white and 1 blue)
11) P(none white)	12) P(at least one red)

A committee of four is selected from a total of 4 freshmen, 5 sophomores, and 6 juniors. Find the probabilities for the following events.

13) At least three freshmen.	14) No sophomores.
15) All four of the same class.	16) Not all four from the same class.
17) Exactly three of the same class.	18) More juniors than freshmen and sophomores combined.

Five cards are drawn from a deck. Find the probabilities for the following events.

19) Two hearts, two spades, and one club.	20) A flush of any suit (*all cards of a single suit*).
21) A full house of nines and tens (*3 nines and 2 tens*).	22) Any full house.
23) A pair of nines and a pair of tens (*and the fifth card is not a nine or ten*).	24) Any two pairs (*two cards of one value, two more cards of another value, and the fifth card does not have the same value as either pair*).

Jorge has 6 rock songs, 7 rap songs and 4 country songs that he likes to listen to while he exercises.
He randomly selects six (6) of these songs to create a playlist to listen to today while he exercises.

Find the following probabilities:

25) P(playlist has 2 songs of each type)	26) P(playlist has no country songs)
27) P(playlist has 3 rock, 2 rap, and 1 country song)	28) P(playlist has 3 or 4 rock songs and the rest are rap songs)

A project is staffed 12 people: 5 engineers, 4 salespeople, and 3 customer service representatives.
A committee of 5 people is selected to make a presentation to senior management.

Find the probabilities of the following events.

29) The committee has 2 engineers, 2 salespeople, and 1 customer service representative.	30) The committee contains 3 engineer and 2 salespeople.
31) The committee has no engineers.	32) The committee has all salespeople.

Do the following birthday problems.

33) If there are 5 people in a room, what is the probability that no two have the same birthday?	34) If there are 5 people in a room, find the probability that at least 2 have the same birthday.

This page titled 1.E: Probability Using Tree Diagrams and Combinations (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **8.3.1: Probability Using Tree Diagrams and Combinations (Exercises)** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

1.E: Tree Diagrams and the Multiplication Axiom (Optional Exercises)

Do the following problems using a tree diagram or the multiplication axiom.

1. A man has 3 shirts, and 2 pairs of pants. Use a tree diagram to determine the number of possible outfits.	2. In a city election, there are 2 candidates for mayor, and 3 for supervisor. Use a tree diagram to find the number of ways to fill the two offices.
3. There are 4 roads from Town A to Town B, 2 roads from Town B to Town C. Use a tree diagram to find the number of ways one can travel from Town A to Town C.	4. Brown Home Construction offers a selection of 3 floor plans, 2 roof types, and 2 exterior wall types. Use a tree diagram to determine the number of possible homes available.
5. For lunch, a small restaurant offers 2 types of soups, three kinds of sandwiches, and two types of soft drinks. Use a tree diagram to determine the number of possible meals consisting of a soup, sandwich, and a soft drink.	6. A California license plate consists of a number from 1 to 5, then three letters followed by three digits. How many such plates are possible?

Do the following problems using the Multiplication Axiom

7. A license plate consists of three letters followed by three digits. How many license plates are possible if no letter may be repeated?	8. How many different 4-letter radio station call letters can be made if the first letter must be K or W and no letters can be repeated?
9. How many seven-digit telephone numbers are possible if the first two digits cannot be ones or zeros?	10. How many 3-letter word sequences can be formed using the letters {a, b, c, d} if no letter is to be repeated?

Use a tree diagram for questions 11 and 12:

11. A family has two children, use a tree diagram to determine all four possibilities of outcomes by gender.	12. A coin is tossed three times and the sequence of heads and tails is recorded. Use a tree diagram to list all the possible outcomes.

Do the following problems using the Multiplication Axiom

13. In how many ways can a 4-question true-false test be answered?	14. In how many ways can three people be arranged to stand in a straight line?
15. A combination lock is opened by first turning to the left, then to the right, and then to the left again. If there are 30 digits on the dial, how many possible combinations are there?	16. How many different answers are possible for a multiple-choice test with 10 questions and five possible answers for each question?
17. In the past, a college required students to use a 4 digit PIN (Personal Identification Number) as their password for its registration system. How many different PINs are possible if each must have 4 digits with no restrictions on selection or arrangement of the digits used?	18. The college decided that a more secure password system is needed. New passwords must have 3 numerical digits followed by 6 letters. There are no restrictions on the selection of the numerical digits. However, the letters I and O are not permitted. How many different passwords are possible?

This page titled 1.E: Tree Diagrams and the Multiplication Axiom (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by Rupinder Sekhon and Roberta Bloom.

- **7.2.1: Tree Diagrams and the Multiplication Axiom (Exercises)** by Rupinder Sekhon and Roberta Bloom is licensed CC BY 4.0. Original source: https://www.deanza.edu/faculty/bloomroberta/math11/afm3files.html.html.

CHAPTER OVERVIEW

5: Discrete Probability Distributions

5.1: Prelude to Discrete Random Variables

5.2: Probability Distribution Function (PDF) for a Discrete Random Variable

5.3: Mean or Expected Value and Standard Deviation

5.4: Binomial Distribution

5.4.1: Binomial Distribution Formula

5.E: Discrete Random Variables (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 5: Discrete Probability Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Prelude to Discrete Random Variables

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.
- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

- A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?
- Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A *random variable* describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.



Figure **5.1.1** You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Random Variable Notation

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If **X is a random variable, then X is written in words, and x is given as a number.**

For example, let $X =$ the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Collaborative Exercise

Toss a coin ten times and record the number of heads. After all members of the class have completed the experiment (tossed a coin ten times and counted the number of heads), fill in Table. Let $X =$ the number of heads in ten tosses of the coin.

x	Frequency of x	Relative Frequency of x

a. Which value(s) of x occurred most frequently?
b. If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
c. What does the relative frequency column sum to?

Glossary

Random Variable (RV)

a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters $X, Y, Z,...$; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if $X =$ hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

This page titled 5.1: Prelude to Discrete Random Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2: Probability Distribution Function (PDF) for a Discrete Random Variable

A discrete probability distribution function has two characteristics:

a. Each probability is between zero and one, inclusive.
b. The sum of the probabilities is one.

Example 5.2.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$.

$P(x)$ = probability that X takes on a value x.

x	$P(x)$
0	$P(x=0) = \frac{2}{50}$
1	$P(x=1) = \frac{11}{50}$
2	$P(x=2) = \frac{23}{50}$
3	$P(x=3) = \frac{9}{50}$
4	$P(x=4) = \frac{4}{50}$
5	$P(x=5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

a. Each $P(x)$ is between zero and one, inclusive.
b. The sum of the probabilities is one, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \quad (5.2.1)$$

Exercise 5.2.1

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$. $P(x)$ = the probability that X takes on value x. Why is this a discrete probability distribution function (two reasons)?

X	$P(x)$
0	$P(x=0) = \frac{4}{50}$
1	$P(x=1) = \frac{8}{50}$
2	$P(x=2) = \frac{16}{50}$
3	$P(x=3) = \frac{14}{50}$
4	$P(x=4) = \frac{6}{50}$
5	$P(x=5) = \frac{2}{50}$

Answer

Each $P(x)$ is between 0 and 1, inclusive, and the sum of the probabilities is 1, that is:

$$\frac{4}{50}+\frac{8}{50}+\frac{16}{50}+\frac{14}{50}+\frac{6}{50}+\frac{2}{50}=1 \quad (5.2.2)$$

Example 5.2.2

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

a. Let X = the number of days Nancy _____.
b. X takes on what values?
c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in Example. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Solutions

a. Let X = the number of days Nancy attends class per week.

b. 0, 1, 2, and 3

c

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

Exercise 5.2.2

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Answer

X is the number of days Jeremiah attends basketball practice per week. X takes on the values 0, 1, and 2.

Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let $X =$ the number of years a new hire will stay with the company.

Let $P(x) =$ the probability that a new hire will stay with the company x years.

Exercise 4.2.3

Complete Table using the data provided.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

Answer

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

Exercise 4.2.4

$P(x = 4) =$ _____

Exercise 4.2.5

$P(x \geq 5) =$ _____

Answer

0.10 + 0.05 = 0.15

Exercise 4.2.6

On average, how long would you expect a new hire to stay with the company?

Exercise 4.2.7

What does the column "$P(x)$" sum to?

Answer

1

Use the following information to answer the next six exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

Exercise 4.2.8

Define the random variable X.

Exercise 4.2.9

What is the probability the baker will sell more than one batch? $P(x > 1) =$ _____

Answer

0.35 + 0.40 + 0.10 = 0.85

Exercise 4.2.10

What is the probability the baker will sell exactly one batch? $P(x = 1) =$ _____

Exercise 4.2.11

On average, how many batches should the baker make?

Answer

1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45

Use the following information to answer the next four exercises: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

Exercise 4.2.12

Define the random variable X.

Exercise 4.2.13

Construct a probability distribution table for the data.

Answer

x	$P(x)$
0	0.03
1	0.04
2	0.08
3	0.85

Exercise 4.2.14

We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

Exercise 4.2.15

Define the random variable X.

Answer

Let $X =$ the number of events Javier volunteers for each month.

Exercise 4.2.16

What values does x take on?

Exercise 4.2.17

Construct a PDF table.

Answer

x	$P(x)$
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

Exercise 4.2.18

Find the probability that Javier volunteers for less than three events each month. $P(x < 3) =$ _____

Exercise 4.2.19

Find the probability that Javier volunteers for at least one event each month. $P(x > 0) =$ _____

Answer

1 – 0.05 = 0.95

Glossary

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

This page titled 5.2: Probability Distribution Function (PDF) for a Discrete Random Variable is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Mean or Expected Value and Standard Deviation

The expected value is often referred to as the "long-term" average or mean. This means that over the long term of doing an experiment over and over, you would expect this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As you learned in Chapter 3, probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together). When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This "long-term average" is known as the mean or expected value of the experiment and is denoted by the Greek letter μ. In other words, after conducting many trials of an experiment, you would expect this average value.

To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

Example 5.3.1

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.

Solution

To do the problem, first let the random variable $X =$ the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x * P(x)$. In this column, you will multiply each x value by its probability.

Expected Value Table This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	$P(x)$	$x * P(x)$
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Add the last column $x * P(x)$ to find the long term average or expected value:

$$(0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1.$$

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$.

Example 5.3.2

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

x	$P(x)$	$x * P(x)$	$(x - \mu)^2 \cdot P(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x = 1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x = 2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x = 3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x = 4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x = 5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

Add the values in the third column of the table to find the expected value of X:

$$\mu = \text{Expected Value} = \frac{105}{50} = 2.1$$

Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x, multiply the square of its deviation by its probability. (Each deviation has the format $x-\mu$.

Add the values in the fourth column of the table:

$$0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$$

The standard deviation of X is the square root of this sum: $\sigma = \sqrt{1.05} \approx 1.0247$

The mean, μ, of a discrete probability function is the expected value.

$$\mu = \sum(x \bullet P(x))$$

The standard deviation, Σ, of the PDF is the square root of the variance.

$$\sigma = \sqrt{\sum[(x-\mu)2 \bullet P(x)]}$$

When all outcomes in the probability distribution are equally likely, these formulas coincide with the mean and standard deviation of the set of possible outcomes.

Exercise 5.3.2

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Answer

The expected value is 2.24

$$(0)\frac{4}{50}+(1)\frac{8}{50}+(2)\frac{16}{50}+(3)\frac{14}{50}+(4)\frac{6}{50}+(5)\frac{2}{50}=0+\frac{8}{50}+\frac{32}{50}+\frac{42}{50}+\frac{24}{50}+\frac{10}{50}=\frac{116}{50}=2.32 \quad (5.3.1)$$

Example 5.3.2

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay $2 to play and could profit $100,000 if you match all five numbers in order (you get your $2 back plus $100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let $X =$ the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and −2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)=(1)(10^{-5})$$
$$=0.00001.$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

$$1-0.00001=0.99999.1-0.00001=0.99999.$$

The expected value table is as follows:

Add the last column. –1.99998 + 1 = –0.99998

	x	$P(x)$	$xP(x)$
Loss	–2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

Since –0.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Exercise 5.3.3

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

Answer

Let $X =$ the amount of money you profit. The x-values are –$1 and $256.

The probability of guessing the right suit each time is $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)=\frac{1}{256}=0.0039$

The probability of losing is $1-\frac{1}{256}=\frac{255}{256}=0.9961$

$(0.0039)256+(0.9961)(-1)=0.9984+(-0.9961)=0.0023$ or 0.23 cents.

Example 5.3.4

Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

a. Define a random variable X.
b. Complete the following expected value table.
c. What is the expected value, μ? Do you come out ahead?

Solutions

a.

X = amount of profit

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

b.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

c.

Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Exercise 5.3.4

Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay $10. If you land on blue, you don't pay or win anything. If you land on green, you win $10. Complete the following expected value table.

	x	$P(x)$	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Answer

	x	$P(x)$	$x * P(x)$
Red	–10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$
Green	10	$\frac{1}{5}$	$\frac{1}{5}$

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x)–\mu^2P(x)$ and take the square root.

x	$P(x)$	$x * P(x)$	$(x–\mu)^2P(x)$
0	0.2	(0)(0.2) = 0	$(0 – 1.1)^2(0.2) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1 – 1.1)^2(0.5) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2 – 1.1)^2(0.3) = 0.243$

Add the last column in the table. $0.242 + 0.005 + 0.243 = 0.490$. The standard deviation is the square root of 0.49, or $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating μ and σ.

Example 5.3.5

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like Table and calculate the mean μ and standard deviation σ of X.

Solution

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

Calculating μ and σ.

x	$P(x)$	$xP(x)$	$(x–\mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0–1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1–1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(1–1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum:

$$\sigma = \sqrt{\frac{18}{36}} \approx 0.7071. \quad (5.3.2)$$

Example 5.3.6

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20. Let X = the amount of profit from a bet.

$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 21.42$

$P(\text{loss}) = P(\text{one moderate earthquake will not occur}) = 100$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X? Construct a table similar to Table and Table to help you answer these questions.

Answer

	x	$P(x)$	$xP(x)$	$(x–\mu^2)P(x)$
win	50	0.2142	10.71	$[50 – (–5.006)]^2(0.2142)$ = 648.0964
loss	–20	0.7858	–15.716	$[–20 – (–5.006)]^2(0.7858)$ = 176.6636

Mean = Expected Value = 10.71 + (–15.716) = –5.006.

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of \$5.01 per bet.

Standard Deviation $= \sqrt{648.0964 + 176.6636} \approx 28.7186$

Exercise 5.3.6

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. You bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X.

Answer

	x	$P(x)$	$x \cdot P(x)$	$(x - \mu^2) \cdot P(x)$
win	100	0.0108	1.08	$[100 – (–8.812)]^2 \cdot 0.0108$ = 127.8726
loss	–10	0.9892	–9.892	$[–10 – (–8.812)]^2 \cdot 0.9892$ = 1.3961

Mean = Expected Value $= \mu = 1.08 + (–9.892) = –8.812$

If you make this bet many times under the same conditions, your long term outcome will be an average loss of \$8.81 per bet.

Standard Deviation $= \sqrt{127.7826 + 1.3961} \approx 11.3696$

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

Summary

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

1. Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$
2. Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$

Glossary

Expected Value

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$.

Mean

a number that measures the central tendency; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (detonated by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Mean of a Probability Distribution

the long-term average of many trials of a statistical experiment

Standard Deviation of a Probability Distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

The Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

References

1. Class Catalogue at the Florida State University. Available online at apps.oti.fsu.edu/RegistrarCo...archFormLegacy (accessed May 15, 2013).
2. "World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. www.world-earthquakes.com/ind...thq_prediction (accessed May 15, 2013).

This page titled 5.3: Mean or Expected Value and Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **4.3: Mean or Expected Value and Standard Deviation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

5.4: Binomial Distribution

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn *with replacement* from a population of size N.

Three characteristics of a binomial experiment

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the n independent trials. The mean, μ, and variance, σ^2, for the binomial probability distribution are

$$\mu = np \tag{5.4.1}$$

and

$$\sigma^2 = npq. \tag{5.4.2}$$

The standard deviation, σ, is then

$$\sigma = \sqrt{npq}. \tag{5.4.3}$$

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example 5.4.1

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable $X =$ the number of students who withdraw from the randomly selected elementary physics class.

Exercise 5.4.1

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

Answer

a school that offers fruit in their lunch every day

Example 5.4.2

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the

values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Exercise 5.4.2

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Answer

$P(x = 12)$

Example 5.4.3

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let $X =$ the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Solution

$P(x > 10)$

Exercise 5.4.4

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Answer

$P(x > 3)$

Example 5.4.5

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

a. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
b. If we are interested in the number of students who do their homework on time, then how do we define X?
c. What values does x take on?
d. What is a "failure," in words?
e. If $p + q = 1$, then what is q?
f. The words "at least" translate as what kind of inequality for the probability question $P(x$ ____ $40)$.

Solution

a. failure
b. X = the number of statistics students who do their homework on time
c. 0, 1, 2, …, 50
d. Failure is defined as a student who does not complete his or her homework on time. The probability of a success is $p = 0.70$. The number of trials is $n = 50$.
e. $q = 0.30$
f. greater than or equal to (≥). The probability question is $P(x \geq 40)$.

Exercise 5.4.5

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Answer

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

Notation for the Binomial: B = Binomial Probability Distribution Function

$$X \sim B(n, p) \tag{5.4.4}$$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p; n = number of trials, p = probability of a success on each trial.

Example 5.4.6

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20, p = 0.41$, and $q = 1–0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate $(x = \text{value})$: $\text{binompdf}(n, p, \text{number})$ if "number" is left out, the result is the binomial probability table.

To calculate $P(x \leq \text{value})$: $\text{binomcdf}(n, p, \text{number})$ if "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,0.41,12). The result is $P(x \leq 12) = 0.9738$.

If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find $P(x > 12)$, use $1 - \text{binomcdf}(20, 0.41, 12)$.

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:

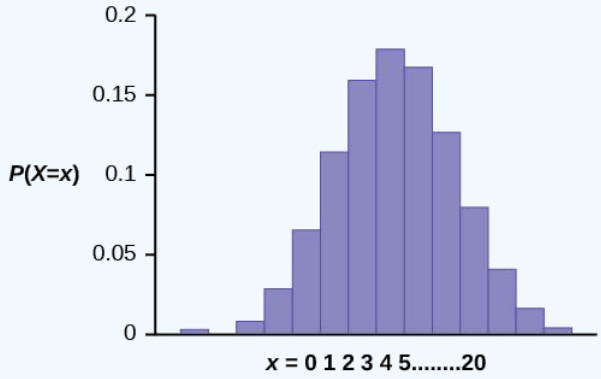


Figure 5.4.1 : The graph of $X \sim B(20, 0.41)$.

The *y*-axis contains the probability of x, where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.

$$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20. \quad (5.4.5)$$

Exercise 4.4.5

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Answer

$P(x \leq 14) = 0.9695$

Example 5.4.7

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let $X =$ the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution? Find the following probabilities:
   i. the probability that two pages feature signature artists
   ii. the probability that at most six pages feature signature artists
   iii. the probability that more than three pages feature signature artists.
c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Answer

a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
b. $X \sim B(100, 8560)(100, 8560)$
   i. $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = 0.2466$
   ii. $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = 0.9994$
   iii. $P(x > 3) = 1\text{–}P(x \leq 3) = 1\text{–binomcdf}\left(100, \frac{8}{560}, 3\right) = 1\text{–}0.9443 = 0.0557$
c. i. Mean $= np = (100)\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
   ii. Standard Deviation $= \sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

Exercise 5.4.7

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?
b. Use your calculator to find the following probabilities:
   i. the probability that 25 adults in the sample prefer saving over spending
   ii. the probability that at most 20 adults prefer saving
   iii. the probability that more than 30 adults prefer saving
c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

Answer

a. $X \sim B(50, 0.6)$
b. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:
   i. $P(x = 25) = \text{binompdf}(50, 0.6, 25) = 0.0405$
   ii. $P(x \leq 20) = \text{binomcdf}(50, 0.6, 20) = 0.0034$
   iii. $(x > 30) = 1 - \text{binomcdf}(50, 0.6, 30) = 1–0.5535 = 0.4465$
c. i. Mean $= np = 50(0.6) = 30$
   ii. Standard Deviation $= \sqrt{npq} = \sqrt{50(0.6)(0.4)} \approx 3.4641$

Example 5.4.8

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that at most eight people develop pancreatic cancer
d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Answer

a. $X \sim B(200, 0.0128)$
b. i. Mean $= np = 200(0.0128) = 2.56$
   ii. Standard Deviation $= \sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$
c. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:
   $P(x \leq 8) = \text{binomcdf}(200, 0.0128, 8) = 0.9988$
d. $P(x = 5) = \text{binompdf}(200, 0.0128, 5) = 0.0707$
   $P(x = 6) = \text{binompdf}(200, 0.0128, 6) = 0.0298$
   So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Exercise 5.4.8

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let $X =$ the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

Answer

a. $X \sim B(80, 0.613)$
b. i. Mean $= np = 80(0.613) = 49.04$
   ii. Standard Deviation $= \sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$
c. Using the TI-83, 83+, 84 calculator with instructions as provided in Example:
   $P(x = 60) = \text{binompdf}(80, 0.613, 60) = 0.0036$
d. $P(x > 50) = 1–P(x \leq 50) = 1–\text{binomcdf}(80, 0.613, 50) = 1–0.6282 = 0.3718$

Example 5.4.9

The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and

a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

? Exercise 5.4.9

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Answer

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

References

1. "Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...first&sort=asc (accessed May 15, 2015).
2. "Distance Education." Wikipedia. Available online at http://en.Wikipedia.org/wiki/Distance_education (accessed May 15, 2013).
3. "NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).
4. Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/am...-spending.aspx (accessed May 15, 2013).
5. Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/N...eshman2011.pdf (accessed May 15, 2013).
6. "The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publicat...k/geos/af.html (accessed May 15, 2013).
7. "What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at www.cancer.org/cancer/pancrea...key-statistics (accessed May 15, 2013).

Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

There are a fixed number of trials, n.

There are only two possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.

The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable $X =$ the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

Formula Review

- $X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.

- $X =$ the number of successes in n independent trials
- $n =$ the number of independent trials
- X takes on the values $x = 0, 1, 2, 3, \ldots, n$
- $p =$ the probability of a success for any trial
- $q =$ the probability of a failure for any trial
- $p + q = 1$
- $q = 1–p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

Exercise 4.4.9

In words, define the random variable X.

Answer

$X =$ the number that reply “yes”

Exercise 4.4.10

$X \sim$ _____(_____,_____)

Exercise 4.4.11

What values does the random variable X take on?

Answer

0, 1, 2, 3, 4, 5, 6, 7, 8

Exercise 4.4.12

Construct the probability distribution function (PDF).

x	$P(x)$

Exercise 4.4.13

On average (μ), how many would you expect to answer yes?

Answer

5.7

Exercise 4.4.14

What is the standard deviation (σ)?

Exercise 4.4.15

What is the probability that at most five of the freshmen reply “yes”?

Answer

0.4151

Exercise 4.4.16

What is the probability that at least two of the freshmen reply “yes”?

Glossary

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

Bernoulli Trials

an experiment with the following characteristics:

1. There are only two possible outcomes called “success” and “failure” for each trial.
2. The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Probability Distribution

a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \ B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is
$P(X = x) = \binom{n}{x} p^x q^{n-x}$.

This page titled 5.4: Binomial Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.4.1: Binomial Distribution Formula

Example 5.4.1.1: Shock Study

Suppose we randomly selected four individuals to participate in the "shock" study. What is the chance exactly one of them will be a success? Let's call the four people Allen (A), Brittany (B), Caroline (C), and Damian (D) for convenience. Also, suppose 35% of people are successes as in the previous version of this example.

Solution

Let's consider a scenario where one person refuses:

$$\begin{aligned} P(A = \text{refuse}; B = \text{shock}; C = \text{shock}; D = \text{shock}) &= P(A = \text{refuse})P(B = \text{shock}) \\ &\quad P(C = \text{shock})P(D = \text{shock}) \\ &= (0.35)(0.65)(0.65)(0.65) \\ &= (0.35)^1(0.65)^3 \\ &= 0.096 \end{aligned}$$

But there are three other scenarios: Brittany, Caroline, or Damian could have been the one to refuse. In each of these cases, the probability is again

$$P = (0.35)^1(0.65)^3.$$

These four scenarios exhaust all the possible ways that exactly one of these four people could refuse to administer the most severe shock, so the total probability is

$$4 \times (0.35)^1(0.65)^3 = 0.38.$$

Exercise 5.4.1.1

Verify that the scenario where Brittany is the only one to refuse to give the most severe shock has probability $(0.35)^1(0.65)^3$.

Answer

$$\begin{aligned} P(A = shock; B = refuse; C = shock; D = shock) &= (0.65)(0.35)(0.65)(0.65) \\ &= (0.35)^1(0.65)^3. \end{aligned}$$

The Binomial Distribution

The scenario outlined in Example 5.4.1.1 is a special case of what is called the binomial distribution. The binomial distribution describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p (in Example 5.4.1.1, n = 4, k = 1, p = 0.35). We would like to determine the probabilities associated with the binomial distribution more generally, i.e. we want a formula where we can use n, k, and p to obtain the probability. To do this, we reexamine each part of the example.

There were four individuals who could have been the one to refuse, and each of these four scenarios had the same probability. Thus, we could identify the nal probability as

Misplaced '#'

The first component of this equation is the number of ways to arrange the k = 1 successes among the n = 4 trials. The second component is the probability of any of the four (equally probable) scenarios.

Consider P(single scenario) under the general case of k successes and n-k failures in the n trials. In any such scenario, we apply the Multiplication Rule for independent events:

$$p^k(1-p)^{n-k} \tag{5.4.1.1}$$

This is our general formula for P(single scenario).

Secondly, we introduce a general formula for the number of ways to choose k successes in n trials, i.e. arrange k successes and n - k failures:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5.4.1.2}$$

The quantity $\binom{n}{k}$ is read **n choose k**.[30] The exclamation point notation (e.g. k!) denotes a **factorial** expression.

$$\begin{aligned} 0! &= 1 \\ 1! &= 1 \\ 2! &= 2 \times 1 = 2 \\ 3! &= 3 \times 2 \times 1 = 6 \\ 4! &= 4 \times 3 \times 2 \times 1 = 24 \\ &\vdots \\ n! &= n \times (n-1) \times \cdots \times 3 \times 2 \times 1 \end{aligned} \tag{5.4.1.3}$$

Substituting Equation 5.4.1.3 into Equation 5.4.1.2, we can compute the number of ways to choose $k = 1$ successes in $n = 4$ trials:

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} \tag{5.4.1.4}$$

$$= \frac{4!}{1!3!} \tag{5.4.1.5}$$

$$= \frac{4 \times 3 \times 2 \times 1}{(1)(3 \times 2 \times 1)} \tag{5.4.1.6}$$

$$= 4 \tag{5.4.1.7}$$

This result is exactly what we found by carefully thinking of each possible scenario in Example 5.4.1.1.

Other notations

Other notation for **n choose k** includes nC_k, C_n^k, and $C(n,k)$.

Substituting n choose k for the number of scenarios and $p^k(1-p)^{n-k}$ for the single scenario probability in Equation ??? yields the general binomial formula (Equation 5.4.1.8).

Definition: Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{5.4.1.8}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \sigma^2 = np(1-p) \sigma = \sqrt{np(1-p)} \tag{5.4.1.9}$$

TIP: Four conditions to check if it is binomial?

1. The trials independent.
2. The number of trials, n, is fixed.
3. Each trial outcome can be classified as a success or failure.
4. The probability of a success, p, is the same for each trial.

Example 5.4.1.2

What is the probability that 3 of 8 randomly selected students will refuse to administer the worst shock, i.e. 5 of 8 will?

Solution

We would like to apply the binomial model, so we check our conditions. The number of trials is fixed (n = 8) (condition 2) and each trial outcome can be classi ed as a success or failure (condition 3). Because the sample is random, the trials are independent (condition 1) and the probability of a success is the same for each trial (condition 4).

In the outcome of interest, there are k = 3 successes in n = 8 trials, and the probability of a success is p = 0.35. So the probability that 3 of 8 will refuse is given by

$$\binom{8}{3}(0.35)^k(1-0.35)^{8-3} = \frac{8!}{3!(8-3)!}(0.35)^k(1-0.35)^{8-3}$$
$$= \frac{8!}{3!5!}(0.35)^3(0.65)^5$$

Dealing with the factorial part:

$$\frac{8!}{3!5!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(3\times2\times1)(5\times4\times3\times2\times1)}$$
$$= \frac{8\times7\times6}{3\times2\times1}$$
$$= 56$$

Using $(0.35)^3(0.65)^5 \approx 0.005$, the final probability is about 56 * 0.005 = 0.28.

TIP: computing binomial probabilities

The rst step in using the binomial model is to check that the model is appropriate. The second step is to identify n, p, and k. The final step is to apply the formulas and interpret the results.

TIP: computing n choose k

In general, it is useful to do some cancelation in the factorials immediately. Alternatively, many computer programs and calculators have built in functions to compute n choose k, factorials, and even entire binomial probabilities.

Exercise 5.4.1.2A

If you ran a study and randomly sampled 40 students, how many would you expect to refuse to administer the worst shock? What is the standard deviation of the number of people who would refuse? Equation 5.4.1.9 may be useful.

Answer

We are asked to determine the expected number (the mean) and the standard deviation, both of which can be directly computed from the formulas in Equation 5.4.1.9:

$$\mu = np = 40\times0.35 = 14$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{40\times0.35\times0.65} = 0.02.$$

Because very roughly 95% of observations fall within 2 standard deviations of the mean (see Section 1.6.4), we would probably observe at least 8 but less than 20 individuals in our sample who would refuse to administer the shock.

Exercise 5.4.1.2B

The probability that a random smoker will develop a severe lung condition in his or her lifetime is about 0:3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?

Answer

One possible answer: if the friends know each other, then the independence assumption is probably not satis ed. For example, acquaintances may have similar smoking habits.

Example 5.4.1.3

Suppose these four friends do not know each other and we can treat them as if they were a random sample from the population. Is the binomial model appropriate? What is the probability that

a. none of them will develop a severe lung condition?
b. One will develop a severe lung condition?
c. That no more than one will develop a severe lung condition?

Solution

To check if the binomial model is appropriate, we must verify the conditions. (i) Since we are supposing we can treat the friends as a random sample, they are independent. (ii) We have a fixed number of trials (n = 4). (iii) Each outcome is a success or failure. (iv) The probability of a success is the same for each trials since the individuals are like a random sample (p = 0.3 if we say a "success" is someone getting a lung condition, a morbid choice). Compute parts (a) and (b) from the binomial formula in Equation 5.4.1.8:

$$P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$$

$$P(1) = \binom{4}{1}(0.3)^1(0.7)^3 = 0.4116.$$

Note: 0! = 1.

Part (c) can be computed as the sum of parts (a) and (b):

$$P(0) + P(1) = 0.2401 + 0.4116 = 0.6517. \tag{5.4.1.10}$$

That is, there is about a 65% chance that no more than one of your four smoking friends will develop a severe lung condition.

Exercise 5.4.1.3A

What is the probability that at least 2 of your 4 smoking friends will develop a severe lung condition in their lifetimes?

Answer

The complement (no more than one will develop a severe lung condition) as computed in Example **5.4.1.3** as 0.6517, so we compute one minus this value: 0.3483.

Exercise 5.4.1.3B

Suppose you have 7 friends who are smokers and they can be treated as a random sample of smokers.

a. How many would you expect to develop a severe lung condition, i.e. what is the mean?
b. What is the probability that at most 2 of your 7 friends will develop a severe lung condition.

Answer a

μ = 0.3 \times 7 = 2.1.

Answer b

P(0, 1, or 2 develop severe lung condition) = P(k = 0)+P(k = 1)+P(k = 2) = 0:6471.

Below we consider the first term in the binomial probability, **n choose k** under some special scenarios.

Exercise 5.4.1.3C

Why is it true that $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ for any number n?

Solution

Frame these expressions into words. How many different ways are there to arrange 0 successes and n failures in n trials? (1 way.) How many different ways are there to arrange n successes and 0 failures in n trials? (1 way.)

Exercise 5.4.1.3D

How many ways can you arrange one success and n -1 failures in n trials? How many ways can you arrange n -1 successes and one failure in n trials?

Solution

One success and n - 1 failures: there are exactly n unique places we can put the success, so there are n ways to arrange one success and n - 1 failures. A similar argument is used for the second question. Mathematically, we show these results by verifying the following two equations:

$$\binom{n}{1} = n, \binom{n}{n-1} = n \tag{5.4.1.11}$$

Contributors and Attributions

- David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)

This page titled 5.4.1: Binomial Distribution Formula is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

5.E: Discrete Random Variables (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

4.1: Introduction

4.2: Probability Distribution Function (PDF) for a Discrete Random Variable

Q 4.2.1

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in Table.

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. In words, define the random variable X.
b. What does it mean that the values zero, one, and two are not included for x in the PDF?

Exercise 4.3.5

Complete the expected value table.

x	$P(x)$	$x * P(x)$
0	0.2	
1	0.2	
2	0.4	
3	0.2	

Exercise 4.3.6

Find the expected value from the expected value table.

x	$P(x)$	$x * P(x)$
2	0.1	
4	0.3	4(0.3) = 1.2
6	0.4	6(0.4) = 2.4
8	0.2	8(0.2) = 1.6

Answer

$0.2 + 1.2 + 2.4 + 1.6 = 5.4$

Exercise 4.3.7

Find the standard deviation.

x	$P(x)$	$x * P(x)$	$(x–\mu)^2P(x)$
2	0.1	2(0.1) = 0.2	$(2–5.4)^2(0.1) = 1.156$
4	0.3	4(0.3) = 1.2	$(4–5.4)^2(0.3) = 0.588$
6	0.4	6(0.4) = 2.4	$(6–5.4)^2(0.4) = 0.144$
8	0.2	8(0.2) = 1.6	$(8–5.4)^2(0.2) = 1.352$

Exercise 4.3.8

Identify the mistake in the probability distribution table.

x	$P(x)$	$x * P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

Answer

The values of $P(x)$ do not sum to one.

Exercise 4.3.9

Identify the mistake in the probability distribution table.

x	$P(x)$	$x * P(x)$
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	$P(x)$	$x * P(x)$
1	0.35	
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

Exercise 4.3.10

Define the random variable X.

Answer

Let $X =$ the number of years a physics major will spend doing post-graduate research.

Exercise 4.3.11

Define $P(x)$, or the probability of x.

Exercise 4.3.12

Find the probability that a physics major will do post-graduate research for four years. $P(x = 4) =$ _____

Answer

$1–0.35–0.20–0.15–0.10–0.05 = 0.15$

Exercise 4.3.13

Find the probability that a physics major will do post-graduate research for at most three years. $P(x \leq 3) =$ _____

Exercise 4.3.14

On average, how many years would you expect a physics major to spend doing post-graduate research?

Answer

$1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 0.35 + 0.40 + 0.45 + 0.60 + 0.50 + 0.30 = 2.6$ years

Use the following information to answer the next seven exercises: A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let $X =$ the number of years a student will study ballet with the teacher.
- Let $P(x) =$ the probability that a student will study ballet x years.

Exercise 4.3.15

Complete Table using the data provided.

x	$P(x)$	$x * P(x)$
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

Exercise 4.3.16

In words, define the random variable X.

Answer

X is the number of years a student studies ballet with the teacher.

Exercise 4.3.17

$P(x = 4) =$ _____

Exercise 4.3.18

$P(x < 4) =$ _____

Answer

$0.10 + 0.05 + 0.10 = 0.25$

Exercise 4.3.19

On average, how many years would you expect a child to study ballet with this teacher?

Exercise 4.3.20

What does the column "$P(x)$" sum to and why?

Answer

The sum of the probabilities sum to one because it is a probability distribution.

Exercise 4.3.21

What does the column "$x * P(x)$" sum to and why?

Exercise 4.3.22

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

Answer

$-2\left(\frac{40}{52}\right) + 30\left(\frac{12}{52}\right) = -1.54 + 6.92 = 5.38$

Exercise 4.3.23

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

4.3: Mean or Expected Value and Standard Deviation

Q 4.3.1

A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.

a. What are you interested in here?
b. In words, define the random variable X.
c. List the values that X may take on.
d. Construct a PDF.
e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

Q 4.3.2

A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win $6
- If the card is a face card, and the coin lands on Tails, you win $2
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

S 4.3.2

The variable of interest is X, or the gain or loss, in dollars.

The face cards jack, queen, and king. There are $(3)(4) = 12$ face cards and $52–12 = 40$ cards that are not face cards.

We first need to construct the probability distribution for X. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Card Event	X net gain/loss	$P(X)$
Face Card and Heads	6	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right) = \left(\frac{6}{52}\right)$
Face Card and Tails	2	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right) = \left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	–2	$\left(\frac{40}{52}\right)(1) = \left(\frac{40}{52}\right)$

- $\text{Expected value} = (6)\left(\frac{6}{52}\right) + (2)\left(\frac{6}{52}\right) + (-2)\left(\frac{40}{52}\right) = -\frac{32}{52}$
- Expected value = –$0.62, rounded to the nearest cent
- If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

Q 4.3.3

You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

Q 4.3.4

Complete the PDF and answer the questions.

x	$P(x)$	$xP(x)$
0	0.3	
1	0.2	
2		
3	0.4	

a. Find the probability that $x = 2$.
b. Find the expected value.

S 4.3.4

a. 0.1
b. 1.6

Q 4.3.5

Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

a. What are you ultimately interested in here (the value of the roll or the money you win)?
b. In words, define the Random Variable X.
c. List the values that X may take on.
d. Construct a PDF.
e. Over the long run of playing this game, what are your expected average winnings per game?
f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

Q 4.3.6

A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

S 4.3.6

a.

Software Company	
x	$P(x)$
5,000,000	0.10
1,000,000	0.30
–1,000,000	0.60

Hardware Company	
x	$P(x)$
3,000,000	0.20
1,000,000	0.40
–1,000,00	0.40

Biotech Firm	
x	$P(x)$
6,00,000	0.10
0	0.70
–1,000,000	0.20

b. $200,000; $600,000; $400,000
c. third investment because it has the lowest probability of loss
d. first investment because it has the highest probability of loss
e. second investment

Q 4.3.7

Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let $X =$ the number of children married people have.

x	$P(x)$	$xP(x)$
0	0.10	
1	0.20	
2	0.30	
3		
4	0.10	
5	0.05	
6 (or more)	0.05	

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

Q 4.3.8

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in the Table.

x	$P(x)$
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

On average, how many years do you expect it to take for an individual to earn a B.S.?

S 4.3.8

4.85 years

Q 4.3.9

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.70
5	0.04

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs. Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	$P(x)$
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

e. At which store is the expected number of DVDs rented per customer higher?
f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
g. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Q 4.3.10

A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn’t matter. I expect to break even.

S 4.3.10

b

Q 4.3.11

Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Define the PDF for X.
c. Find the mean of X.
d. Find the standard deviation of X.

Q 4.3.12

In a lottery, there are 250 prizes of $5, 50 prizes of $25, and ten prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

S 4.3.12

Let $X =$ the amount of money to be won on a ticket. The following table shows the PDF for X.

x	$P(x)$
0	0.969
5	$\frac{250}{10,000} = 0.025$
25	$\frac{50}{10,000} = 0.005$
100	$\frac{10}{10,000} = 0.001$

Calculate the expected value of X.

$$0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35 \quad (5.E.1)$$

A fair price for a ticket is $0.35. Any price over $0.35 will enable the lottery to raise money.

4.4: Binomial Distribution

Q 4.4.1

According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery.

Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

Q 4.4.2

Define the random variable and list its possible values.

S 4.4.2

$X =$ the number of patients calling in claiming to have the flu, who actually have the flu.

$X = 0, 1, 2, ...25$

Q 4.4.3

State the distribution of X.

Q 4.4.4

Find the probability that at least four of the 25 patients actually have the flu.

S 4.4.4

0.0165

Q 4.4.5

On average, for every 25 patients calling in, how many do you expect to have the flu?

Q 4.4.6

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given Table. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

S 4.4.6

a. $X =$ the number of DVDs a Video to Go customer rents
b. 0.12
c. 0.11
d. 0.77

Q 4.4.7

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.
f. Find the probability that more than two students will attend.

Use the following information to answer the next three exercises: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

Q 4.4.8

The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

S 4.4.8

d. 4.43

Let $X =$ the number of games won in that upcoming month.

Q 4.4.9

What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476
b. 0.2336
c. 0.7664
d. 0.8903

Q 4.4.10

What is the probability that the San Jose Sharks win at least five games in that upcoming month?

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

S 4.4.10

c

Q 4.4.11

A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

Q 4.4.12

A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

S 4.4.13

- $X =$ number of questions answered correctly
- $X \sim B(32, \frac{1}{3})$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."
- Using your calculator's distribution menu: $1–\text{binomcdf}(32, \frac{1}{3}, 24)$
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

Q 4.4.14

Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

Q 4.4.15

More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most ten offer such courses.
f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

S 4.4.15

a. $X =$ the number of college and universities that offer online offerings.
b. 0, 1, 2, …, 13
c. $X \sim B(13, 0.96)$
d. 12.48
e. 0.0135
f. $P(x = 12) = 0.3186 P(x = 13) = 0.5882$ More likely to get 13.

Q 4.4.16

Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Q 4.4.17

At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to **not** to use the foil as their main weapon?
e. Find the probability that six do **not** use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

S 4.4.17

a. $X =$ the number of fencers who do **not** use the foil as their main weapon
b. 0, 1, 2, 3,... 25
c. $X \sim B(25, 0.40)$
d. 10
e. 0.0442
f. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

Q 4.4.18

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Q 4.4.19

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

S 4.4.19

a. $X =$ the number of audits in a 20-year period
b. 0, 1, 2, …, 20
c. $X \sim B(20, 0.02)$
d. 0.4
e. 0.6676
f. 0.0071

Q 4.4.20

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

Q 4.4.21

There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let $X =$ number of matches and $Y =$ profit per game.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player.
g. Determine who has the advantage, the player or the house.

S 4.4.21

a. $X =$ the number of matches
b. 0, 1, 2, 3
c. $X \sim B(3, 16)(3, 16)$
d. In dollars: −1, 1, 2, 3
e. $\frac{1}{2}$
f. Multiply each Y value by the corresponding X probability from the PDF table. The answer is −0.0787. You lose about eight cents, on average, per game.
g. The house has the advantage.

Q 4.4.22

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let $X =$ the number of people who have access to electricity.

a. What is the probability distribution for X?
b. Using the formulas, calculate the mean and standard deviation of X.
c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most ten people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

Q 4.4.23

The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let $X =$ the number of people who are literate.

1. Sketch a graph of the probability distribution of X.
2. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
3. Find the probability that more than five people in the sample are literate. Is it is more likely that three people or four people are literate.

S 4.4.23

1. $X \sim B(15, 0.281)$

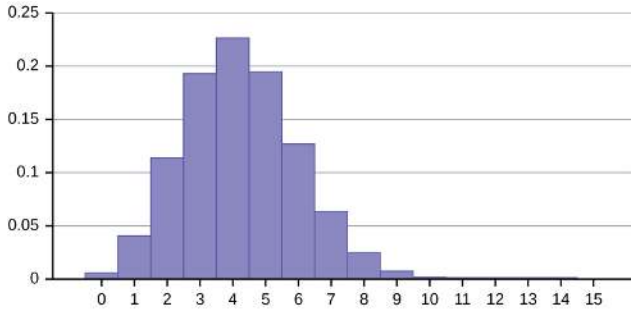


Figure 4.4.1.

2. 1. Mean $= \mu = np = 15(0.281) = 4.215$
   2. Standard Deviation $= \sigma = \sqrt{npq} = \sqrt{15(0.281)(0.719)} = 1.7409$
3. $P(x > 5) = 1 – P(x \leq 5) = 1 – \text{binomcdf}(15, 0.281, 5) = 1 – 0.7754 = 0.2246$
   $P(x = 3) = \text{binompdf}(15, 0.281, 3) = 0.1927$
   $P(x = 4) = \text{binompdf}(15, 0.281, 4) = 0.2259$
   It is more likely that four people are literate that three people are.

4.5: Geometric Distribution

Q 4.5.1

A consumer looking to buy a used red Miata car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28%. We are interested in the number of dealerships she must call.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many dealerships would we expect her to have to call until she finds one that has the car?
e. Find the probability that she must call at most four dealerships.
f. Find the probability that she must call three or four dealerships.

Q 4.5.2

Suppose that the probability that an adult in America will watch the Super Bowl is 40%. Each person is considered independent. We are interested in the number of adults in America we must survey until we find one who will watch the Super Bowl.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many adults in America do you expect to survey until you find one who will watch the Super Bowl?
e. Find the probability that you must ask seven people.
f. Find the probability that you must ask three or four people.

S 4.5.2

a. $X =$ the number of adults in America who are surveyed until one says he or she will watch the Super Bowl.
b. $X \sim G(0.40)$
c. 2.5
d. 0.0187
e. 0.2304

Q 4.5.3

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose we are interested in the number of California residents we must survey until we find a resident who does **not** have adequate earthquake supplies.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. What is the probability that we must survey just one or two residents until we find a California resident who does not have adequate earthquake supplies?
e. What is the probability that we must survey at least three California residents until we find a California resident who does not have adequate earthquake supplies?
f. How many California residents do you expect to need to survey until you find a California resident who **does not** have adequate earthquake supplies?
g. How many California residents do you expect to need to survey until you find a California resident who **does** have adequate earthquake supplies?

Q 4.5.4

In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked more than once.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many pages do you expect to advertise footwear on them?
e. Is it probable that all twenty will advertise footwear on them? Why or why not?
f. What is the probability that fewer than ten will advertise footwear on them?
g. Reminder: A page may be picked more than once. We are interested in the number of pages that we must randomly survey until we find one that has footwear advertised on it. Define the random variable X and give its distribution.
h. What is the probability that you only need to survey at most three pages in order to find one that advertises footwear on it?
i. How many pages do you expect to need to survey in order to find one that advertises footwear?

S 4.5.4

a. $X =$ the number of pages that advertise footwear
b. X takes on the values 0, 1, 2, ..., 20
c. $X \sim B(20, \frac{29}{192})$
d. 3.02
e. No
f. 0.9997
g. $X =$ the number of pages we must survey until we find one that advertises footwear. $X \sim G(\frac{29}{192})$
h. 0.3881
i. 6.6207 pages

Q 4.5.5

Suppose that you are performing the probability experiment of rolling one fair six-sided die. Let $\mathbf{F}$ be the event of rolling a four or a five. You are interested in how many times you need to roll the die in order to obtain the first four or five as the outcome.

- $p =$ probability of success (event $\mathbf{F}$ occurs)
- $q =$ probability of failure (event $\mathbf{F}$ does not occur)

a. Write the description of the random variable X.
b. What are the values that X can take on?
c. Find the values of p and q.
d. Find the probability that the first occurrence of event $\mathbf{F}$ (rolling a four or five) is on the second trial.

Q 4.5.5

Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random. What values does X take on?

S 4.5.5

0, 1, 2, and 3

Q 4.5.6

The World Bank records the prevalence of HIV in countries around the world. According to their data, "Prevalence of HIV refers to the percentage of people ages 15 to 49 who are infected with HIV."[1] In South Africa, the prevalence of HIV is 17.3%. Let $X =$ the number of people you test until you find a person infected with HIV.

a. Sketch a graph of the distribution of the discrete random variable X.
b. What is the probability that you must test 30 people to find one with HIV?
c. What is the probability that you must ask ten people?
d. Find the (i) mean and (ii) standard deviation of the distribution of X.

Q 4.5.7

According to a recent Pew Research poll, 75% of millenials (people born between 1981 and 1995) have a profile on a social networking site. Let $X =$ the number of millenials you ask until you find a person without a profile on a social networking site.

a. Describe the distribution of X.
b. Find the (i) mean and (ii) standard deviation of X.
c. What is the probability that you must ask ten people to find one person without a social networking site?
d. What is the probability that you must ask 20 people to find one person without a social networking site?
e. What is the probability that you must ask *at most* five people?

S 4.5.7

a. $X \sim \mathrm{G}(0.25)$
b. i. Mean $= \mu = \frac{1}{p} = \frac{1}{0.25} = 4$
   ii. Standard Deviation $= \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.25}{0.25^2}} \approx 3.4641$
c. $P(x = 10) = \mathrm{geometpdf}(0.25, 10) = 0.0188$
d. $P(x = 20) = \mathrm{geometpdf}(0.25, 20) = 0.0011$
e. $P(x \leq 5) = \mathrm{geometcdf}(0.25, 5) = 0.7627$

Footnotes

1. "Prevalence of HIV, total (% of populations ages 15-49)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...last&sort=desc (accessed May 15, 2013).

4.6: Hypergeometric Distribution

Q 4.6.1

A group of Martial Arts students is planning on participating in an upcoming demonstration. Six are students of Tae Kwon Do; seven are students of Shotokan Karate. Suppose that eight students are randomly picked to be in the first demonstration. We are interested in the number of Shotokan Karate students in that first demonstration.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many Shotokan Karate students do we expect to be in that first demonstration?

Q 4.6.2

In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked at most once.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many pages do you expect to advertise footwear on them?
e. Calculate the standard deviation.

S 4.6.2

a. $X =$ the number of pages that advertise footwear
b. 0, 1, 2, 3, ..., 20
c. $X \sim \mathrm{H}(29, 163, 20); r = 29, b = 163, n = 20$
d. 3.03
e. 1.5197

Q 4.6.3

Suppose that a technology task force is being formed to study technology awareness among instructors. Assume that ten people will be randomly chosen to be on the committee from a group of 28 volunteers, 20 who are technically proficient and eight who are not. We are interested in the number on the committee who are **not** technically proficient.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many instructors do you expect on the committee who are **not** technically proficient?
e. Find the probability that at least five on the committee are not technically proficient.
f. Find the probability that at most three on the committee are not technically proficient.

Q 4.6.4

Suppose that nine Massachusetts athletes are scheduled to appear at a charity benefit. The nine are randomly chosen from eight volunteers from the Boston Celtics and four volunteers from the New England Patriots. We are interested in the number of Patriots picked.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. Are you choosing the nine athletes with or without replacement?

S 4.6.4

a. $X =$ the number of Patriots picked
b. 0, 1, 2, 3, 4
c. $X \sim H(4, 8, 9)$
d. Without replacement

Q 4.6.5

A bridge hand is defined as 13 cards selected at random and without replacement from a deck of 52 cards. In a standard deck of cards, there are 13 cards from each suit: hearts, spades, clubs, and diamonds. What is the probability of being dealt a hand that does not contain a heart?

a. What is the group of interest?
b. How many are in the group of interest?
c. How many are in the other group?

d. Let $X =$ _____. What values does X take on?
e. The probability question is P(_____).
f. Find the probability in question.
g. Find the (i) mean and (ii) standard deviation of X.

4.7: Poisson Distribution

Q 4.7.1

The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let $X =$ the number of calls received at noon.

a. Find the mean and standard deviation of X.
b. What is the probability that the office receives at most six calls at noon on Monday?
c. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
d. What is the probability that the office receives more than eight calls at noon?

S 4.7.1

a. $X \sim P(5.5); \mu = 5.5; \sigma = \sqrt{5.5} \approx 2.3452$
b. $P(x \leq 6) = \text{poissoncdf}(5.5, 6) \approx 0.6860$
c. There is a 15.7% probability that the law staff will receive more calls than they can handle.
d. $P(x > 8) = 1 – P(x \leq 8) = 1 – \text{poissoncdf}(5.5, 8) \approx 1 – 0.8944 = 0.1056$

Q 4.7.2

The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let $X =$ the number of births in an hour.

a. Find the mean and standard deviation of X.
b. Sketch a graph of the probability distribution of X.
c. What is the probability that the maternity ward will deliver three babies in one hour?
d. What is the probability that the maternity ward will deliver at most three babies in one hour?
e. What is the probability that the maternity ward will deliver more than five babies in one hour?

Q 4.7.3

A manufacturer of Christmas tree light bulbs knows that 3% of its bulbs are defective. Find the probability that a string of 100 lights contains at most four defective bulbs using both the binomial and Poisson distributions.

S 4.7.3

Let $X =$ the number of defective bulbs in a string.

Using the Poisson distribution:

- $\mu = np = 100(0.03) = 3$
- $X \sim P(3)$
- $P(x \leq 4) = \text{poissoncdf}(3, 4) \approx 0.8153$

Using the binomial distribution:

- $X \sim \text{B}(100, 0.03)$
- $P(x \leq 4) = \text{binomcdf}(100, 0.03, 4) \approx 0.8179$

The Poisson approximation is very good—the difference between the probabilities is only 0.0026.

Q 4.7.4

The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Japanese average.
f. Find the probability that she has more children than the Japanese average.

Q 4.7.5

The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.

a. In words, define the Random Variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Spanish average.
f. Find the probability that she has more children than the Spanish average .

S 4.7.5

a. $X =$ the number of children for a Spanish woman
b. 0, 1, 2, 3,...
c. $X \sim P(1.47)$
d. 0.2299
e. 0.5679
f. 0.4321

Q 4.7.6

Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____
d. Find the probability that she has no litters in one year.
e. Find the probability that she has at least two litters in one year.
f. Find the probability that she has exactly three litters in one year.

Q 4.7.7

he chance of having an extra fortune in a fortune cookie is about 3%. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem, but only use one distribution to solve the problem.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many cookies do we expect to have an extra fortune?
e. Find the probability that none of the cookies have an extra fortune.
f. Find the probability that more than three have an extra fortune.
g. As n increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

S 4.7.7

a. $X =$ the number of fortune cookies that have an extra fortune
b. 0, 1, 2, 3,... 144

c. $X \sim B(144, 0.03)$ or $P(4.32)$
d. 4.32
e. 0.0124 or 0.0133
f. 0.6300 or 0.6264
g. As n gets larger, the probabilities get closer together.

Q 4.7.8

According to the South Carolina Department of Mental Health web site, for every 200 U.S. women, the average number who suffer from anorexia is one. Out of a randomly chosen group of 600 U.S. women determine the following.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to suffer from anorexia?
e. Find the probability that no one suffers from anorexia.
f. Find the probability that more than four suffer from anorexia.

Q 4.7.9

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to answer the following questions.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to be audited?
e. Find the probability that no one was audited.
f. Find the probability that at least three were audited.

S 4.7.9

a. $X =$ the number of people audited in one year
b. 0, 1, 2, ..., 100
c. $X \sim P(2)$
d. 2
e. 0.1353
f. 0.3233

Q 4.7.10

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Q 4.7.11

On average, Pierre, an amateur chef, drops three pieces of egg shell into every two cake batters he makes. Suppose that you buy one of his cakes.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many pieces of egg shell do you expect to be in the cake?
e. What is the probability that there will not be any pieces of egg shell in the cake?
f. Let's say that you buy one of Pierre's cakes each week for six weeks. What is the probability that there will not be any egg shell in any of the cakes?
g. Based upon the average given for Pierre, is it possible for there to be seven pieces of shell in the cake? Why?

S 4.7.11

a. $X =$ the number of shell pieces in one cake
b. 0, 1, 2, 3,...
c. $X \sim P(1.5)$
d. 1.5
e. 0.2231
f. 0.0001
g. Yes

Use the following information to answer the next two exercises: The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is ten. We are interested in the number of times her cats wake her up each week.

Q 4.7.12

In words, the random variable $X =$ _____

a. the number of times Mrs. Plum's cats wake her up each week.
b. the number of times Mrs. Plum's cats wake her up each hour.
c. the number of times Mrs. Plum's cats wake her up each night.
d. the number of times Mrs. Plum's cats wake her up.

Q 4.7.13

Find the probability that her cats will wake her up no more than five times next week.

a. 0.5000
b. 0.9329
c. 0.0378
d. 0.0671

S 4.7.13

d

4.8: Discrete Distribution (Playing Card Experiment)

4.9: Discrete Distribution (Lucky Dice Experiment)

This page titled 5.E: Discrete Random Variables (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **4.E: Discrete Random Variables (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

6: Continuous Random Variables and the Normal Distribution

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

6.1: Introduction

6.1.1: Continuous Probability Functions

6.1.2: The Uniform Distribution

6.2: The Normal Distribution

6.2.1: The Standard Normal Distribution

6.3: Applications of the Normal Distribution

6.4: The Central Limit Theorem

6.5: Normal Approximation to the Binomial Distribution

6.E: The Normal Distribution (Optional Exercises)

1.E: The Central Limit Theorem for Sample Means (Optional Exercises)

1.E: The Standard Normal Distribution (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 6: Continuous Random Variables and the Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.1: Introduction

Note: We won't be studying the **Exponential Distribution** in this course. It is here just to demonstrate that there are many different possible continuous distributions.

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.



Figure 6.1.1: The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve. The curve is called the probability density function (abbreviated as pdf). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution. Area under the curve is given by a different function called the cumulative distribution function(abbreviated as cdf). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability

is also zero.

- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook. There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

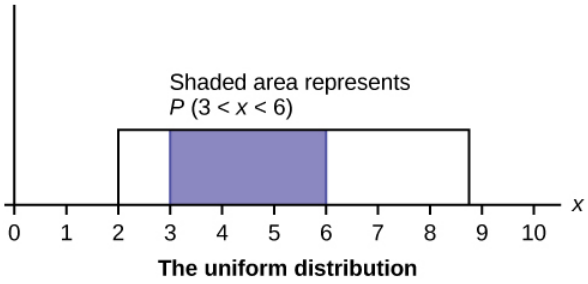


Figure **6.1.2**: The graph shows a Uniform Distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

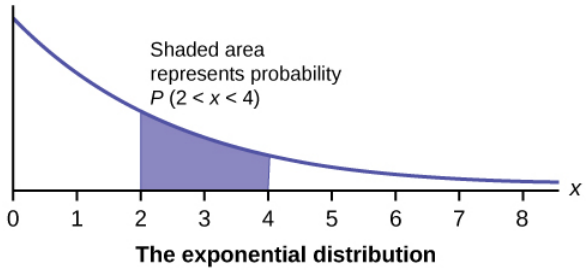


Figure **6.1.3**: The graph shows an Exponential Distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

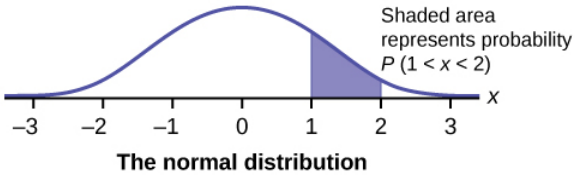


Figure **6.1.4**: The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Glossary

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim \text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and

the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{mx}$.

This page titled 6.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.1.1: Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

Example 6.1.1.1

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. $x =$ a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.



Figure 6.1.1.1

$$f(x) = \frac{1}{20} \text{ for } 0 \leq x \leq 20. \tag{6.1.1.1}$$

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

$$AREA = 20\left(\frac{1}{20}\right) = 1 \tag{6.1.1.2}$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $0 < x < 2$.



Figure 6.1.1.2

$$AREA = (2-0)\left(\frac{1}{20}\right) = 0.1 \tag{6.1.1.3}$$

$$(2-0) = 2 = \text{base of a rectangle}$$

REMINDER: area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $4 < x < 15$.

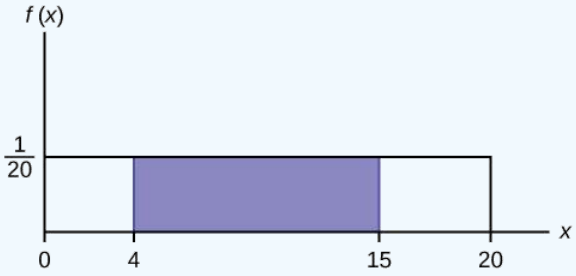


Figure 6.1.1.3

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$(15–4) = 11 = \text{the base of a rectangle} (15–4) = 11 = \text{the base of a rectangle}$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$

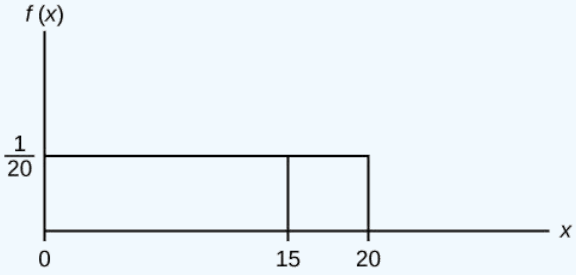


Figure 6.1.1.4

$P(X \leq x)$ (can be written as $P(X < x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1–P(X < x)$.

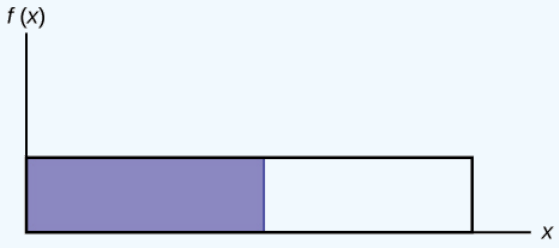


Figure 6.1.1.5

Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.

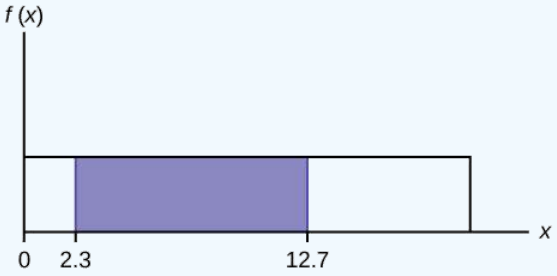


Figure 6.1.1.6

$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52 \quad (6.1.1.4)$$

Exercise 6.1.1.1

Consider the function $f(x) = \frac{1}{8}$ for $0 \le x \le 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

Answer

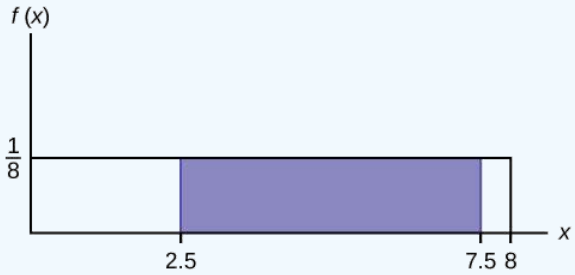


Figure 6.1.1.7

$P(2.5 < x < 7.5) = 0.625$

Summary

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.

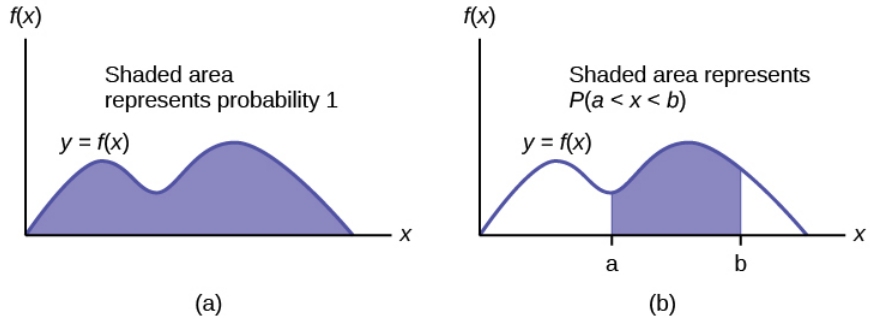


Figure 6.1.1.8

The cumulative distribution function (cdf) of X is defined by $P(X \le x)$. It is a function of x that gives the probability that the random variable is less than or equal to x.

Formula Review

Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

This page titled 6.1.1: Continuous Probability Functions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.1.2: The Uniform Distribution

The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive.

Example 5.3.1

The data in Table **6.1.2.1** are 55 smiling times, in seconds, of an eight-week-old baby.

Table **6.1.2.1**

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is equally likely. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let $X =$ length, in seconds, of an eight-week-old baby's smile.

The notation for the uniform distribution is

$X \sim U(a, b)$ where $a =$ the lowest value of x and $b =$ the highest value of x.

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

For this example, $X \sim U(0, 23)$ and $f(x) = \frac{1}{23-0}$ for $0 \leq X \leq 23$.

Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2}$$

and

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

For this problem, the theoretical mean and standard deviation are

$$\mu = \frac{0+23}{2} = 11.50 \, seconds$$

and

$$\sigma = \frac{(23-0)^2}{12} = 6.64 \, seconds.$$

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

Exercise 6.1.2.1

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are

equally likely. State the values of a and b. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

Table **6.1.2.2**

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

Answer

a is zero; b is 14; $X \sim U(0, 14)$; $\mu = 7$ passengers; $\sigma = 4.04$ passengers

Example 5.3.2A

a. Refer to Example 5.3.1. What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

Answer

a. Find $P(2 < x < 18)$.

$P(2 < x < 18) = (\text{base})(\text{height}) = (18–2)\left(\frac{1}{23}\right) = \left(\frac{16}{23}\right)$.

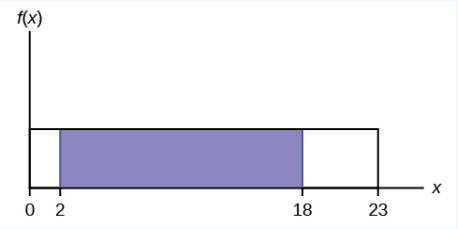


Figure **6.1.2.1** This graph shows a uniform distribution. The horizontal axis ranges from 0 to 15. The distribution is modeled by a rectangle extending from x = 0 to x = 15. A region from x = 2 to x = 18 is shaded inside the rectangle.

Exercise 6.1.2.2B

b. Find the 90th percentile for an eight-week-old baby's smiling time.

Answer

b. Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x < k) = 0.90$

$$P(x < k) = 0.90 \quad (6.1.2.1)$$

$$(\text{base})(\text{height}) = 0.90 \quad (6.1.2.2)$$

$$(k - 0)\left(\frac{1}{23}\right) = 0.90 \quad (6.1.2.3)$$

$$k = (23)(0.90) = 20.7 \quad (6.1.2.4)$$

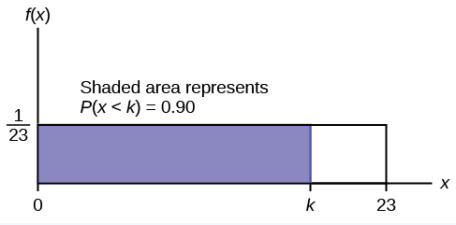


Figure **6.1.2.2** This shows the graph of the function f(x) = 1/15. A horiztonal line ranges from the point (0, 1/15) to the point (15, 1/15). A vertical line extends from the x-axis to the end of the line at point (15, 1/15) creating a rectangle. A region is shaded inside the rectangle from x = 0 to x = k. The shaded area represents P(x < k) = 0.90.

? Exercise 6.1.2.3C

c. Find the probability that a random eight-week-old baby smiles more than 12 seconds **KNOWING** that the baby smiles **MORE THAN EIGHT SECONDS**.

Answer

c. This probability question is a **conditional**. You are asked to find the probability that an eight-week-old baby smiles more than 12 seconds when you **already know** the baby has smiled for more than eight seconds.

Find $P(x > 12|x > 8)$ There are two ways to do the problem. **For the first way**, use the fact that this is a **conditional** and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than eight seconds.

Write a new $f(x)$: $f(x) = \frac{1}{23-8} = \frac{1}{15}$

for $8 < x < 23$

$P(x > 12|x > 8) = (23 - 12)\left(\frac{1}{15}\right) = \left(\frac{11}{15}\right)$

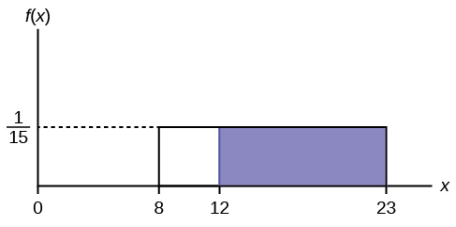


Figure **6.1.2.3** f(X)=1/15 graph displaying a boxed region consisting of a horizontal line extending to the right from point 1/15 on the y-axis, a vertical upward line from points 8 and 23 on the x-axis, and the x-axis. A shaded region from points 12-23 occurs within this area.

For the second way, use the conditional formula from Probability Topics with the original distribution $X \sim U(0, 23)$:

$P(\text{A}|\text{B}) = \frac{P(\text{A AND B})}{P(\text{B})}$

For this problem, **A** is $(x > 12)$ and **B** is $(x > 8)$.

So, $P(x > 12|x > 8) = \frac{(x>12 \text{ AND } x>8)}{P(x>8)} = \frac{P(x>12)}{P(x>8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = \frac{11}{15}$

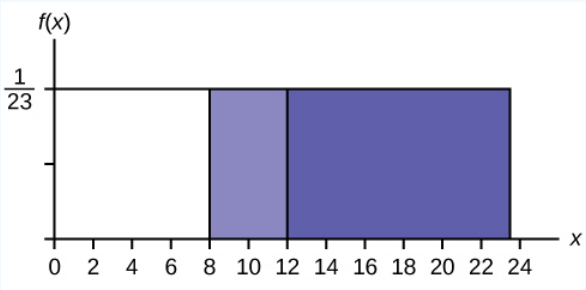


Figure **6.1.2.4**: Darker shaded area represents $P(x > 12)$. Entire shaded area shows $P(x > 8)$.

Exercise 6.1.2.2

A distribution is given as $X \sim U(0, 20)$. What is $P(2 < x < 18)$? Find the 90thpercentile.

Answer

$P(2 < x < 18) = 0.8$; 90th percentile $= 18$

Example 5.3.3

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

Exercise 6.1.2.3.1

a. What is the probability that a person waits fewer than 12.5 minutes?

Answer

a. Let $X =$ the number of minutes a person must wait for a bus. $a = 0$ and $b = 15$. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P(x < 12.5)$. Draw a graph.

$$P(x < k) = (\text{base})(\text{height}) = (12.5 - 0)\left(\frac{1}{15}\right) = 0.8333 \quad (6.1.2.5)$$

The probability a person waits less than 12.5 minutes is 0.8333.

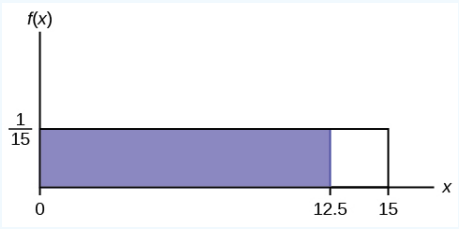


Figure 6.1.2.5. This shows the graph of the function f(x) = 1/15. A horiztonal line ranges from the point (0, 1/15) to the point (15, 1/15). A vertical line extends from the x-axis to the end of the line at point (15, 1/15) creating a rectangle. A region is shaded inside the rectangle from x = 0 to x = 12.5.

Exercise 6.1.2.3.2

b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Answer

b. $\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7.5$. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(12-0)^2}{12}} = 4.3$. The Standard deviation is 4.3 minutes.

Exercise 6.1.2.3.3

c. Ninety percent of the time, the time a person must wait falls below what value?

Note 5.3.3.3.1

This asks for the 90th percentile.

Answer

c. Find the 90th percentile. Draw a graph. Let $k =$ the 90th percentile.

$P(x < k) = (\text{base})(\text{height}) = (k - 0)\left(\frac{1}{15}\right)$
$0.90 = (k)\left(\frac{1}{15}\right)$
$k = (0.90)(15) = 13.5$
k is sometimes called a critical value.
The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.

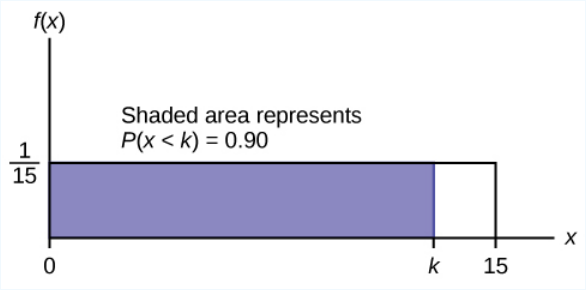


Figure 6.1.2.6.

Exercise 6.1.2.4

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

a. Find a and b and describe what they represent.
b. Write the distribution.
c. Find the mean and the standard deviation.
d. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
e. What is the 65th percentile for the duration of games for a team for the 2011 season?

Answer

a. a is 447, and b is 521. a is the minimum duration of games for a team for the 2011 season, and b is the maximum duration of games for a team for the 2011 season.
b. $X \sim U(447, 521)$.
c. $\mu = 484$, and $\sigma = 21.36$
d. $P(480 < x < 500) = 0.2703$
e. 65th percentile is 495.1 hours.

Example 5.3.4

Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let $X =$ the time, in minutes, it takes a nine-year old child to eat a donut. Then $X \sim U(0.5, 4)$.

a. The probability that a randomly selected nine-year old child eats a donut in at least two minutes is _____.

Solution

a. 0.5714

Exercise 6.1.2.4.1

b. Find the probability that a different nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes.

The second question has a conditional probability. You are asked to find the probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways (see Example). You must reduce the sample space. **First way**: Since you know the child has

already been eating the donut for more than 1.5 minutes, you are no longer starting at a = 0.5 minutes. Your starting point is 1.5 minutes.

Write a new $f(x)$:

$f(x) = \frac{1}{4-1.5} = \frac{2}{5}$ for $1.5 \leq x \leq 4$.

Find $P(x > 2|x > 1.5)$. Draw a graph.

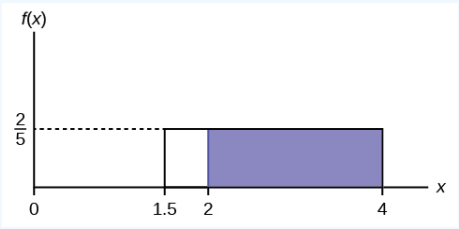


Figure **6.1.2.2**.f(X)=2/5 graph displaying a boxed region consisting of a horizontal line extending to the right from point 2/5 on the y-axis, a vertical upward line from points 1.5 and 4 on the x-axis, and the x-axis. A shaded region from points 2-4 occurs within this area.

$P(x > 2|x > 1.5) = (\text{base})(\text{new height}) = (4-2)(25)\left(\frac{2}{5}\right) = ?$

Answer

b. $\frac{4}{5}$

The probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $X \sim U(0.5, 4)$. Use the conditional formula

$P(x > 2|x > 1.5) = \frac{P(x>2 \text{AND} x>1.5)}{P(x>1.5)} = \frac{P(x>2)}{P(x>1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$

Exercise 6.1.2.5

Suppose the time it takes a student to finish a quiz is uniformly distributed between six and 15 minutes, inclusive. Let X = the time, in minutes, it takes a student to finish a quiz. Then $X \sim U(6, 15)$.

Find the probability that a randomly selected student needs at least eight minutes to complete the quiz. Then find the probability that a different student needs at least eight minutes to finish the quiz given that she has already taken more than seven minutes.

Answer

$P(x > 8) = 0.7778$

$P(x > 8|x > 7) = 0.875$

Example 5.3.5

Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let x = the time needed to fix a furnace. Then $x \sim U(1.5, 4)$.

a. Find the probability that a randomly selected furnace repair requires more than two hours.
b. Find the probability that a randomly selected furnace repair requires less than three hours.
c. Find the 30th percentile of furnace repair times.
d. The longest 25% of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25% of repair times.) What percentile does this represent?
e. Find the mean and standard deviation

Solution

a. To find $f(x)$: $f(x) = \frac{1}{4-1.5} = \frac{1}{2.5}$ so $f(x) = 0.4$

$P(x > 2) = (\text{base})(\text{height}) = (4–2)(0.4) = 0.8$

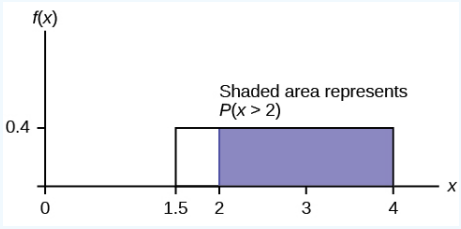


Figure 6.1.2.3. Uniform Distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time x is greater than two

b. $P(x < 3) = (\text{base})(\text{height}) = (3–1.5)(0.4) = 0.6$

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between $x = 1.5$ and $x = 3$. Note that the shaded area starts at $x = 1.5$ rather than at $x = 0$; since $X \sim U(1.5, 4)$, x can not be less than 1.5.

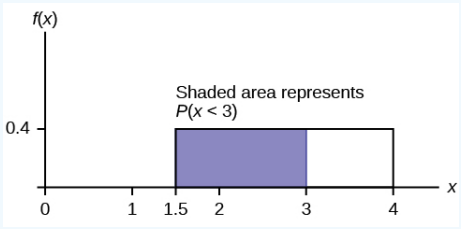


Figure 6.1.2.4. Uniform Distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time x is less than three

c.

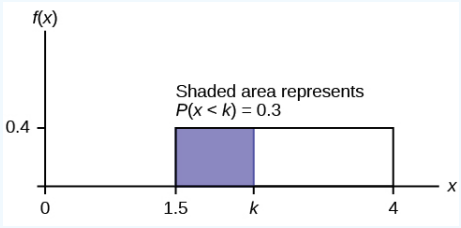


Figure 6.1.2.5. Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

$P(x < k) = 0.30$

$P(x < k) = (\text{base})(\text{height}) = (k–1.5)(0.4)$

$0.3 = (k–1.5)(0.4)$; Solve to find k:

$0.75 = k–1.5$, obtained by dividing both sides by 0.4

$k = 2.25$, obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.25 hours or less.

d.

fig-ch05_03_11N.jpg

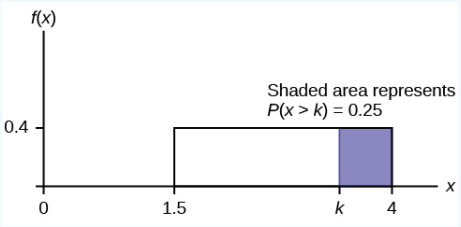


Figure **6.1.2.6**. Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.

$P(x > k) = 0.25$
$P(x > k) = (\text{base})(\text{height}) = (4–k)(0.4)$
$0.25 = (4–k)(0.4)$; Solve for k:
$0.625 = 4 – k$,
obtained by dividing both sides by 0.4
$–3.375 = –k$,
obtained by subtracting four from both sides: $k = 3.375$
The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the 75th **percentile** of furnace repair times.

e. $\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$\mu = \frac{1.5+4}{2} = 2.75$ hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

? Exercise 6.1.2.6

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let $X =$ the time needed to change the oil on a car.

a. Write the random variable X in words. $X =$ _____.
b. Write the distribution.
c. Graph the distribution.
d. Find $P(x > 19)$.
e. Find the 50th percentile.

Answer

a. Let $X =$ the time needed to change the oil in a car.
b. $X \sim U(11, 21)$.
c. This graph shows a uniform distribution. The horizontal axis ranges from 405 to 525. The distribution is modeled by a rectangle extending from x = 447 to x = 521.
d. $P(x > 19) = 0.2$
e. the 50th percentile is 16 minutes.

Review

If X has a uniform distribution where $a < x < b$ or $a \leq x \leq b$, then X takes on values between a and b (may include a and b). All values x are equally likely. We write $X \sim U(a, b)$. The mean of X is $\mu = \frac{a+b}{2}$. The standard deviation of X is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function of X is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$. The cumulative distribution function of X is $P(X \leq x) = \frac{x-a}{b-a}$. X is continuous.

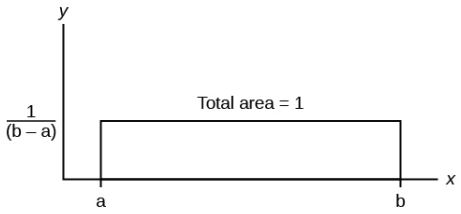


Figure 6.1.2.8.

The probability $P(c < X < d)$ may be found by computing the area under $f(x)$, between c and d. Since the corresponding area is a rectangle, the area may be found simply by multiplying the width and the height.

Formula Review

$X =$ a real number between a and b (in some instances, X can take on the values a and b). $a =$ smallest X; $b =$ largest X

$X \sim U(a, b)$

The mean is $\mu = \frac{a+b}{2}$

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f(x) = \frac{1}{b-a}$ for $a \leq X \leq b$

Area to the Left of x: $P(X < x) = (x–a)\left(\frac{1}{b-a}\right)$

Area to the Right of x: P(X > x) = (b – x)$\left(\frac{1}{b-a}\right)$

Area Between c and d: $P(c < x < d) = (\text{base})(\text{height}) = (d–c)\left(\frac{1}{b-a}\right)$

Uniform: $X \sim U(a, b)$ where $a < x < b$

- pdf: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- cdf: $P(X \leq x) = \frac{x-a}{b-a}$
- mean $\mu = \frac{a+b}{2}$
- standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d–c)\left(\frac{1}{b-a}\right)$

References

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

Use the following information to answer the next ten questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes.

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be written as $X \sim U(1.5, 4.5)$.

? Exercise 6.1.2.7

What type of distribution is this?

Exercise 6.1.2.8

In this distribution, outcomes are equally likely. What does this mean?

Answer

It means that the value of x is just as likely to be any number between 1.5 and 4.5.

Exercise 6.1.2.9

What is the height of $f(x)$ for the continuous probability distribution?

Exercise 6.1.2.10

What are the constraints for the values of x?

Answer

$1.5 \leq x \leq 4.5$

Exercise 6.1.2.11

Graph $P(2 < x < 3)$.

Exercise 6.1.2.12

What is $P(2 < x < 3)$?

Answer

0.3333

Exercise 6.1.2.13

What is $P(x < 3.5 | x < 4)$?

Exercise 6.1.2.14

What is $P(x = 1.5)$?

Answer

zero

Exercise 6.1.2.15

What is the 90th percentile of square footage for homes?

Exercise 6.1.2.16

Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.

Answer

0.6

Exercise 6.1.2.17

What is a? What does it represent?

Exercise 6.1.2.18

What is b? What does it represent?

Answer

b is 12, and it represents the highest value of x.

Exercise 6.1.2.19

What is the probability density function?

Exercise 6.1.2.20

What is the theoretical mean?

Answer

six

Exercise 6.1.2.21

What is the theoretical standard deviation?

Exercise 6.1.2.22

Draw the graph of the distribution for $P(x > 9)$.

Answer

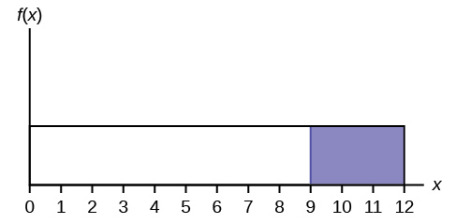


Figure 6.1.2.9.

Exercise 6.1.2.23

Find $P(x > 9)$.

Exercise 6.1.2.24

Find the 40th percentile.

Answer

4.8

Use the following information to answer the next eleven exercises. The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

Exercise 6.1.2.25

What is being measured here?

Exercise 6.1.2.26

In words, define the random variable X.

Answer

X = The age (in years) of cars in the staff parking lot

Exercise 6.1.2.27

Are the data discrete or continuous?

Exercise 6.1.2.28

The interval of values for x is _____.

Answer

0.5 to 9.5

Exercise 6.1.2.29

The distribution for X is _____.

Exercise 6.1.2.30

Write the probability density function.

Answer

$f(x) = \frac{1}{9}$ where x is between 0.5 and 9.5, inclusive.

Exercise 6.1.2.31

Graph the probability distribution.

a. Sketch the graph of the probability distribution.

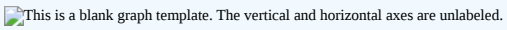


***Figure* 6.1.2.10.**

b. Identify the following values:

i. Lowest value for $\bar{x}$: _____
ii. Highest value for $\bar{x}$: _____
iii. Height of the rectangle: _____
iv. Label for *x*-axis (words): _____
v. Label for *y*-axis (words): _____

Exercise 6.1.2.32

Find the average age of the cars in the lot.

Answer

$\mu = 5$

Exercise 6.1.2.33

Find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, and shade the area of interest.



***Figure* 6.1.2.11.**

b. Find the probability. $P(x < 4) =$ _____

Exercise 6.1.2.34

Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, shade the area of interest.

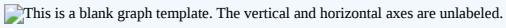


***Figure* 6.1.2.12.**

b. Find the probability. $P(x < 4|x < 7.5) =$ _____

Answer

a. Check student's solution.
b. $\frac{3.5}{7}$

Exercise 6.1.2.35

What has changed in the previous two problems that made the solutions different

Exercise 6.1.2.36

Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75%, of the cars are at most (less than or equal to) that age.

a. Sketch the graph, and shade the area of interest.



***Figure* 6.1.2.13.**

b. Find the value k such that $P(x < k) = 0.75$.
c. The third quartile is _____

Answer

a. Check student's solution.
b. $k = 7.25$
c. 7.25

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

This page titled 6.1.2: The Uniform Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.2: The Normal Distribution

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.



Figure **6.2.1**: If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.2.1}$$

The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

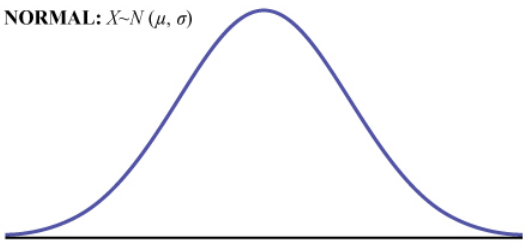


Figure **6.2.2**: The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

COLLABORATIVE CLASSROOM ACTIVITY

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

- $X \sim N(\mu, \sigma)$
- $\mu =$ the mean $\sigma =$ the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x \cdot \mu)^2}{2\sigma^2}} \tag{6.2.2}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

This page titled 6.2: The Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.2.1: The Standard Normal Distribution

Z-Scores

The standard normal distribution is a normal distribution of standardized values called *z-scores*. A z-score is measured in units of the standard deviation.

Definition: Z-Score

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:

$$z = \frac{x - \mu}{\sigma} \tag{6.2.1.1}$$

The *z*-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$\begin{aligned} x &= \mu + (z)(\sigma) \\ &= 5 + (3)(2) = 11 \end{aligned}$$

The *z*-score is three.

Since the mean for the standard normal distribution is zero and the standard deviation is one, then the transformation in Equation 6.2.1.1 produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

> *A z-score is measured in units of the standard deviation.*

Example 6.2.1.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then (via Equation 6.2.1.1):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two** standard deviations (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Exercise 6.2.1.1

What is the z-score of x, when $x = 1$ and $X \sim N(12, 3)$?

Answer

$$z = \frac{1-12}{3} \approx -3.67$$

Example 6.2.1.2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $x = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. This z-score tells you that $x = 10$ is 2.5 standard deviations to the right of the mean five.

b. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$$z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$$

where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of their own) standard deviations to the right of their respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Exercise 6.2.1.2

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is -1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

Answer

1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the *Empirical Rule* says the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are +1 and –1, respectively.
- The z-scores for $+2\sigma$ and -2σ are +2 and –2, respectively.
- The z-scores for $+3\sigma$ and -3σ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.

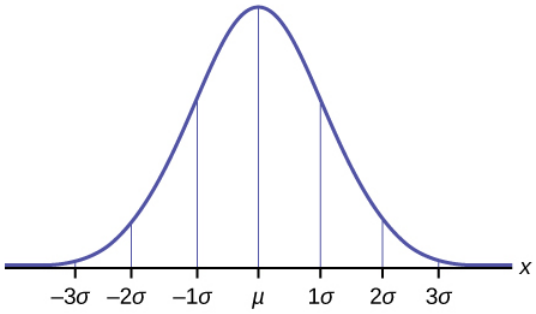


Figure **6.2.1.1**

Example 6.2.1.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when $x = 168$ cm is $z =$ _____. This z-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. –0.32, 0.32, left, 170

b. 177.98, 1.27, right

Exercise 6.2.1.3

Use the information in Example **6.2.1.3** to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z =$ _____. This z-score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = -2$. What is the male's height? The z-score ($z = -2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answer

Solve the equation $z = \frac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$z = \frac{176 - 170}{6.28}$, This z-score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.

Answer

Solve the equation $z = \frac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$X = 157.44$ cm, The z-score($z = -2$) tells you that the male's height is two standard deviations to the left of the mean.

Example 6.2.1.4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Answer

- The z-score (Equation 6.2.1.1) for $x = 160.58$ is $z = -1.5$.
- The z-score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Exercise 6.2.1.4

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let $X =$ a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Answer

The z-score (Equation 6.2.1.1) for $x_1 = 325$ is $z_1 = -1.15$.

The z-score (Equation 6.2.1.1) for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example 6.2.1.5

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie within one standard deviation of the mean. Therefore, about 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation from the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie within two standard deviations of the mean. Therefore, about 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations from the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie within three standard deviations of the mean. Therefore, about 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ from the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Exercise 6.2.1.5

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Answer

between 20 and 30.

Example 6.2.1.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____ respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.
b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Exercise 6.2.1.6

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer a

About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.

Answer b

About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.

Answer c

About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

Summary

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

$Z \sim N(0, 1)$

$z = a$ standardized value (z-score)

mean = 0; standard deviation = 1

To find the Kth percentile of X when the z-scores is known:

$k = \mu + (z)\sigma$

z-score: $z = \frac{x - \mu}{\sigma}$

$Z =$ the random variable for z-scores

$Z \sim N(0, 1)$

Glossary

Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as \(Z \sim N(0, 1)\.

z-score

the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

References

1. “Blood Pressure of Males and Females.” StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...reportid=11960 (accessed May 14, 2013).
2. “The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores.” London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).
3. “2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at media.collegeboard.com/digita...Group-2012.pdf (accessed May 14, 2013).
4. “Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at nces.ed.gov/programs/digest/d...s/dt09_147.asp (accessed May 14, 2013).
5. Data from the *San Jose Mercury News*.
6. Data from *The World Almanac and Book of Facts*.
7. “List of stadiums by capacity.” Wikipedia. Available online at en.Wikipedia.org/wiki/List_o...ms_by_capacity (accessed May 14, 2013).
8. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

This page titled 6.2.1: The Standard Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **6.2: The Standard Normal Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

6.3: Applications of the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.

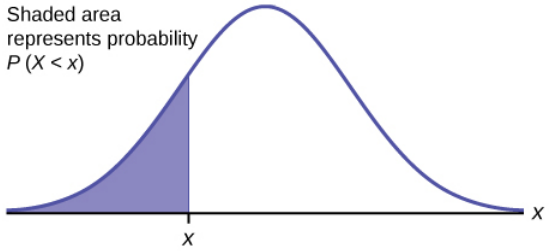


Figure 6.3.1.

The area to the right is then $P(X > x) = 1–P(X < x)$. Remember, $P(X < x) =$ **Area to the left** of the vertical line through x. $P(X > x) = 1–P(X < x) =$ **Area to the right** of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.To calculate the probability, use the probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

Example 6.3.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Exercise 6.3.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer

$1 - 0.012 = 0.988$

Example 6.3.2

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.
c. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Answer

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$$P(x > 65) = 0.3446$$

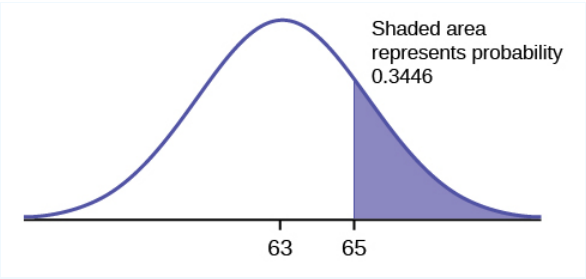


Figure **6.3.2**.

The probability that any student selected at random scores more than 65 is 0.3446.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into `2nd DISTR` .

After pressing `2nd DISTR` , press `2:normalcdf` .

The syntax for the instructions are as follows:

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1` , the `EE` key (a 2nd key) and then `99` . Or, you can enter `10^ 99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Historical Note

The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65-63}{5} = 0.4$$

Area to the left is 0.6554.

$$P(x > 65) = P(z > 0.4) = 1-0.6554 = 0.3446$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Find the percentile for a student scoring 65:

*Press `2nd Distr`
*Press `2:normalcdf` (
*Enter lower bound, upper bound, mean, standard deviation followed by)
*Press `ENTER` .
For this Example, the steps are
`2nd Distr`
`2:normalcdf` (65,1,2nd EE,99,63,5) `ENTER`

The probability that a selected student scored more than 65 is 0.3446.
To find the probability that a selected student scored *more than* 65, subtract the percentile from 1.

Answer

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

$\text{normalcdf}(0, 85, 63, 5) = 1$ (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Answer

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let $k =$ the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a critical value.

$k = 69.4$

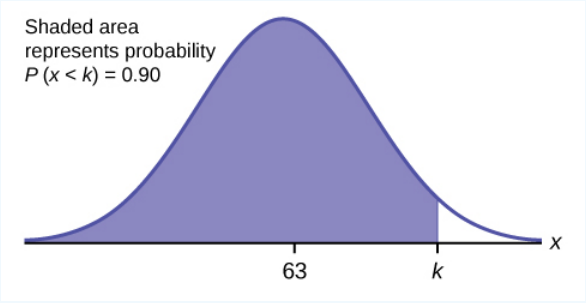


Figure 6.3.3.

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

`invNorm` in `2nd DISTR` . invNorm(area to the left, mean, standard deviation)

For this problem, $\text{invNorm}(0.90, 63, 5) = 69.4$

Answer

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.6 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

? Exercise 6.3.2

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a randomly selected golfer scored less than 65.

Answer

$\text{normalcdf}(10^{99}, 65, 68, 3) = 0.1587$

✓ Example 6.3.3

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is

used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

a. Let $X =$ the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

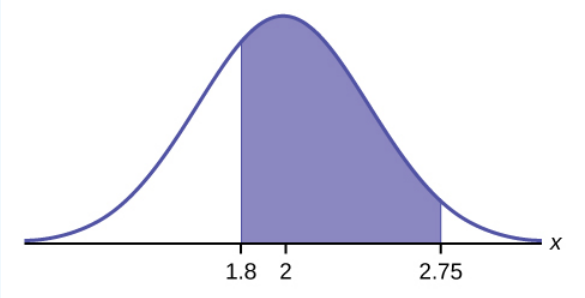


Figure 6.3.4.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b.

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.

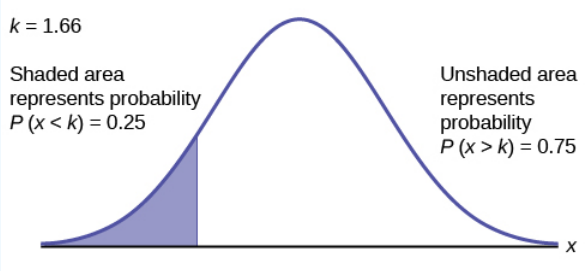


Figure 6.3.5.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Exercise 6.3.3

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Answer

$$\text{normalcdf}(66, 70, 68, 3) = 0.4950$$

Example 6.3.4

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Answer

a. $\textbf{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$
b. $\textbf{normalcdf}(-10^{99}, 50.8, 36.9, 13.9) = 0.8413$
c. $\textbf{invNorm}(0.80, 36.9, 13.9) = 48.6$

The 80th percentile is 48.6 years.

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Use the information in Example to answer the following questions.

Exercise 6.3.4

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old and at least 0 years old?

70.

Answer

Let $X =$ a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

To find the 30th percentile, find k such that $P(x < k) = 0.30$.
$\textbf{invNorm}(0.30, 36.9, 13.9) = 29.6$ years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years. Find $P(x < 27)$
(Note that $\textbf{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$. The two answers differ only by 0.0040.)

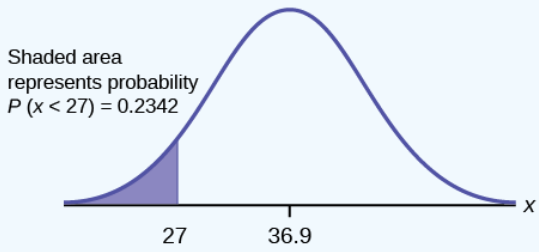


Figure 6.3.6.

$\textbf{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$

Example 6.3.5

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information, answer the following questions (round answers to one decimal place).

a. Calculate the interquartile range (IQR).
b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Answer

a.

$$IQR = Q_3 - Q_1$$

Calculate $Q_3 = 75^{\text{th}}$ percentile and $Q_1 = 25^{\text{th}}$ percentile.

$$\text{invNorm}(0.75, 36.9, 13.9) = Q_3 = 46.2754$$
$$\text{invNorm}(0.25, 36.9, 13.9) = Q_1 = 27.5246$$
$$IQR = Q_3 - Q_1 = 18.7508$$

b.

Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")

$0.40 =$ the area to the right.

Area to the left $= 1 - 0.40 = 0.60$.

The area to the left of $k = 0.60$.

$\text{invNorm}(0.60, 36.9, 13.9) = 40.4215$.

$k = 40.42$.

Forty percent of the smartphone users from 13 to 55+ are at least 40.4 years.

Exercise 6.3.5

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Answer

a. $Q_1 = 25^{\text{th}}$ percentile $= \text{invNorm}(0.25, 81, 15) = 70.9$
   $Q_3 = 75^{\text{th}}$ percentile $= \text{invNorm}(0.75, 81, 15) = 91.1$
b. The middle 50% of the scores are between 70.9 and 91.1.

Example 6.3.6

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
c. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$

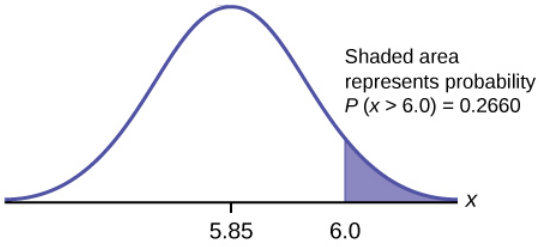


Figure 6.3.7.

Answer

b.

$1–0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

Find $k1$, the 40th percentile, and $k2$, the 60th percentile ($0.40 + 0.20 = 0.60$).

$k1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79$ cm

$k2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91$ cm

Answer

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Exercise 6.3.6

Using the information from Example, answer the following:

a. The middle 45% of mandarin oranges from this farm are between _____ and _____.
b. Find the 16th percentile and interpret it in a complete sentence.

Answer a

The middle area $= 0.40$, so each tail has an area of 0.30.

$–0.40 = 0.60$

The tails of the graph of the normal distribution each have an area of 0.30.

Find $k1$, the 30th percentile and $k2$, the 70th percentile ($0.40 + 0.30 = 0.70$).

$k1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72$ cm

$k2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98$ cm

Answer b

$\text{normalcdf}(5, 10^{99}, 5.85, 0.24) = 0.9998$

References

1. “Naegele’s rule.” Wikipedia. Available online at http://en.Wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).
2. “403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).
3. “Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at www.winatthelottery.com/publi...partment40.cfm (accessed May 14, 2013).
4. “Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).
5. “Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebo...tics/(accessed May 14, 2013).

Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

- Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.
- Standard Normal Distribution: $Z \sim N(0, 1)$.
- Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)
- Calculator function for the kth percentile: $k = \text{invNorm}$ (area to the left of k, mean, standard deviation)

? Exercise 6.3.7

How would you represent the area to the left of one in a probability statement?

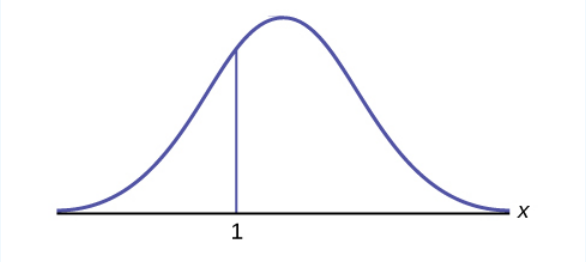


Figure 6.3.8.

Answer

$P(x < 1)$

? Exercise 6.3.8

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

? Exercise 6.3.9

How would you represent the area to the left of three in a probability statement?

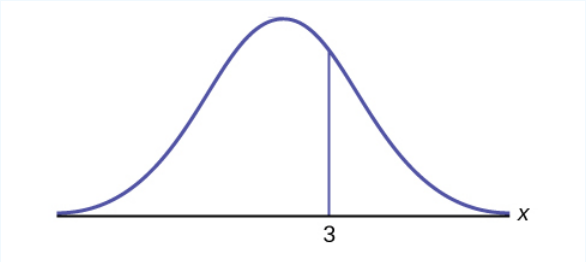


Figure 6.3.10.

Exercise 6.3.10

What is the area to the right of three?

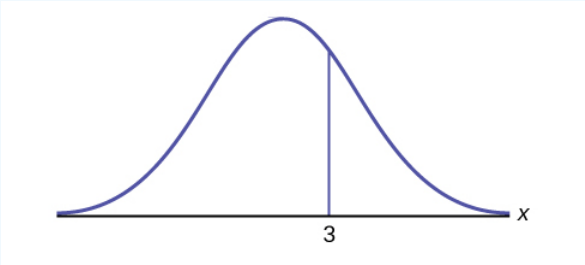


Figure 6.3.11.

Answer

$1–P(x < 3)$ or $P(x > 3)$

Exercise 6.3.11

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise 6.3.12

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Answer

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise 6.3.13

Find the probability that $x > 56$.

Exercise 6.3.14

Find the probability that $x < 30$.

Answer

0.0013

Exercise 6.3.15

Find the 80th percentile.

Exercise 6.3.16

Find the 60th percentile.

Answer

56.03

Exercise 6.3.17

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise 6.3.18

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Answer

0.1186

Exercise 6.3.19

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Exercise 6.3.20

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

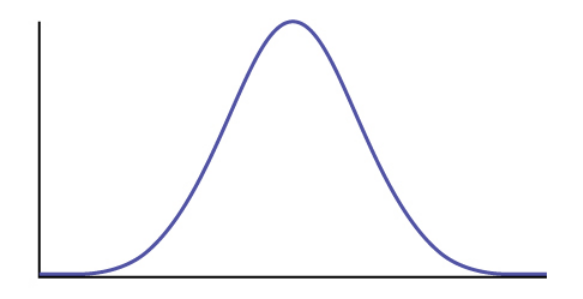


***Figure* 6.3.12.**

$P(0 < x <$ _____$) =$ _____ (Use zero for the minimum value of x.)

Answer

a. Check student's solution.
b. 3, 0.1979

Exercise 6.3.21

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

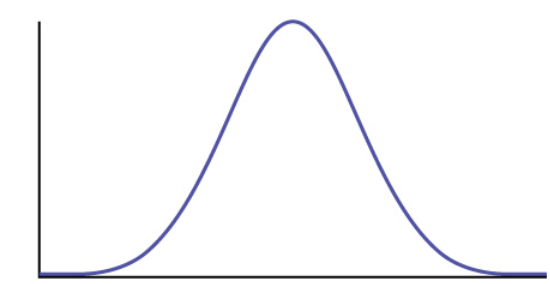


***Figure* 6.3.13.**

$P($_____ $< x <$ _____$)$ = _____

Exercise 6.3.22

Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.



***Figure* 6.3.14.**

$P(x < k) =$ _____ Therefore, $k =$ _____

Answer

a. Check student's solution.
b. 0.70, 4.78 years

This page titled 6.3: Applications of the Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **6.3: Using the Normal Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

6.4: The Central Limit Theorem

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. $\mu_x =$ the mean of X
b. $\sigma_x =$ the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\bar{X}$ which consists of sample means, tends to be normally distributed and

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right). \tag{6.4.1}$$

The central limit theorem for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, the sample means form their own *normal distribution* (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by, the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\bar{X}$, which consists of sample means, is called the *sampling distribution of the mean*. The sampling distribution of the mean approaches a normal distribution as n, the sample size, increases.

The random variable $\bar{X}$ has a different z-score associated with it from that of the random variable X. The mean $\bar{x}$ is the value of $\bar{X}$ in one sample.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)} \tag{6.4.2}$$

- μ_x is the average of both X and $\bar{X}$.
- $\sigma\bar{x} = \frac{\sigma_x}{\sqrt{n}} =$ standard deviation of $\bar{X}$ and is called the standard error of the mean.

Howto: Find probabilities for means on the calculator

2nd DISTR

2:normalcdf

$$\text{normalcdf}\left(\text{lower value of the area, upper value of the area, mean, } \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$$

where:

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* $= n$

Example 6.4.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

a. Find the probability that the sample mean is between 85 and 92.
b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Answer

a.

Let $X =$ one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\bar{X} =$ the mean of a sample of size 25. Since $\mu_x = 90, \sigma_x = 15$, and $n = 25$,

$$\bar{X} \sim N(90, \frac{15}{\sqrt{25}}).$$

Find $P(85 < x < 92)$. Draw a graph.

$$P(85 < x < 92) = 0.6997$$

The probability that the sample mean is between 85 and 92 is 0.6997.

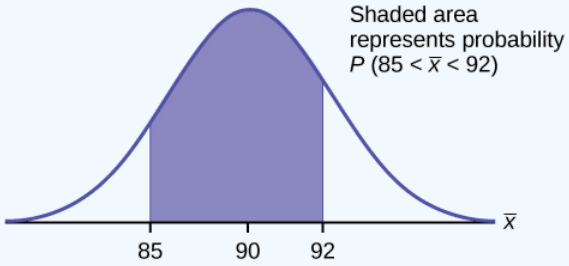


Figure 6.4.1.

`normalcdf` (lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf` $(85, 92, 90, \frac{15}{\sqrt{25}}) = 0.6997$

b.

To find the value that is two standard deviations above the expected value 90, use the formula:

$$\text{value} = \mu_x + (\#\text{ofTSDEVs})\left(\frac{\sigma_x}{\sqrt{n}}\right)$$

$$= 90 + 2\left(\frac{15}{\sqrt{25}}\right) = 96$$

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is

$$\frac{\sigma_x}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3.$$

Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

Exercise 6.4.1

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size n = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Answer

$$P(42 < \bar{x} < 50) = \left(42, 50, 45, \frac{8}{\sqrt{30}}\right) = 0.9797$$

Example 6.4.2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size** $n = 50$ is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Answer

Let $X =$ the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let $\bar{X} =$ the mean time, in hours, it takes to play one soccer match.

If $\mu_x =$ _____, $\sigma_x =$ _____, and $n =$ _____, then $X \sim N$(_____, _____) by the central limit theorem for means.

$\mu_x = 2, \sigma_x = 0.5, n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$P(1.8 < \bar{x} < 2.3) = 0.9977$

normalcdf $\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Exercise 6.4.2

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Answer

$$P(2 < \bar{x} < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$$

Calculator SKills

To find percentiles for means on the calculator, follow these steps.

- 2nd DIStR
- 3:invNorm

$$k = \text{invNorm}\left(\text{area to the left of} k, \text{mean}, \frac{\text{standard deviation}}{\sqrt{samplesize}}\right)$$

where:

- k = the kth percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

Example 6.4.3

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

Answer

a. Since the sample mean tends to target the population mean, we have $\mu_x = \mu = 34$. The sample standard deviation is given by:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

b. The central limit theorem states that for large sample sizes (n), the sampling distribution will be approximately normal.
c. The probability that the sample mean age is more than 30 is given by:

$$P(\mathrm{X} > 30) = \mathrm{normalcdf}(30, E99, 34, 1.5) = 0.9962$$

d. Let k = the 95th percentile.

$$k = \mathrm{invNorm}\left(0.95, 34, \frac{15}{\sqrt{100}}\right) = 36.5$$

Exercise 6.4.3

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Answer

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$$P(29 < \bar{x} < 35) = \mathrm{normalcdf}\left(29, 35, 28, \frac{4.8}{\sqrt{100}}\right) = 0.0186 \quad (6.4.3)$$

You can conclude there is approximately a 1.9% chance that your game will be played by men whose mean age is between 29 and 35.

Example 6.4.4

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Answer

a. $\mu = \mu = 8.2 \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$

b. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.

c. Let k = the 90th percentile

$k = \text{invNorm}\left(0.90, 8.2, \frac{1}{\sqrt{60}}\right) = 8.37$. This values indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.

d. $P(8 < \bar{x} < 8.5) = \text{normalcdf}\left(8, 8.5, 8.2, \frac{1}{\sqrt{60}}\right) = 0.9293$

Exercise 6.4.4

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\bar{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Answer

We have $P(\bar{x} > 16.01) = \text{normalcdf}\left(16.01, E99, 16, \frac{0.143}{\sqrt{34}}\right) = 0.3417$. Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

Summary

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

Formula Review

- The Central Limit Theorem for Sample Means:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

- The Mean $\bar{X}$: σ_x
- Central Limit Theorem for Sample Means z-score and standard error of the mean:

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

- Standard Error of the Mean (Standard Deviation ($\bar{X}$)):

$$\frac{\sigma_x}{\sqrt{n}}$$

Glossary

Average

a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called a **standard normal distribution**.

Standard Error of the Mean

the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

References

1. Baran, Daya. “20 Percent of Americans Have Never Used Email.”WebGuild, 2010. Available online at www.webguild.org/20080519/20-...ver-used-email (accessed May 17, 2013).
2. Data from The Flurry Blog, 2013. Available online at blog.flurry.com (accessed May 17, 2013).
3. Data from the United States Department of Agriculture.

This page titled 6.4: The Central Limit Theorem is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.5: Normal Approximation to the Binomial Distribution

Normal Approximation to the Binomial Distribution

Historical Note: Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n(say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n, you had to use the binomial formula, which could be very complicated. Using the normal approximation to the binomial distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a binomial distribution:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$); the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor and is used in the following example.

Example 6.5.5

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that **at least 150** favor a charter school.
b. Find the probability that **at most 160** favor a charter school.
c. Find the probability that **more than 155** favor a charter school.
d. Find the probability that **fewer than 147** favor a charter school.
e. Find the probability that **exactly 175** favor a charter school.

Let $X =$ the number that favor a charter school for grades K trough 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is X. $Y \sim N(159, 8.6447)$. See The Normal Distribution for help with calculator instructions.

For part a, you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

normalcdf $(149.5, 10^{99}, 159, 8.6447) = 0.8641$.

For part b, you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

normalcdf $(0, 160.5, 159, 8.6447) = 0.5689$

For part c, you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$.

normalcdf $(155.5, 10^{99}, 159, 8.6447) = 0.6572$.

For part d, you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

normalcdf $(0, 146.5, 159, 8.6447) = 0.0741$

For part e, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

normalcdf $(174.5, 175.5, 159, 8.6447) = 0.0083$

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial.

For Example, the probabilities are calculated using the following binomial distribution: $(n = 300 and p = 0.53)$. Compare the binomial and normal distribution answers. See Discrete Random Variables for help with calculator instructions for the binomial.

$P(X \geq 150)$: `1 - binomialcdf` $(300, 0.53, 149) = 0.8641$

$P(X \leq 160)$: `binomialcdf` $(300, 0.53, 160) = 0.5684$

$P(X > 155)$: `1 - binomialcdf` $(300, 0.53, 155) = 0.6576$

$P(X < 147)$: `binomialcdf` $(300, 0.53, 146) = 0.0742$

$P(X = 175)$:(You use the binomial pdf.) `binomialpdf` $(300, 0.53, 175) = 0.0083$

Exercise 6.5.5

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

Answer

0.0401

References

- Data from the Wall Street Journal.
- “National Health and Nutrition Examination Survey.” Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Glossary

Exponential Distribution

a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital, notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

Uniform Distribution

a continuous random variable (RV) that has equally likely outcomes over the domain, \(a < x < b\); often referred as the **Rectangular Distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{a+b}{2}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

This page titled 6.5: Normal Approximation to the Binomial Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.E: The Normal Distribution (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

6.1: Introduction

6.2: The Standard Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Q 6.2.1

What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

Q 6.2.2

What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

S 6.2.2

c

Q 6.2.3

The length of time to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution..
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

Q 6.2.4

The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, μ = 79 inches and a standard deviation, σ = 3.89 inches. For each of the following heights, calculate the *z*-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a *z*-score of 3.5, would you believe him? Explain your answer.

S 6.2.4

a. Use the z-score formula. $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.

b. Use the z-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
c. Height $= 79 + 3.5(3.89) = 90.67$ inches, which is over 7.7 feet tall. There are very few NBA players this tall so the answer is no, not likely.

Q 6.2.5

The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

Q 6.2.6

Kyle's doctor told him that the z-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If $X =$ a systolic blood pressure score then $X \sim N(125, 14)$.

a. Which answer(s) **is/are** correct?
   i. Kyle's systolic blood pressure is 175.
   ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
b. Calculate Kyle's blood pressure.

S 6.2.6

a. iv
b. Kyle's blood pressure is equal to $125 + (1.75)(14) = 149.5$.

Q 6.2.7

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

Q 6.2.8

In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

a. Calculate the z-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

S 6.2.8

Let $X =$ an SAT math score and $Y =$ an ACT math score.

a. $X = 720 \frac{720-520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.

b. $z = 1.5$
The math SAT score is $520 + 1.5(115) \approx 692.5$. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
c. $\frac{X-\mu}{\sigma} = \frac{700-514}{117} \approx 1.59$, the *z*-score for the SAT. $\frac{Y-\mu}{\sigma} = \frac{30-21}{5.3} \approx 1.70$, the *z*-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher *z*-score).

6.3: Using the Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Q 6.3.1

What is the probability of spending more than two days in recovery?

a. 0.0580
b. 0.8447
c. 0.0553
d. 0.9420

Q 6.3.2

The 90th percentile for recovery times is?

a. 8.89
b. 7.07
c. 7.99
d. 4.32

S 6.3.2

c

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

Q 6.3.3

Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

a. Yes
b. No
c. Unable to determine

Q 6.3.4

Find the probability that it takes at least eight minutes to find a parking space.

a. 0.0001
b. 0.9270
c. 0.1862
d. 0.0668

S 6.3.4

d

Q 6.3.5

Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41

c. 3.95
d. 6.05

Q 6.3.6

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let $X =$ height of the individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
d. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

S 6.3.6

a. $X \sim N(66, 2.5)$
b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

Q 6.3.7

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let $X =$ IQ of an individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Q 6.3.8

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let $X =$ percent of fat calories.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

S 6.3.8

a. $X \sim N(36, 10)$
b. The probability that a person consumes more than 40% of their calories as fat is 0.3446.
c. Approximately 25% of people consume less than 29.26% of their calories as fat.

Q 6.3.9

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If $X =$ distance in feet for a fly ball, then $X \sim$ _____(_____,_____)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

Q 6.3.10

In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

S 6.3.10

a. $X =$ number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. $X\ N(3, 1.5)$
c. The probability that the child spends less than one hour a day unsupervised is 0.0918.
d. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
e. 2.21 hours

Q 6.3.11

In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let $X =$ number of votes for President Clinton for an election district.

a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
e. Find the third quartile for votes for President Clinton.

Q 6.3.12

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

S 6.3.12

a. $X =$ the distribution of the number of days a particular type of criminal trial will take
b. $X \sim N(21, 7)$
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

Q 6.3.13

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)

c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under _____.
e. The middle 80% of her laps are from _____ seconds to _____ seconds.

Q 6.3.14

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X = time in line. Table displays the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

S 6.3.14

a. $\text{mean} = 5.51$, $s = 2.15$
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. $X \sim N(5.51, 2.15)$
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

Q 6.3.15

Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his *z*-score is zero.
c. Anita is in the 70th percentile of students at her college.

Q 6.3.16

Table shows a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

a. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____).
f. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
g. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
h. Why aren't the answers to part f and part g exactly the same?

S 6.3.16

a. $\text{mean} = 60,136$, $s = 10,468$
b. Answers will vary.
c. Answers will vary.
d. Answers will vary.
e. $X \sim N(60136, 10468)$
f. 0.7440
g. The cumulative relative frequency is $\frac{43}{60} = 0.717$.
h. The answers for part f and part g are not the same, because the normal distribution is only an approximation.

https://stats.libretexts.org/@go/page/10949

Q 6.3.17

An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the *z*-scores first, and then use those to calculate the probability.

Q 6.3.18

A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of n = 100 cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

S 6.3.18

- $n = 100; p = 0.1; q = 0.9$
- $\mu = np = (100)(0.10) = 10$
- $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

a. $z = \pm : x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x_2 = \mu = z\sigma = 10 - 1(3) = 7.68$ of the defective cars will fall between seven and 13.
b. $z = \pm : x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x_2 = \mu = z\sigma = 10 - 2(3) = 4.95$ of the defective cars will fall between four and 16
c. $z = \pm : x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x_2 = \mu = z\sigma = 10 - 3(3) = 1.997$ of the defective cars will fall between one and 19.

Q 6.3.19

We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

Q 6.3.20

A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

S 6.3.21

- $n = 190; p = 1515 = 0.2; q = 0.8$
- $\mu = np = (190)(0.2) = 38$
- $\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)} = 5.5136$

a. For this problem: $P(34 < x < 54) = \text{normalcdf}(34, 54, 48, 5.5136) = 0.7641$
b. For this problem: $P(54 < x < 64) = \text{normalcdf}(54, 64, 48, 5.5136) = 0.0018$
c. For this problem: $P(x > 64) = \text{normalcdf}(64, 10^{99}, 48, 5.5136) = 0.0000012$ (approximately 0)

Q 6.3.22

Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

a. Find the probability that the percent of 18 to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
b. Find the 95th percentile, and express it in a sentence.

6.4: Normal Distribution (Lap Times)

6.5: Normal Distribution (Pinkie Length)

This page titled 6.E: The Normal Distribution (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **6.E: The Normal Distribution (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

1.E: The Central Limit Theorem for Sample Means (Optional Exercises)

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\bar{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

? Example 1.E.1

What is the mean, standard deviation, and sample size?

Answer

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

Exercise 1.E.2

Complete the distributions.

1. $X \sim$ _____(_____,_____)
2. $\bar{X} \sim$ _____(_____,_____)

? Example 1.E.3

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



***Figure* 1.E.2.**

2. P(_____ $< x <$ _____) = _____

Answer

1. Check student's solution.
2. 3.5, 4.25, 0.2441

Exercise 1.E.4

Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



***Figure* 1.E.3.**

2. P(_____) = _____

Example 1.E.5

What causes the probabilities in Exercise and Exercise to be different?

Answer

The fact that the two distributions are different accounts for the different probabilities.

Exercise 1.E.6

Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.



***Figure* 1.E.4.**

a. The 95th Percentile =_____

1.E: The Central Limit Theorem for Sample Means (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: The Standard Normal Distribution (Optional Exercises)

Exercise 1.E.7

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ _____.

Answer

ounces of water in a bottle

Exercise 1.E.8

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise 1.E.9

$X \sim N(1, 2)$

$\sigma =$ _____

Answer

2

Exercise 1.E.10

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ _____.

Exercise 1.E.11

$X \sim N(-4, 1)$

What is the median?

Answer

–4

Exercise 1.E.12

$X \sim N(3, 5)$

$\sigma =$ _____

Exercise 1.E.13

$X \sim N(-2, 1)$

$\mu =$ _____

Answer

–2

Exercise 1.E.14

What does a z-score measure?

Exercise 1.E.15

What does standardizing a normal distribution do to the mean?

Answer

The mean becomes zero.

Exercise 1.E.16

Is $X \sim N(0,1)$ a standardized normal distribution? Why or why not?

Exercise 1.E.17

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Answer

$z = 2$

Exercise 1.E.18

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Exercise 1.E.19

What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

Answer

$z = 2.78$

Exercise 1.E.20

What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

Exercise 1.E.21

Suppose $X \sim N(2,6)$. What value of x has a z-score of three?

Answer

$x = 20$

Exercise 1.E.22

Suppose $X \sim N(8,1)$. What value of x has a z-score of –2.25?

Exercise 1.E.23

Suppose $X \sim N(9,5)$. What value of x has a z-score of –0.5?

Answer

$x = 6.5$

Exercise 1.E.24

Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

Exercise 1.E.25

Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

Answer

$x = 1$

Exercise 1.E.26

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Exercise 1.E.27

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Answer

$x = 1.97$

Exercise 1.E.28

Suppose $X \sim N(-1, 12)$. What is the z-score of $x = 2$?

Exercise 1.E.29

Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Answer

$z = –1.67$

Exercise 1.E.30

Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

Exercise 1.E.31

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Answer

$z \approx –0.33$

Exercise 1.E.32

In a normal distribution, $x = 5$ and $z = –1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 1.E.33

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

0.67, right

Exercise 1.E.34

In a normal distribution, $x = -2$ and $z = 6$. This tells you that $z = -2$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 1.E.35

In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

3.14, left

Exercise 1.E.36

In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 1.E.37

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Answer

about 68%

Exercise 1.E.38

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise 1.E.39

About what percent of x values lie between the second and third standard deviations (both sides)?

Answer

about 4%

Exercise 1.E.40

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise 1.E.41

Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution (i.e., –3).

Answer

between –5 and –1

Exercise 1.E.42

Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

Exercise 1.E.43

About what percent of x values lie between the mean and three standard deviations?

Answer

about 50%

Exercise 1.E.44

About what percent of x values lie between the mean and one standard deviation?

Exercise 1.E.45

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Answer

about 27%

Exercise 1.E.46

About what percent of x values lie between the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

Exercise 1.E.47

Define the random variable X in words. $X =$ _____.

Answer

The lifetime of a Sunshine CD player measured in years.

Exercise 1.E.48

$X \sim$ _____(_____,_____)

1.E: The Standard Normal Distribution (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

7: Confidence Intervals and Sample Size

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

7.1: Confidence Intervals

7.2: Confidence Intervals for the Mean with Known Standard Deviation

7.3: Confidence Intervals for the Mean with Unknown Standard Deviation

7.4: Confidence Intervals and Sample Size for Proportions

7.5: Confidence Intervals (Summary)

7.E: Confidence Intervals (Optional Exercises)

1.E: Confidence Intervals for the Mean with Known Standard Deviation (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 7: Confidence Intervals and Sample Size is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

7.1: Confidence Intervals

Learning Objectives

By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's *t* distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.



Figure **7.1.1**. Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\bar{x}$ and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \quad (7.1.1)$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \quad (7.1.2)$$

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \quad (7.1.3)$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

> Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval $\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

---

This page titled 7.1: Confidence Intervals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.1: Prelude to Confidence Intervals** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

7.2: Confidence Intervals for the Mean with Known Standard Deviation

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated *EBM*). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

Example 7.2.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5) and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Exercise 7.2.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

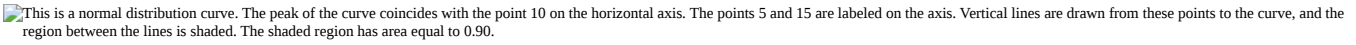


Figure 7.2.1

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_x, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0,1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

`invNorm` $(0.975, 0, 1) = 1.96$

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0,1)$.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

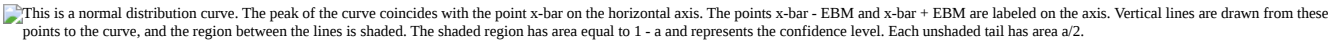

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 1 - a and represents the confidence level. Each unshaded tail has area a/2.

Figure 8.2.2.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 7.2.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% (CL = 0.90)

$$CL = 0.90$$

so

$$\alpha = 1–CL = 1–0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225).**

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `7:ZInterval` .
Press `ENTER` .
Arrow to `Stats` and press `ENTER` .
Arrow down and enter three for σ, 68 for $\bar{x}$, 36 for n, and .90 for `C-level` .
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Exercise 7.2.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.1347, 37.8653)

Example 7.2.3: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

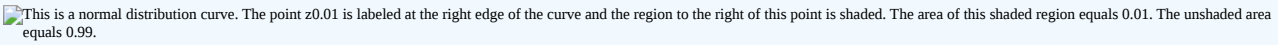

This is a normal distribution curve. The point z0.01 is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$\bar{x} - EBM = 1.024–0.1431 = 0.8809$

$\bar{x} - EBM = 1.024–0.1431 = 1.1671$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:Z Interval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:

- σ : 0.337
- $\bar{x}$: 1024
- n : 30
- C-level: 0.98

- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

Exercise 7.2.3

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 7.2.4

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3; n = 36$; The confidence level is 95% (CL = 0.95).

$CL = 0.95$ so $\alpha = 1–CL = 1–0.95 = 0.05$

$$\frac{\alpha}{2} = 0.025 z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1–0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x}–EBM = 68–0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

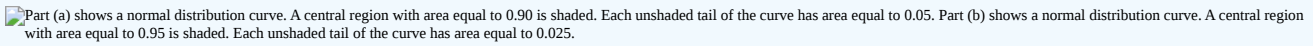

Part (a) shows a normal distribution curve. A central region with area equal to 0.90 is shaded. Each unshaded tail of the curve has area equal to 0.05. Part (b) shows a normal distribution curve. A central region with area equal to 0.95 is shaded. Each unshaded tail of the curve has area equal to 0.025.

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

? Exercise 7.2.4

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Answer

(33.37, 38.63)

✔ Example 7.2.5

Suppose we change the original problem in Example to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Answer

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

? Exercise 7.2.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 7.2.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is $68: EBM = 68.82–68 = 0.82$.
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82–0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Exercise 7.2.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z\frac{a}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n. In Equation , z is $z\frac{a}{2}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 7.2.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.

- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

? Exercise 7.2.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer

35 students

References

1. "American Fact Finder." U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?refresh=t (accessed July 2, 2013).
2. "Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2, 2013).
3. "Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at research.fhda.edu/factbook/FH...phicTrends.htm (accessed September 30,2013).
4. Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at www.cdc.gov/growthcharts/2000...thchart-us.pdf (accessed July 2, 2013).
5. La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).
6. "Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...prodType=table (accessed July 2, 2013).
7. "Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at www.fec.gov/finance/disclosur...esummary.shtml (accessed July 2, 2013).
8. "National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

Glossary

Confidence Level (CL)

the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the $CL = 90$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

This page titled 7.2: Confidence Intervals for the Mean with Known Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: A Single Population Mean using the Normal Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

7.3: Confidence Intervals for the Mean with Unknown Standard Deviation

In practice, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ. If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}, \tag{7.3.1}$$

then the t-scores follow a Student's t-distribution with n–1 degrees of freedom. The t-score has the same interpretation as the z-score. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The degrees of freedom, n–1, come from the calculation of the sample standard deviation s. Previously, we used n deviations ($x - \bar{x}$ values) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other n–1 deviations. The other n–1 deviations can change or vary freely. We call the number n–1 the degrees of freedom (df).

> *For each sample size n, there is a different Student's t-distribution.*

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's t-Distribution.) When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t-table gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**

The notation for the Student's t-distribution (using T as the random variable) is:

- $T \sim t_{df}$ where $df = n$–1.
- For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the error bound for a population mean is:

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n$–1 degrees of freedom, and
- $s =$ sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM). \tag{7.3.2}$$

To calculate the confidence interval directly:

Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 7.3.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 8.2267$$

$$s = 1.6722\ n = 15$$

$$df = 15\text{–}1 = 14 CL so \alpha = 1\text{–}CL = 1\text{–}0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$
$$= (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

Now it is just a direct application of Equation 7.3.2:

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$
$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq` .
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER` .
The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

? Exercise 7.3.1

You do a study of hypnotherapy to determine how effective it is in increasing the number of hourse of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Answer

(8.1634, 9.8032)

✓ Example 7.3.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table **7.3.1** shows how many of the targeted chemicals were found in each infant's cord blood.

Table **7.3.1**

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution A

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20–1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so

$$\alpha = 1–CL = 1–0.90 = 0.10 \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} \tag{7.3.3}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1–0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right) = 1.729 \left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$$\bar{x}–EBM = 127.45–10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER` .

The 90% confidence interval is (117.41, 137.49).

? Example 7.3.3

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table 7.3.2. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

Table 7.3.2

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A

- $\bar{x} = 6.133$,
- $s = 5.514$,
- $n = 15$, and
- $df = 15–1 = 14$.

$$CL = 0.98, \text{ so } \alpha = 1 - CL = 1 - 0.98 = 0.02$$

$$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} 2.624$$

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.624\left(\frac{5.514}{\sqrt{15}}\right) - 3.736$$

$$\bar{x}–EBM = 6.133–3.736 = 2.397$$

$$\bar{x} + EBM = 6.133 + 3.736 = 9.869$$

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` .
Press `ENTER` .
Arrow to `Data` and press `ENTER` .
Arrow down and enter the name of the list where the data is stored.
Enter `Freq` : 1
Enter `C-Level` : 0.98
Arrow down to `Calculate` and press `Enter` .
The 98% confidence interval is (2.3965, 9,8702).

Reference

1. “America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from http://www.businessweek.com/.
4. Data from http://www.forbes.com/.
5. “Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at www.fec.gov/data/index.jsp (accessed July 2,2013).
6. “Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
7. “Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

Glossary

Degrees of Freedom (df)

the number of objects in a sample that are free to vary

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.

- It approaches the standard normal distribution as n get larger.
- There is a "family" of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

This page titled 7.3: Confidence Intervals for the Mean with Unknown Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.3: A Single Population Mean using the Student t-Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

7.4: Confidence Intervals and Sample Size for Proportions

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03,0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the error bound, and the confidence level for a proportion is similar to that for the population mean, but the formulas are different. How do you know you are dealing with a proportion problem? First, the underlying distribution is a binomial distribution. (There is no mention of a mean or average.) If X is a binomial random variable, then

$$X \sim B(n, p)$$

where n is the number of trials and p is the probability of a success.

To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the binomial.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

P' follows a normal distribution for proportions:

$$\frac{X}{n} = P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

The confidence interval has the form

$$(p'-EBP, p'+EBP).$$

where

- EBP is error bound for the proportion.
- $p' = \frac{x}{n}$
- $p' =$ the estimated proportion of successes (p' is a point estimate for p, the true proportion.)
- $x =$ the number of successes
- $n =$ the size of the sample

The error bound (EBP) for a proportion is

$$EBP = (z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right)$$

where $q = 1 - p'$.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is

$$\sqrt{\frac{pq}{n}}.$$

However, in the error bound formula, we use

$$\sqrt{\frac{p'q'}{n}}$$

as the standard deviation, instead of

$$\sqrt{\frac{pq}{n}}.$$

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

Normal Distribution of Proportions

For the normal distribution of proportions, the z-score formula is as follows.

If

$$P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{7.4.1}$$

then the z-score formula is

$$z = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \tag{7.4.2}$$

Example 7.4.1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let $X =$ the number of people in the sample who have cell phones. X is binomial.

$$X \sim B(500, \frac{421}{500}).$$

To calculate the confidence interval, you must find p', q', and EBP.

- $n = 500$
- $x =$ the number of successes $= 421$

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Since $CL = 0.95$, then

$$\alpha = 1 - CL = 1 - 0.95 = 0.05 \left(\frac{\alpha}{2}\right) = 0.025.$$

Then

$$z\frac{\alpha}{2} = z_{0.025 = 1.96}$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 421.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (0.81003, 0.87397).

? Exercise 7.4.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Answer

(0.3315, 0.4525)

Example 7.4.2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $x = 300$ and
- $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \left(\frac{\alpha}{2}\right) = 0.05 \tag{7.4.3}$$

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 300.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER` .

The confidence interval is (0.564, 0.636).

Exercise 7.4.2

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.
b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Answer a

(0.7731, 0.8269); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

Answer b

Sixty-eight percent (68%) of students own an iPod and a smart phone.

$$p' = 0.68$$

$$q' = 1 - p' = 1 - 0.68 = 0.32$$

Since $CL = 0.97$, we know

$$\alpha = 1 - 0.97 = 0.03$$

and

$$\frac{\alpha}{2} = 0.015.$$

The area to the left of $z_{0.05}$ is 0.015, and the area to the right of $z_{0.05}$ is 1 – 0.015 = 0.985.

Using the TI 83, 83+, or 84+ calculator function InvNorm(0.985,0,1),

$$z_{0.05} = 2.17$$

$$EPB = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = 2.17\sqrt{\frac{0.68(0.32)}{300}} \approx 0.0269$$

$$p' - EPB = 0.68 - 0.0269 = 0.6531$$

$$p' + EPB = 0.68 + 0.0269 = 0.7069$$

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.6531 and 0.7069.

Calculator

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 300*0.68.
Arrow down to n and enter 300.
Arrow down to C-Level and enter 0.97.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

"Plus Four" Confidence Interval for p

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$. Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example 7.4.3

A random sample of 25 statistics students was asked: "Have you smoked a cigarette in the past week?" Six students reported smoking within the past week. Use the "plus-four" method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution A

Six students out of 25 reported smoking within the past week, so $x = 6$ and $n = 25$. Because we are using the "plus-four" method, we will use $x = 6 + 2 = 8$ and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$$q' = 1 - p' = 1 - 0.276 = 0.724$$

Since $CL = 0.95$, we know $\alpha = 1 - 0.95 = 0.05$ and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96$$

$$EPB = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$$

$$p' - EPB = 0.276 - 0.163 = 0.113$$

$$p' + EPB = 0.276 + 0.163 = 0.439$$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

REMINDER

Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to x and enter eight.

Arrow down to n and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.113, 0.439).

Exercise 7.4.3

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the "plus-four" method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution A

Using "plus four," we have $x = 31 + 2 = 33$ and $n = 65 + 4 = 69$.

$$p' = \frac{33}{69} \approx 0.478$$

$$q' = 1 - p' = 1 - 0.478 = 0.522$$

Since $CL = 0.96$, we know $\alpha = 1 - 0.96 = 0.04$ and $\frac{\alpha}{2} = 0.02$.

$$z_{0.02} = 2.054$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (2.054)\left(\sqrt{\frac{(0.478)(0.522)}{69}}\right) = 0.124$$

$$p' - EPB = 0.478 - 0.124 = 0.354$$

$$p' + EPB = 0.478 + 0.124 = 0.602$$

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 33.
Arrow down to n and enter 69.
Arrow down to C-Level and enter 0.96.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.355, 0.602).

Example 7.4.4

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the "plus four" method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution A

Using "plus-four," we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$$p' = \frac{15}{54} \approx 0.278$$

$$q' = 1 - p' = 1 - 0.241 = 0.722$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$.

$$z_{0.05} = 1.645$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

$$p' - EPB = 0.278 - 0.100 = 0.178$$

$$p' + EPB = 0.278 + 0.100 = 0.378$$

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.178, 0.378).

Exercise 7.4.4

The Berkman Center Study referenced in Example talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the "plus-four" method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in Example.

Answer

Solution A

Using "plus-four," we have $x = 159 + 2 = 161$ and $n = 588 + 4 = 592$.

$$p' = \frac{161}{592} \approx 0.272$$

$$q' = 1 - p' = 1 - 0.272 = 0.728$$

Since CL = 0.90, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030$$

$$p' - EPB = 0.272 - 0.030 = 0.242$$

$$p' + EPB = 0.272 + 0.030 = 0.302$$

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint. Press ENTER.
- Arrow down to x and enter 161.
- Arrow down to n and enter 592.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.
- The confidence interval is (0.242, 0.302).

Conclusion: The confidence interval for the larger sample is narrower than the interval from Example. Larger samples will always yield more precise confidence intervals than smaller samples. The "plus four" method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$$

Solving for n gives you an equation for the sample size.

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$$

Example 7.4.5

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

From the problem, we know that $\mathbf{EBP = 0.03}$ (3%=0.03) and $z_{\frac{\alpha}{2}} z_{0.05} = 1.645$ because the confidence level is 90%.

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that $q' = 1 – p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p'q'}{EBP^2}$$

gives

$$n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Exercise 7.4.5

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Answer

271 customers should be surveyed. Check the Real Estate section in your local

Glossary

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x = \left(\binom{n}{x}\right))p^x q^{n-x}$.

Error Bound for a Population Proportion (EBP)

the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

This page titled 7.4: Confidence Intervals and Sample Size for Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.4: A Population Proportion** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

7.5: Confidence Intervals (Summary)

Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

$$\text{(lower bound, upper bound)} = (\text{point estimate} - EBM, \text{point estimate} + EBM)$$

The calculation of EBM depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding EBM increases as well. As the sample size increases, the EBM decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (EBM) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the EBM formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

Formula Review

$\bar{X} - N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

$$\text{(lower bound, upper bound)} = (\text{point estimate} - EBM, \text{point estimate} + EBM) \quad (7.5.1)$$

$$= \bar{x} - EBM, \bar{x} + EBM \quad (7.5.2)$$

$$= \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right) \quad (7.5.3)$$

$EBM = z\frac{\sigma}{\sqrt{n}} =$ the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

$CL =$ confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL =$ the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of "EBM where $\alpha = 1 - CL$".

$n = \frac{z^2\sigma^2}{EBM^2}$ the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

$$\begin{gathered}\text{(lower value, upper value)} \\ = (\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})\end{gathered} \quad (7.5.4)$$

To find the error bound when you know the confidence interval

$$\text{error bound} = \text{upper value} - \text{point estimate} \quad (7.5.5)$$

OR

$$\text{error bound} = \frac{\text{upper value - lower value}}{2} \tag{7.5.6}$$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

? Exercise 8.2.8

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. 244
b. 15
c. 50

? Exercise 8.2.9

In words, define the random variables X and $\bar{X}$.

? Exercise 8.2.10

Which distribution should you use for this problem?

Answer

$N\left(244, \frac{15}{\sqrt{50}}\right)$

? Exercise 8.2.11

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

? Exercise 8.2.12

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Answer

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Exercise 8.2.13

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Exercise 8.2.14

In words, define the random variables X and $\bar{X}$.

Answer

X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Exercise 8.2.15

Which distribution should you use for this problem?

Exercise 8.2.16

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : (7.9441, 8.4559)

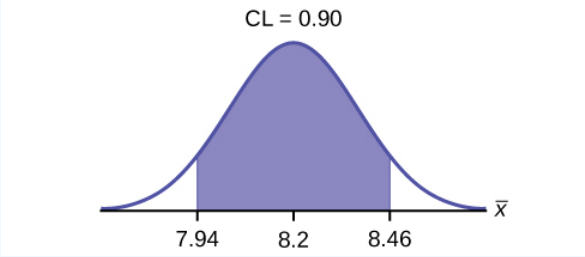


Figure 7.5.3.

$EBM = 0.26$

Exercise 8.2.17

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

Exercise 8.2.18

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Answer

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

Exercise 8.2.19

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Exercise 8.2.20

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. $\bar{x} = 2.2$
b. $\sigma = 0.2$
c. $n = 20$

Exercise 8.2.21

In words, define the random variable X.

Exercise 8.2.22

In words, define the random variable $\bar{X}$.

Answer

$\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

Exercise 8.2.23

Which distribution should you use for this problem?

Exercise 8.2.24

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$EBM = 0.07$

$CI : (2.1264, 2.2736)$

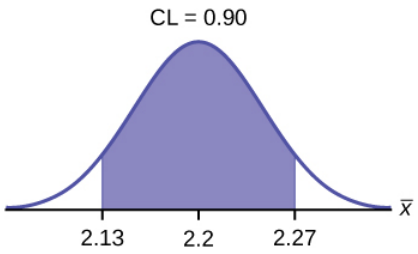


Figure 7.5.4.

? Exercise 8.2.25

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

? Exercise 8.2.26

In complete sentences, explain why the confidence interval in Exercise is larger than in Exercise.

Answer

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

? Exercise 8.2.27

In complete sentences, give an interpretation of what the interval in Exercise means.

? Exercise 8.2.28

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Answer

The confidence level would increase.

? Exercise 8.2.29

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

? Exercise 8.2.30

$\bar{x} =$ _____

Answer

30.4

Exercise 8.2.31

$n =$ _____

Exercise 8.2.32

_____ $= 15$

Answer

σ

Exercise 8.2.33

In words, define the random variable $\bar{X}$.

Exercise 8.2.34

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.2.35

Is σ_x known?

Exercise 8.2.36

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

Exercise 8.2.37

How much area is in both tails (combined)? $\alpha =$ _____

Exercise 8.2.38

How much area is in each tail? $\frac{\alpha}{2} =$ _____

Answer

0.025

Exercise 8.2.39

Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

Exercise 8.2.40

The 95% confidence interval is:_____.

Answer

(24.52,36.28)

Exercise 8.2.41

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

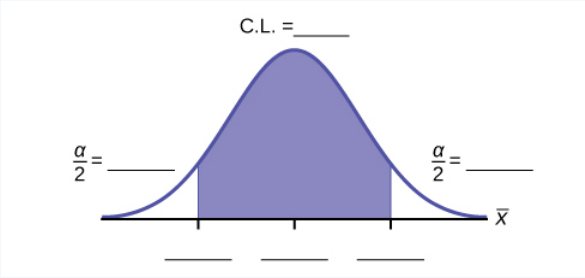


Figure 7.5.5.

Exercise 8.2.42

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

Exercise 8.2.43

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise 8.2.44

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Answer

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{7.5.7}$$

The t-score follows the Student's t-distribution with n–1 degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

$s =$ the standard deviation of sample values.

$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} =$ the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t-score in the Student's t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

$$= (\text{point estimate} - EBM, \text{point estimate} + EBM) \tag{7.5.8}$$

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}}\right) \tag{7.5.9}$$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

Exercise 8.3.3

Identify the following:

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

Exercise 8.3.4

Define the random variables X and $\bar{X}$ in words.

Answer

X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

Exercise 8.3.5

Which distribution should you use for this problem?

Exercise 8.3.6

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : (1.3808, 1.6192)

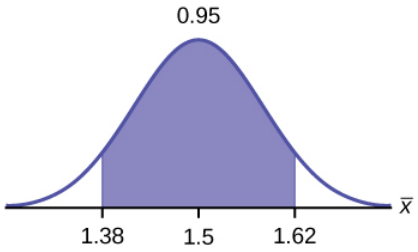


Figure 7.5.1.

$EBM = 0.12$

Exercise 8.3.7

Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Exercise 8.3.8

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

Answer

a. $\bar{x} = 151$
b. $s_x = 32$
c. $n = 108$
d. $n - 1 = 107$

Exercise 8.3.9

Define the random variable X in words.

Exercise 8.3.10

Define the random variable $\bar{X}$ in words.

Answer

$\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

Exercise 8.3.11

Which distribution should you use for this problem?

Exercise 8.3.12

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

Answer

CI : (142.92, 159.08)

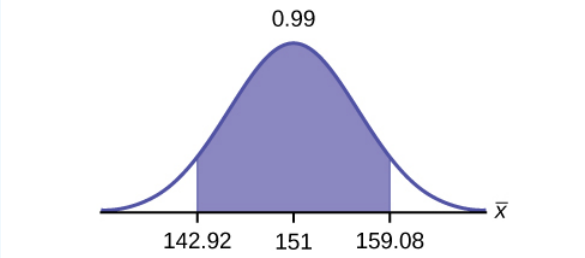


Figure 7.5.2.

$EBM = 8.08$

Exercise 8.3.13

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in Table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let $X =$ the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Exercise 8.3.14

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____

Answer

a. 3.26
b. 1.02
c. 39

Exercise 8.3.15

Define the random variable $\bar{X}$ in words.

Exercise 8.3.16

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.3.17

Is σ_x known?

Exercise 8.3.18

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

Exercise 8.3.19

How much area is in both tails (combined)?

Exercise 8.3.20

How much area is in each tail?

Answer

0.025

Exercise 8.3.21

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.3.22

The 95% confidence interval is_____.

Answer

(2.93, 3.59)

Exercise 8.3.23

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

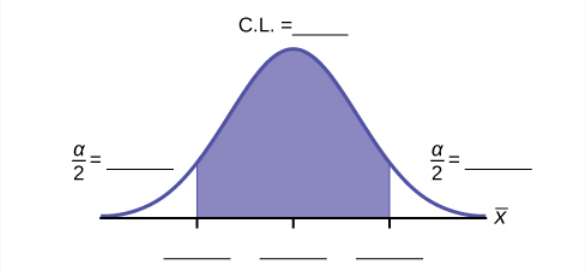


Figure 7.5.3.

Exercise 8.3.24

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Exercise 8.3.25

Using the same $\bar{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Answer

The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

Exercise 8.3.26

Using the same $\bar{x}$, s_x, and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

References

1. Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
2. Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at www.pewinternet.org/Reports/2...d-Privacy.aspx (accessed July 2, 2013).
3. Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at www.pewinternet.org/~/media//...al%20Media.pdf (accessed July 2, 2013).
4. Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/th...ement-age.aspx (accessed July 2, 2013).
5. The Field Poll. Available online at field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
6. Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/2...analytics-poll (accessed July 2, 2013).
7. “52% Say Big-Time College Athletics Corrupt Education Process.” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/publ...cation_process (accessed July 2, 2013).

Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, $\frac{x}{n}$, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}\right)$$

The “plus four” method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n_4}$, and proceed to find the confidence interval. When sample

sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = \frac{x}{n}$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p' \tag{7.5.10}$$

$$p' - N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{7.5.11}$$

The variable p' has a binomial distribution that can be approximated with the normal distribution shown here.

$EBP =$ the error bound for a proportion $= z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$

Confidence interval for a proportion:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}\right), p' + z\sqrt{\frac{p'q'}{n}}$$

$n = \frac{z_{\frac{\alpha}{2}}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error EBP.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} p' + q' = 1$

The confidence interval has the format $(p' - EBP, p' + EBP)$.

- $\bar{x}$ is a point estimate for μ
- p' is a point estimate for ρ
- s is a point estimate for σ

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

? Exercise 8.4.6

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

? Exercise 8.4.7

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Answer

It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

? Exercise 8.4.8

Identify the following:

a. $x =$ _____

b. $n =$ _____

c. $p' =$ _____

Exercise 8.4.9

Define the random variables X and P' in words.

Answer

X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

Exercise 8.4.10

Which distribution should you use for this problem?

Exercise 8.4.11

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.5321, 0.6679)$

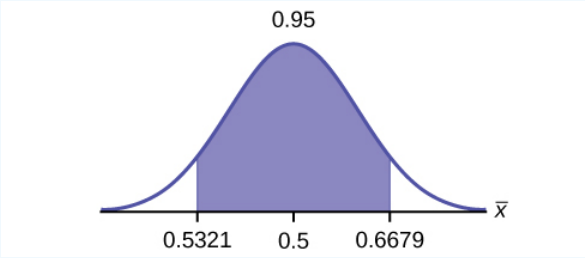


Figure 7.5.1.

EBM : 0.0679

Exercise 8.4.12

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

Exercise 8.4.13

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

Answer

X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

Exercise 8.4.14

Which distribution should you use for this problem?

Exercise 8.4.15

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.19432, 0.33068)$

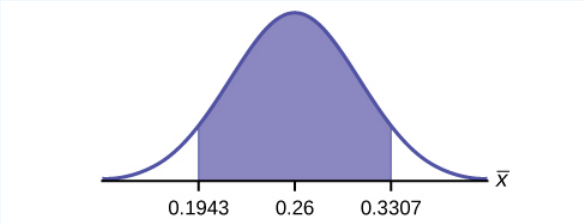


Figure 7.5.2.

EBM : 0.0707

Exercise 8.4.16

Suppose we want to lower the sampling error. What is one way to accomplish that?

Exercise 8.4.17

The sampling error given in the survey is ± 2. Explain what the ± 2 means.

Answer

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

Exercise 8.4.18

Define the random variable X in words.

Exercise 8.4.19

Define the random variable P' in words.

Answer

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Exercise 8.4.20

Which distribution should you use for this problem?

Exercise 8.4.21

Construct a 90% confidence interval, and state the confidence interval and the error bound.

Answer

CI : $(0.62735, 0.67265)$

EBM : 0.02265

Exercise 8.4.22

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

Exercise 8.4.23

What is being counted?

Answer

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

Exercise 8.4.24

In words, define the random variable X.

Exercise 8.4.25

Calculate the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

Answer

a. $x = 64$
b. $n = 80$
c. $p' = 0.8$

Exercise 8.4.26

State the estimated distribution of X. $X \sim$_____

Exercise 8.4.27

Define a new random variable P'. What is p' estimating?

Answer

p

Exercise 8.4.28

In words, define the random variable P'.

Exercise 8.4.29

State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Answer

$P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. $(0.72171, 0.87829)$.

Exercise 8.4.30

How much area is in both tails (combined)?

Exercise 8.4.31

How much area is in each tail?

Answer

0.04

Exercise 8.4.32

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.4.33

The 92% confidence interval is _____.

Answer

(0.72; 0.88)

Exercise 8.4.34

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

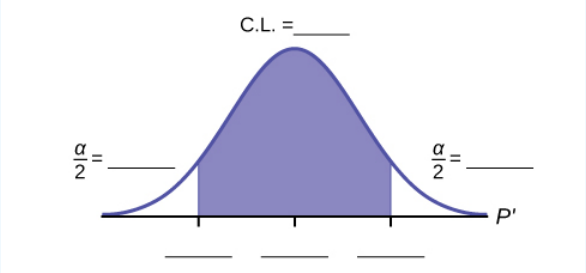


Figure 7.5.3.

Exercise 8.4.35

In one complete sentence, explain what the interval means.

Answer

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Exercise 8.4.36

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise 8.4.37

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Answer

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

Exercise 8.4.38

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

This page titled 7.5: Confidence Intervals (Summary) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.S: Confidence Intervals (Summary)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

7.E: Confidence Intervals (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

8.1: Introduction

8.2: A Single Population Mean using the Normal Distribution

Q 8.2.1

Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean height of male Swedes.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound obtained if 1,000 male Swedes are surveyed instead of 48? Why?

S 8.2.1

a. i. 71
   ii. 3
   iii. 48

b. X is the height of a Swedish male, and is the mean height from a sample of 48 Swedish males.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.849)

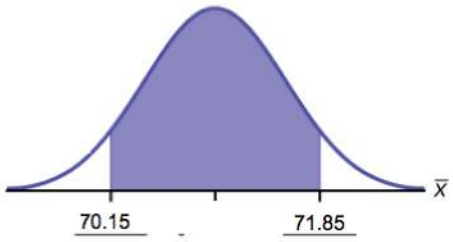


Figure 8.2.1.

ii. EBM = 0.849

e. The error bound will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

Q 8.2.2

Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

1. In words, define the random variables X and $\bar{X}$.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population mean length of engineering conferences.
   1. State the confidence interval.

2. Sketch the graph.
3. Calculate the error bound.

Q 8.2.3

Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. $\bar{x} =$ _____
   ii. $\sigma =$ _____
   iii. $n =$ _____
b. In words, define the random variables X and $\bar{X}$.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

S 8.2.3

a. i. $\bar{x} = 23.6$
   ii. $\sigma = 7$
   iii. $n = 100$
b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.
c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.
d. i. (22.228, 24.972)
   ii.

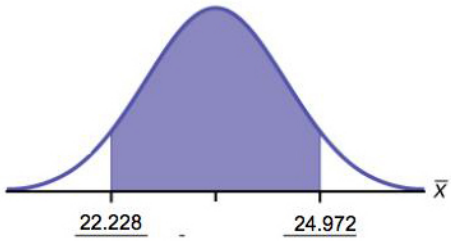


Figure 8.2.7.

   iii. $EBM = 1.372$
e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

Q 8.2.4

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. $\bar{x} =$_____
   ii. $\sigma =$_____
   iii. $s_x =$_____
b. In words, define the random variable X.
c. In words, define the random variable $\bar{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Construct a 98% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
h. In complete sentences, give an interpretation of what the interval in part f means.

Q 8.2.5

A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. i. $\bar{x} =$_____
   ii. $\sigma =$_____
   iii. $x =$_____
b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

S 8.2.5

a. i. 7.9
   ii. 2.5
   iii. 20
b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.
c. $N7.9\left(\frac{2.5}{\sqrt{20}}\right)$
d.
   i. CI: (6.98, 8.82)

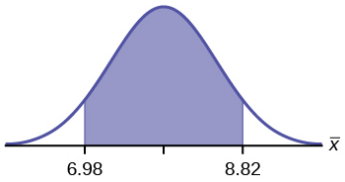


Figure 8..2.2: Copy and Paste Caption here. (Copyright; author via source)

i. EBM : 0.92

e. The error bound and confidence interval will decrease.

Q 8.2.6

What is meant by the term “90% confident” when constructing a confidence interval for a mean?

1. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
2. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
3. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
4. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Q 8.2.7

The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. Table shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

a. Find the point estimate for the population mean.

b. Using 95% confidence, calculate the error bound.

c. Create a 95% confidence interval for the mean total individual contributions.

d. Interpret the confidence interval in the context of the problem.

S 8.2.7

a. $\bar{x} = \$568,873$

b. $CL = 0.95 \alpha = 1 - 0.95 = 0.05 z_{\frac{\alpha}{2}} = 1.96$

$EBM = z_{0.025} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{909200}{\sqrt{40}} = \$281,764$

c. $\bar{x}$ − EBM = 568,873 − 281,764 = 287,109

$\bar{x}$ + EBM = 568,873 + 281,764 = 850,637

Alternate solution:

https://stats.libretexts.org/@go/page/10967

1. Press STAT and arrow over to TESTS .
2. Arrow down to 7:ZInterval .
3. Press ENTER .
4. Arrow to Stats and press ENTER .
5. Arrow down and enter the following values:
   - $\sigma : 909,200$
   - $\bar{x} : 568,873$
   - $n : 40$
   - $CL : 0.95$
6. Arrow down to Calculate and press ENTER .
7. The confidence interval is ($287,114, $850,632).
8. Notice the small difference between the two solutions–these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

Q 8.2.8

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

Q 8.2.9

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

S 8.2.9

Use the formula for EBM, solved for n:

$n = \frac{z^2\sigma^2}{EBM^2}$

From the statement of the problem, you know that σ = 2.5, and you need $EBM = 1$.

$z = z_{0.035} = 1.812$

(This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$

You need to measure at least 21 male students to achieve your goal.

8.3: A Single Population Mean using the Student t Distribution

Q 8.3.1

In six packages of “The Flintstones® Real Fruit Snacks” there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice
c. Calculate p'.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

Q 8.3.2

A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n–1 =$ _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

S 8.3.2

a. i. 8629
   ii. 6944
   iii. 35
   iv. 34

b. t_{34}

c. i. $CI : (6244, 11,014)$

   ii.

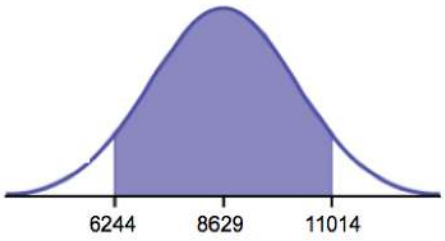


Figure 8.3.4.

   iii. $EB = 2385$

d. It will become smaller

Q 8.3.3

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n–1 =$ _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean time wasted.

i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

Q 8.3.4

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. i. $\bar{x} =$ _____
ii. $s_x =$ _____
iii. $n =$ _____
iv. $n–1 =$ _____

b. Define the random variable X in words.
c. Define the random variable $\bar{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.
i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.

f. What does it mean to be “95% confident” in this problem?

S 8.3.4

a. i. $\bar{x} = 2.51$
ii. $s_x = 0.318$
iii. $n = 9$
iv. $n - 1 = 8$

b. the effective length of time for a tranquilizer
c. the mean effective length of time of tranquilizers from a sample of nine patients
d. We need to use a Student’s-t distribution, because we do not know the population standard deviation.

e. i. $CI : (2.27, 2.76)$
ii. Check student's solution.
iii. $EBM : 0.25$

f. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

Q 8.3.5

Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. $\bar{x} =$ _____
ii. $s_x =$ _____
iii. $n =$ _____
iv. $n–1 =$ _____

b. Define the random variable X in words.
c. Define the random variable $\bar{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99% confidence interval for the population mean length of time using training wheels.
i. State the confidence interval.

ii. Sketch the graph.
iii. Calculate the error bound.

f. Why would the error bound change if the confidence level were lowered to 90%?

Q 8.3.6

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

$\bar{x} = \$251,854.23$

$s = \$521,130.41$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

S 8.3.6

$\bar{x} = \$251,854.23$

$s = \$521,130.41$

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so $n = 30$, and $df = 30 - 1 = 29$

$CL = 0.96$, so $\alpha = 1 - CL = 1 - 0.96 = 0.04$

$\frac{\alpha}{2} = 0.02 t_{0.02} = t_{0.02} = 2.150$

$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = 2.150 \left(\frac{521,130.41}{\sqrt{30}} \right) - \$204,561.66$

$\bar{x} - EBM = \$251,854.23 - \$204,561.66 = \$47,292.57$

$\bar{x} + EBM = \$251,854.23 + \$204,561.66 = \$456,415.89$

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between $47,292.57 and $456,415.89.

Alternate Solution

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` .

Press `ENTER` .

Arrow to Data and press ENTER .

Arrow down and enter the name of the list where the data is stored.

Enter Freq : 1

Enter C-Level : 0.96

Arrow down to Calculate and press Enter .

The 96% confidence interval is ($47,262, $456,447).

The difference between solutions arises from rounding differences.

Q 8.3.7

Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The Table shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's t-distribution.

Q 8.3.8

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n - 1 =$ _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

S 8.3.8

a. i. $\bar{x} =$
   ii. $s_x =$
   iii. $n =$
   iv. $n - 1 =$

b. X is the number of unoccupied seats on a single flight. $\bar{X}$ is the mean number of unoccupied seats from a sample of 225 flights.

c. We will use a Student's t-distribution, because we do not know the population standard deviation.

d. i. CI : (11.12, 12.08)
   ii. Check student's solution.
   iii. EBM: 0.48

Q 8.3.9

In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\bar{X}$ in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Explain what a "95% confidence interval" means for this study.

Q 8.3.10

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

S 8.3.10

a. i. CI: (7.64 , 9.36)

ii.

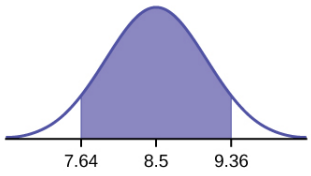


Figure 8.3.5.

iii. EBM : 0.86

b. The sample should have been increased.
c. Answers will vary.
d. Answers will vary.
e. Answers will vary.

Q 8.3.11

A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a. i. $\bar{x} =$ _____
   ii. $s_x =$ _____
   iii. $n =$ _____
   iv. $n - 1 =$ _____
b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Q 8.3.12

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

S 8.3.12

b

Q 8.3.13

What is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

8.4: A Population Proportion

Q 8.4.1

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

S 8.4.1

a. 1,068
b. The sample size would need to be increased since the critical value increases as the confidence level increases.

Q 8.4.2

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. i. $x =$ _____
ii. $n =$ _____
iii. $p' =$ _____

b. Define the random variables X and P', in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

Q 8.4.3

According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

S 8.4.3

a. $X =$ the number of people who feel that the president is doing an acceptable job;

$P' =$ the proportion of people in a sample who feel that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
c. i. $CI : (0.59, 0.63)$
   ii. Check student's solution
   iii. $EBM : 0.02$

Q 8.4.4

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

a. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and P', in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

Q 8.4.5

Refer to the information in Exercise.

a. Construct three 95% confidence intervals.

i. percent of all Asians who would welcome a white person into their families.
ii. percent of all Asians who would welcome a Latino into their families.
iii. percent of all Asians who would welcome a black person into their families.

b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

S 8.4.5

a. i. (0.72, 0.82)
ii. (0.65, 0.76)
iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

Q 8.4.6

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.
d. Explain what a “97% confidence interval” means for this study.

Q 8.4.7

A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
i. State the confidence interval.
ii. Sketch the graph.
iii. Calculate the error bound.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3. In one to three complete sentences, explain what the ±3% represents.

S 8.4.7

a. $X =$ the number of adult Americans who feel that crime is the main problem; $P' =$ the proportion of adult Americans who feel that crime is the main problem

b. Since we are estimating a proportion, given $P' = 0.2$ and $n = 1000$, the distribution we should use is $N\left(0.61, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. $CI : (0.18, 0.22)$
   ii. Check student's solution.
   iii. $EBM : 0.02$

d. One way to lower the sampling error is to increase the sample size.

e. The stated "± 3" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

Q 8.4.8

Refer to Exercise. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ± 3. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

Q 8.4.9

A point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

S 8.4.9

c

Q 8.4.10

A 90% confidence interval for the population proportion is _____.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

Q 8.4.11

The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

S 8.4.11

d

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

Q 8.4.12

Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

Q 8.4.13

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

a. 0.6614
b. 0.3386
c. 173
d. 338

S 8.4.13

a

Q 8.4.14

On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ± 3 margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify CL and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

Q 8.4.15

A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from part a and $n = 1,000$ to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

S 8.4.15

a. $p' = \frac{(0.55+0.49)}{2} = 0.52; EBP = 0.55 - 0.52 = 0.03$
b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
c. $CL = 0.75$, so $\alpha = 1–0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125 z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1–0.125 = 0.875$.)
$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1,000}} \approx 0.018$
$(p'-EBP, p'+EBP) = (0.52–0.018, 0.52 + 0.018) = (0.502, 0.538)$

Alternate Solution

STAT TESTS A: 1-PropZinterval with $x = (0.52)(1,000), n = 1,000, CL = 0.75$.

Answer is (0.502, 0.538)

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

Q 8.4.16

Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

Q 8.4.17

You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

S 8.4.17

$CL = 0.95 \alpha = 1 - 0.95 = 0.05 \frac{\alpha}{2} = 0.025 z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$n = \frac{z_{\frac{\alpha}{2}}^2 p' q'}{EPB^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

Q 8.4.18

In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as "likely" or "very likely." Use the "plus four" method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

8.5: Confidence Interval (Home Costs)

8.6: Confidence Interval (Place of Birth)

8.7: Confidence Interval (Women's Heights)

This page titled 7.E: Confidence Intervals (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **8.E: Confidence Intervals (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

1.E: Confidence Intervals for the Mean with Known Standard Deviation (Optional Exercises)

1.E: Confidence Intervals for the Mean with Known Standard Deviation (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

8: Hypothesis Testing with One Sample

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

8.1: Steps in Hypothesis Testing

8.1.1: Null and Alternative Hypotheses

8.1.2: Outcomes and the Type I and Type II Errors

8.1.3: Distribution Needed for Hypothesis Testing

8.1.4: Rare Events, the Sample, Decision and Conclusion

8.1.5: Additional Information on Hypothesis Tests

8.2: Hypothesis Test Examples for Means

8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation

8.4: Hypothesis Test Examples for Proportions

8.E: Hypothesis Testing (Optional Exercises)

1.E: Distribution Needed for Hypothesis Testing (Optional Exercises)

1.E: Hypothesis Testing with One Sample (Optional Exercises)

1.E: Null and Alternative Hypotheses (Optional Exercises)

1.E: Outcomes and the Type I and Type II Errors (Optional Exercises)

1.E: Rare Events, the Sample, Decision and Conclusion (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8: Hypothesis Testing with One Sample is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1: Steps in Hypothesis Testing

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.



Figure 8.1.1: You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analysis of the data, to reject the null hypothesis. In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.
- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
- Make a decision and write a meaningful conclusion.

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

This page titled 8.1: Steps in Hypothesis Testing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1.1: Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement of no difference between the variables—they are not related. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0. This is usually what the researcher is trying to prove.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Table **8.1.1.1**: Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal ($\neq$) **or** greater than (>) **or** less than (<)
greater than or equal to ($\geq$)	less than (<)
less than or equal to ($\geq$)	more than (>)

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

✔ Example 8.1.1.1

- H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
- H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

? Exercise 8.1.1.1

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Answer

- H_0: The drug reduces cholesterol by 25%. $p = 0.25$
- H_a: The drug does not reduce cholesterol by 25%. $p \neq 0.25$

✔ Example 8.1.1.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

- $H_0 : \mu = 2.0$
- $H_a : \mu \neq 2.0$

Exercise 8.1.1.2

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

- $H_0 : \mu\_66$
- $H_a : \mu\_66$

Answer

- $H_0 : \mu = 66$
- $H_a : \mu \neq 66$

Example 8.1.1.3

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- $H_0 : \mu \geq 5$
- $H_a : \mu < 5$

Exercise 8.1.1.3

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol ($=, \neq, \geq, <, \leq, >$) for the null and alternative hypotheses.

a. $H_0 : \mu\_45$
b. $H_a : \mu\_45$

Answer

a. $H_0 : \mu \geq 45$
b. $H_a : \mu < 45$

Example 8.1.1.4

In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

- $H_0 : p \leq 0.066$
- $H_a : p > 0.066$

Exercise 8.1.1.4

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

a. $H_0 : p\_0.40$
b. $H_a : p\_0.40$

Answer

a. $H_0 : p = 0.40$
b. $H_a : p > 0.40$

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality $(=, \leq \text{ or } \geq)$
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., $(\neq, >, \text{or } <)$.
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_a has:	equal $(=)$	greater than or equal to $(\geq)$	less than or equal to $(\leq)$
then H_a has:	not equal $(\neq)$ **or** greater than ($>$) **or** less than $(<)$	less than $(<)$	greater than $(>)$

- If $\alpha \leq p$-value, then do not reject H_0.
- If$\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

This page titled 8.1.1: Null and Alternative Hypotheses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1.2: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject H_0** when **H_0 is true (correct decision).**
2. The decision is to **reject H_0** when **H_0 is true** (incorrect decision known as aType I error).
3. The decision is **not to reject H_0** when, in fact, **H_0 is false** (incorrect decision known as a Type II error).
4. The decision is to **reject H_0** when **H_0 is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- α = probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- β = probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The *Power of the Test* is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test. The following are examples of Type I and Type II errors.

Example 8.1.2.1: Type I vs. Type II errors

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

- **Type I error**: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

β = **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Exercise 8.1.2.1

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Answer

- **Type I error**: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.
- **Type II error**: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Example 8.1.2.2

Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

- **Type I error**: The emergency crew thinks that the victim is dead when, in fact, the victim is alive.

- **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(**Type I error**).

β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(**Type II error**).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

? Exercise 8.1.2.2

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Answer

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

✓ Example 8.1.2.3

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

- **Type I error**: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
- **Type II error**: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

? Exercise 8.1.2.3

"Red tide" is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Answer

In this scenario, an appropriate null hypothesis would be H_0: the mean level of toxins is at most $800\mu g$, $H_0 : \mu_0 \leq 800\mu g$.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most $800\mu g$. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than $800\mu g$). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example 8.1.2.4

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

- **Type I**: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.
- **Type II**: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Exercise 8.1.2.4

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Answer

Type I error: c

Type II error: b

Summary

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected. The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1-\beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

- α = probability of a Type I error $= P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.
- β = probability of a Type II error $= P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

Glossary

Type 1 Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

This page titled 8.1.2: Outcomes and the Type I and Type II Errors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1.3: Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for *means*:

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{8.1.3.1}$$

or

$$t_{df} \tag{8.1.3.2}$$

The population parameter is μ. The estimated value (point estimate) for μ is $\bar{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p-q}{n}}\right) \tag{8.1.3.3}$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a Student's t-distribution (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a binomial distribution which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1–p$.

Summary

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of successes and the mean number of

failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and
  $sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

This page titled 8.1.3: Distribution Needed for Hypothesis Testing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **9.4: Distribution Needed for Hypothesis Testing** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

8.1.4: Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very *unlikely* to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the p-value. The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 8.1.4.1

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

- The null hypothesis could be $H_0 : \mu \leq 15$
- The alternate hypothesis is $H_a : \mu > 15$

The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ **is known** ($\sigma = 0.5cm.$), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16.$$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

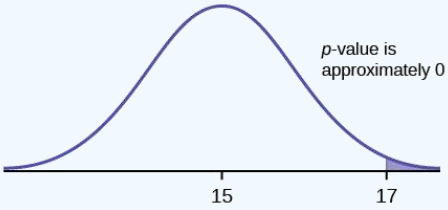


Figure 8.1.4.1

p-value $= P(\bar{x} > 17)$ which is approximately zero.

A p-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Exercise 8.1.4.1

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

- $H_0 : \mu \leq 12$
- $H_a : \mu > 12$

The p-value is 0.0013

Draw a graph that shows the p-value.

Answer

p-value $= 0.0013$

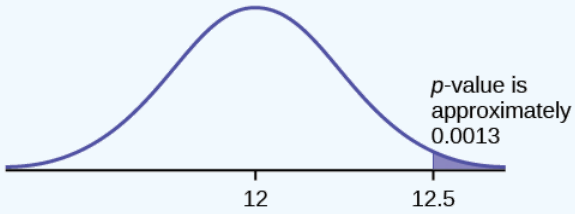


Figure 8.1.4.2

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p-value and a preset or preconceived α (also called a "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a, may be correct.
- If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.

When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses in terms of the given problem.

Example 8.1.4.2

When using the p-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

- If the p-value is low, the null must go.
- If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.

Answer

Reject the null hypothesis when **the p-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the p-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Exercise 8.1.4.2

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50, H_a : p > 0.50$
- $\alpha = 0.01$
- p-value $= 0.025$

Interpret the results and state a conclusion in simple, non-technical terms.

Answer

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

- $\alpha > p - value$, reject the null hypothesis
- $\alpha \leq p - value$, do not reject the null hypothesis

Glossary

Level of Significance of the Test

probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

p-value

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

This page titled 8.1.4: Rare Events, the Sample, Decision and Conclusion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1.5: Additional Information on Hypothesis Tests

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a, tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 8.1.5.1

$H_0 : \mu = 5, H_a : \mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:

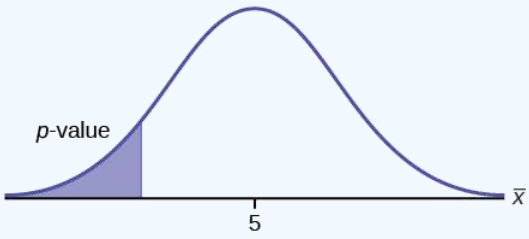


Figure 8.1.5.1

Exercise 8.1.5.1

$H_0 : \mu = 10, H_a : \mu < 10$

Assume the p-value is 0.0935. What type of test is this? Draw the picture of the p-value.

Answer

left-tailed test

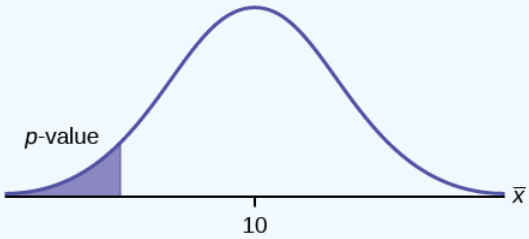


Figure 8.1.5.2

Example 8.1.5.2

$H_0 : \mu \leq 0.2, H_a : \mu < 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:

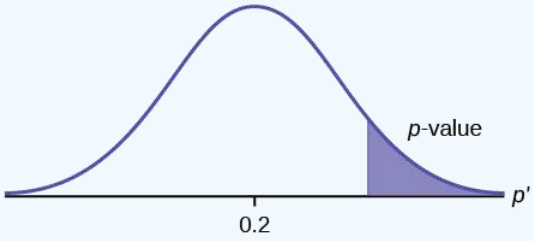


Figure 8.1.5.3

Exercise 8.1.5.2

$H_0 : \mu \leq 1, H_a : \mu > 1$

Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

Answer

right-tailed test

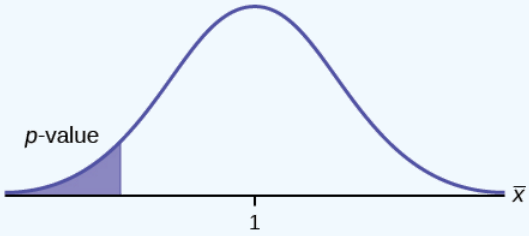


Figure 8.1.5.4

Example 8.1.5.3

$H_0 : \mu = 50, H_a : \mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the *p*-value is as follows.

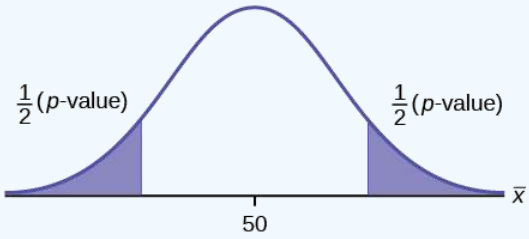


Figure 8.1.5.5

Exercise 8.1.5.3

$H_0 : \mu = 0.5, H_a : \mu \neq 0.5$

Assume the *p*-value is 0.2564. What type of test is this? Draw the picture of the *p*-value.

Answer

two-tailed test

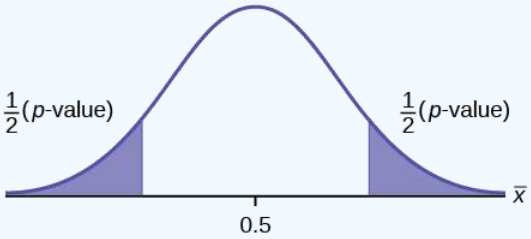


Figure **8.1.5.6**

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the **p-value**. (A *z*-score and a *t*-score are examples of test statistics.)
5. Compare the preconceived α with the *p*-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the **p-value**. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

Exercise 9.6.8

Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.6.9

Assume $H_0 : \mu \leq 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.6.10

Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

Exercise 9.6.11

Draw the general graph of a left-tailed test.

Exercise 9.6.12

Draw the graph of a two-tailed test.

Answer

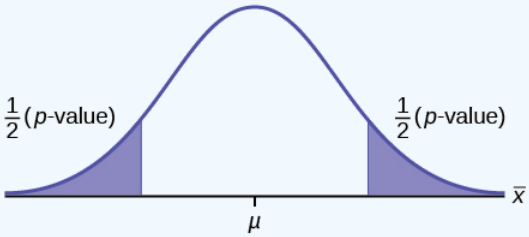


Figure 8.1.5.16

Exercise 9.6.13

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise 9.6.14

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Answer

a right-tailed test

Exercise 9.6.15

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

Exercise 9.6.16

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Answer

a left-tailed test

Exercise 9.6.17

If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

Exercise 9.6.18

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.6.19

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.6.20

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. “Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N\left(n\mu, \sqrt{n}\sigma\right)$. If the size *n* of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.1.5: Additional Information on Hypothesis Tests is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.2: Hypothesis Test Examples for Means

Full Hypothesis Test Examples

Example 8.2.4

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.**Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$H_0 : \mu = 16.43, H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

$\bar{X} - N\left(\mu, \frac{\sigma_x}{\sqrt{n}}\right)$ Therefore, $\bar{X} - N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the $p - \text{value}$ using the normal distribution for a mean:

$p\text{-value} = P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

$p\text{-value} = 0.0187$ (This is called the **actual level of significance**.) The $p - \text{value}$ is the area to the left of the sample mean is given as 16.

Graph:

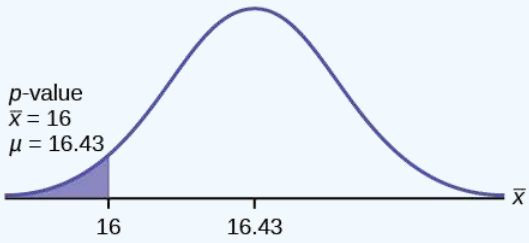


Figure 8.2.7

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the $p - \text{value}$: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the $p - \text{value}$:

$\alpha = 0.05 p\text{-value} = 0.0187 \alpha > p\text{-value}$

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

Press `STAT` and arrow over to `TESTS` . Press `1:Z-Test` . Arrow over to `Stats` and press `ENTER` . Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator not only calculates the *p*-value ($p = 0.0187$) but it also calculates the test statistic (*z*-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ ($p - \text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the *p*-value by doing a normal probability calculation using the central limit theorem:

$P(\bar{X} < 16)$ `2nd DISTR normcdf` $((-10^{99}, 16, 16.43, \frac{0.8}{\sqrt{15}})$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 8.2.4

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Press STAT and arrow over to TESTS. Press 1: Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as $<$, $\neq$, or $>$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (z-score) for the sample mean. Select $<$, $\neq$, or $>$ for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and p-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Answer

Since the problem is about a mean, this is a test of a single population mean.

- $H_0 : \mu = 40$
- $H_a : \mu > 40$
- $p = 0.0062$

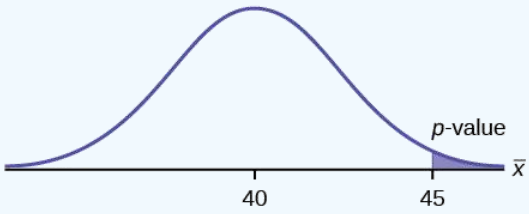


Figure 8.2.8

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Historical Note

The traditional way to compare the two probabilities, α and the $p-\text{value}$, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x}-\mu_{\bar{x}}}{\left(\frac{\sigma_{\bar{x}}}{\sqrt{n}}\right)}$. For this problem, $\bar{x} = 16$, $\mu_{\bar{x}} = 16.43$ from the null hypotheses is, $\sigma_{\bar{x}} = 0.8$, and $n = 15$.) You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p-\text{value}$), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the $p-\text{value}$, 0.0187 is considerably smaller than \(\alpha = 0.05\). You can be confident about your decision to reject. The graph shows α, the $p-\text{value}$, and the test statistics and the critical value.

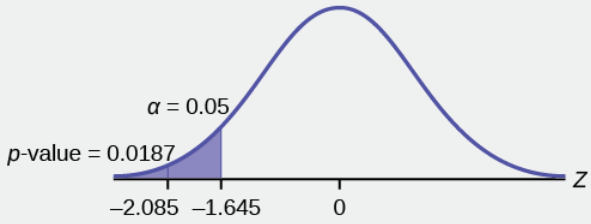


Figure 8.2.9

Example 8.2.5

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a test of a single population mean.

- $H_0 : \mu = 275$
- $H_a : \mu > 275$

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X} =$ the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

- $\bar{X} - N\left(275, \frac{55}{\sqrt{30}}\right)$
- $\bar{x} = 286.2$ pounds (from the data).
- $\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the *p*-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(\bar{x} > 286.2) = 0.1323.$$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

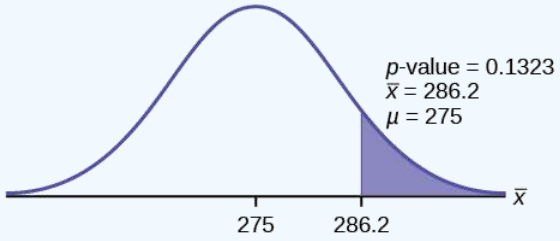


Figure **8.2.10**

Compare α and the $p - \text{value}$:

$\alpha = 0.025 p - value = 0.1323$

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The $p - \text{value}$ can easily be calculated.

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the $p - \text{value}$) (\($p = 0.1331$), a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (*z*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ ($p - \text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. "Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005." Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. "Foothill-De Anza Community College District." De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. "Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark." Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. "How often does sexual assault occur?" RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.2: Hypothesis Test Examples for Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation

Full Hypothesis Test Examples

Example 8.3.6

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Answer

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$H_0 : \mu = 65 \qquad H_a : \mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's t-distribution:

$p\text{-value} = P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

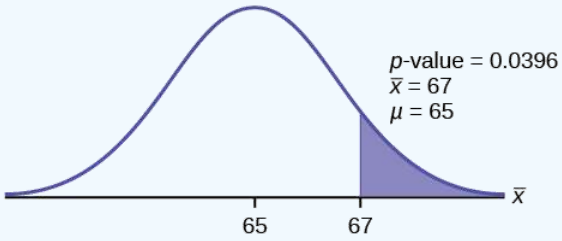


Figure 8.3.11

Compare α and the \(p-\text{value}\):

Since \(α = 0.05\) and $p\text{-value} = 0.0396$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The $p\text{-value}$ can easily be calculated.

Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only

calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER . A shaded graph appears with \(t = 1.9781\) (test statistic) and $p = 0.0396$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Exercise 8.3.6

It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : \mu = 5$
- $H_a : \mu < 5$
- $p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than $5 a week.

- Type I Error: To conclude that the stock price is growing slower than $5 a week when, in fact, the stock price is growing at $5 a week (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that the stock price is growing at a rate of $5 a week when, in fact, the stock price is growing slower than $5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example 8.3.10

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Answer

Let's follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. $H_0 : \mu \leq 1$
   b. $H_a : \mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

Figure 9.6.7.

Figure 9.6.8.

Figure 9.6.9.

Figure 9.6.10.

4. **State the Conclusions**: Since the p-value($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. “Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N\left(n\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population.

The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.4: Hypothesis Test Examples for Proportions

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a, tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

Full Hypothesis Test Examples

Example 8.4.7

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Answer

Set up the hypothesis test:

The 1% level of significance means that α = 0.01. This is a **test of a single population proportion**.

$H_0 : p = 0.50 \qquad H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $P' =$ the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$$P' - N\left(p, \sqrt{\frac{p-q}{n}}\right)$$

Therefore,

$$P' - N\left(0.5, \sqrt{\frac{0.5-0.5}{100}}\right)$$

where $p = 0.50, q = 1 - p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = P(p' < 0.47 or p' > 0.53) = 0.5485$$

where

$$x = 53, p' = \frac{x}{n} = \frac{53}{100} = 0.53$$

.

Interpretation of the \(p\text{-value}\): If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure).

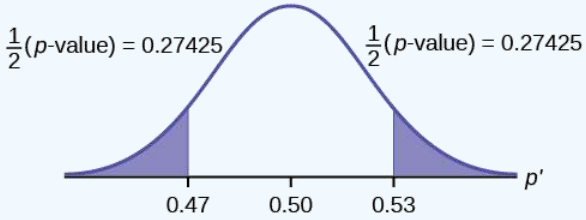


Figure 8.4.12

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 – 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value $= 0.5485$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

Press `STAT` and arrow over to `TESTS` . Press `5:1-PropZTest` . Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with \(z = 0.6\) (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 8.4.7

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Answer

Since the problem is about percentages, this is a test of single population proportions.

- $H_0 : p = 0.85$
- $H_a : p \neq 0.85$
- $p = 0.7554$

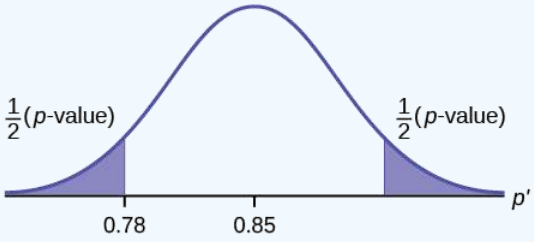


Figure **8.4.13**

Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example 8.4.8

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Answer

Set up the Hypothesis Test:

$H_0 : p = 0.30, H_a : p \neq 0.30$

Determine the distribution needed:

The **random variable** is $P' =$ proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' - N\left(0.30, \sqrt{\frac{(0.30 \cdot 0.70)}{150}}\right)$

Exercise 9.6.8.2

a. The value that helps determine the **p-value** is p'. Calculate p'.

Answer

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

$p' = 43150$

Exercise 9.6.8.3

b. What is a **success** for this problem?

Answer

b. A success is having three cell phones in a household.

Exercise 9.6.8.4

c. What is the level of significance?

Answer

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

Exercise 9.6.8.5

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the p-value.

Answer

d. $p\text{-value} = 0.7216$

Exercise 9.6.8.6

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Answer

e. Assuming that $\alpha = 0.05, \alpha < p\text{-value}$. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Exercise 8.4.8

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : p = 0.92$
- $H_a : p < 0.92$
- $p\text{-value} = 0.0046$

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

- Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example 8.4.9

My dog has so many fleas,

They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!
Before his shampoo
I counted 42.

At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out

Use the new shampoo or go without?

Answer

Set up the hypothesis test:

$H_0 : p \leq 0.25 \qquad H_a : p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\bar{X}$ or P' represents.

$P'=$ The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)1-0.25}{42}}\right)$$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = 0.0103$$

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

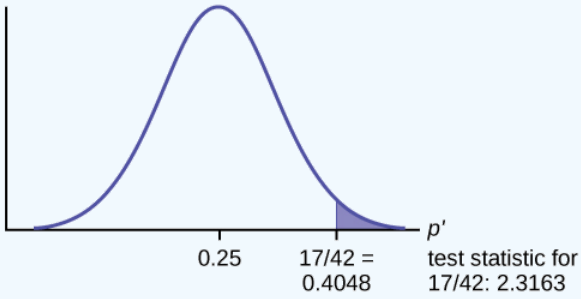


Figure 8.4.14

Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

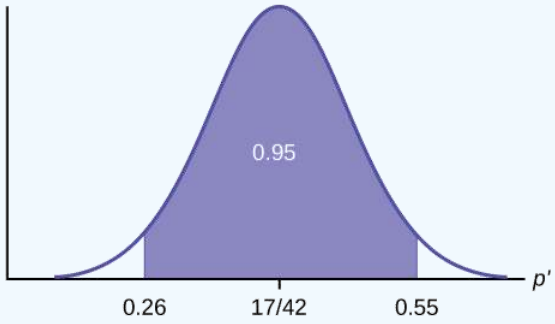


Figure 8.4.15

Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion p of fleas that are killed by the new shampoo is between 26% and 55%.

This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example 8.4.11

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. $H_0 : p \leq 0.00034$
   b. $H_a : p > 0.00034$

If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.

2. We will be testing a sample proportion with $x = 172$ and $n = 420,019$. The sample is sufficiently large because we have $np = 420,019(0.00034) = 142.8$, $nq = 420,019(0.99966) = 419,876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are

Figure 9.6.11.

Figure 9.6.12.

4. Since the p-value $= 0.0073$ is greater than our alpha value $= 0.005$, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example 8.4.12

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Answer

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion z-test. The hypotheses for the test will be
   a. $H_0 : p = 0.00078$
   b. $H_a : p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

Figure 9.6.13.

Figure 9.6.14.

4. Since the p-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at www.census.gov/hhes/socdemo/language/.
13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. “Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N\left(n\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.4: Hypothesis Test Examples for Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.E: Hypothesis Testing (Optional Exercises)

8.E: Hypothesis Testing (Optional Exercises) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

1.E: Distribution Needed for Hypothesis Testing (Optional Exercises)

Exercise 1.E.1

Which two distributions can you use for hypothesis testing for this chapter?

Answer

A normal distribution or a Student's *t*-distribution

Exercise 1.E.2

Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.

Exercise 1.E.3

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

Answer

Use a Student's ***t***-distribution

Exercise 1.E.4

A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Exercise 1.E.5

A population has a mean of 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Answer

a normal distribution for a single population mean

Exercise 1.E.6

It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

Exercise 1.E.7

You are performing a hypothesis test of a single population mean using a Student's ***t***-distribution. What must you assume about the distribution of the data?

Answer

It must be approximately normally distributed.

Exercise 1.E.8

You are performing a hypothesis test of a single population mean using a Student's ***t***-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

Exercise 1.E.9

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

Answer

They must both be greater than five.

Exercise 1.E.10

You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

Exercise 1.E.11

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

Answer

binomial distribution

1.E: Distribution Needed for Hypothesis Testing (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Hypothesis Testing with One Sample (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

9.1: Introduction

9.2: Null and Alternative Hypotheses

Q 9.2.1

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

S 9.2.1

a. $H_0 : \mu = 34; H_a : \mu \neq 34$
b. $H_0 : p \leq 0.60; H_a : p > 0.60$
c. $H_0 : \mu \geq 100,000; H_a : \mu < 100,000$
d. $H_0 : p = 0.29; H_a : p \neq 0.29$
e. $H_0 : p = 0.05; H_a : p < 0.05$
f. $H_0 : \mu \leq 10; H_a : \mu > 10$
g. $H_0 : p = 0.50; H_a : p \neq 0.50$
h. $H_0 : \mu = 6; H_a : \mu \neq 6$
i. $H_0 : p \geq 0.11; H_a : p < 0.11$
j. $H_0 : \mu \leq 20,000; H_a : \mu > 20,000$

Q 9.2.2

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

Q 9.2.3

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$

c. $p < 0.20$
d. $p \leq 0.20$

S 9.2.3

c

Q 9.2.4

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. $H_0 : \bar{x} = 4.5, H_a : \bar{x} > 4.5$
b. $H_0 : \mu \geq 4.5, H_a : \mu < 4.5$
c. $H_0 : \mu = 4.75, H_a : \mu > 4.75$
d. $H_0 : \mu = 4.5, H_a : \mu > 4.5$

9.3: Outcomes and the Type I and Type II Errors

Q 9.3.1

State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

S 9.3.1

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.

h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

Q 9.3.2

For statements a-j in Exercise 9.109, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

Q 9.3.3

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

S 9.3.3

b

Q 9.3.4

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is at least 20%.
d. less than 20%, when in fact, it is less than 20%.

Q 9.3.4

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

S 9.3.4

d

Q 9.3.5

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

9.4: Distribution Needed for Hypothesis Testing

Q 9.4.1

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ _____

a. $N\left(7.24, \frac{1.93}{\sqrt{22}}\right)$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

S 9.4.1

d

9.5: Rare Events, the Sample, Decision and Conclusion

Q 9.5.1

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0 : _____ H_a : _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   i. $x =$ _____
   ii. $n =$ _____
   iii. $p' =$ _____
g. Calculate $\sigma_x =$ _____. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:

ii. Reason for the decision:
iii. Conclusion (write out in a complete sentence):

9.6: Additional Information and Full Hypothesis Test Examples

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's t-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Q 9.6.1.

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using $\alpha = 0.05$, is the data highly inconsistent with the claim?

S 9.6.1

a. $H_0 : \mu \geq 50,000$
b. $H_a : \mu < 50,000$
c. Let $\bar{X} =$ the average lifespan of a brand of tires.
d. normal distribution
e. $z = -2.315$
f. p-value $= 0.0103$
g. Check student's solution.
h. i. alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The p-value is less than 0.05.
iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
i. $(43,537, 49,463)$

Q 9.6.2

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Q 9.6.3

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is \$1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

S 9.6.3

a. $H_0 : \mu = \$1.00$
b. $H_a : \mu \neq \$1.00$
c. Let $\bar{X} =$ the average cost of a daily newspaper.
d. normal distribution
e. $z = -0.866$
f. p-value $= 0.3865$
g. Check student's solution.
h. i. $\alpha : 0.01$
ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The p-value is greater than 0.01.
iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.

i. $(\$0.84, \$1.06)$

Q 9.6.4

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Q 9.6.5

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let $x =$ the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?

S 9.6.5

a. $H_0 : \mu = 10$
b. $H_a : \mu \neq 10$
c. Let $\bar{X}$ the mean number of sick days an employee takes per year.
d. Student's *t*-distribution
e. $t = -1.12$
f. p-value $= 0.300$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05.
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.
i. $(4.9443, 11.806)$

Q 9.6.6

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Q 9.6.7

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

S 9.6.7

a. $H_0 : p \geq 0.6$
b. $H_a : p < 0.6$
c. Let $P' =$ the proportion of students who feel more enriched as a result of taking Elementary Statistics.
d. normal for a single proportion
e. 1.12
f. p-value $= 0.1308$

g. Check student's solution.

h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05.
iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence Interval: $(0.409, 0.654)$

The "plus-4s" confidence interval is $(0.411, 0.648)$

Q 9.6.8

A Nissan Motor Corporation advertisement read, "The average man's I.Q. is 107. The average brown trout's I.Q. is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

Q 9.6.9

Refer to Exercise 9.119. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.

S 9.6.9

a. $H_0 : \mu = 4$
b. $H_a : \mu \neq 4$
c. Let $\bar{X}$ the average I.Q. of a set of brown trout.
d. two-tailed Student's t-test
e. $t = 1.95$
f. p-value $= 0.076$
g. Check student's solution.

h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05
iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. $(3.8865, 5.9468)$

Q 9.6.10

According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Q 9.6.11

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

S 9.6.11

a. $H_0 : p \geq 0.13$

b. $H_a : p < 0.13$
c. Let $P' =$ the proportion of Americans who have seen or sensed angels
d. normal for a single proportion
e. –2.688
f. p-value $= 0.0036$

g. Check student's solution.

h. i. alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The p-value is less than 0.05.
iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

i. $(0, 0.0623)$.

The"plus-4s" confidence interval is (0.0022, 0.0978)

Q 9.6.12

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Q 9.6.13

Use the "Lap time" data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

S 9.6.13

a. $H_0 : \mu \geq 129$
b. $H_a : \mu < 129$
c. Let $\bar{X} =$ the average time in seconds that Terri finishes Lap 4.
d. Student's t-distribution
e. $t = 1.209$
f. 0.8792
g. Check student's solution.

h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The p-value is greater than 0.05.
iv. Conclusion: There is insufficient evidence to conclude that Terri's mean lap time is less than 129 seconds.

i. $(128.63, 130.37)$

Q 9.6.14

Use the "Initial Public Offering" data (see [link]) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note

The following questions were written by past students. They are excellent problems!

Q 9.6.15

"Asian Family Reunion," by Chau Nguyen

Every two years it comes around.

We all get together from different towns.

In my honest opinion,

It's not a typical family reunion.

Not forty, or fifty, or sixty,

But how about seventy companions!

The kids would play, scream, and shout

One minute they're happy, another they'll pout.

The teenagers would look, stare, and compare

From how they look to what they wear.

The men would chat about their business

That they make more, but never less.

Money is always their subject

And there's always talk of more new projects.

The women get tired from all of the chats

They head to the kitchen to set out the mats.

Some would sit and some would stand

Eating and talking with plates in their hands.

Then come the games and the songs

And suddenly, everyone gets along!

With all that laughter, it's sad to say

That it always ends in the same old way.

They hug and kiss and say "good-bye"

And then they all begin to cry!

I say that 60 percent shed their tears

But my mom counted 35 people this year.

She said that boys and men will always have their pride,

So we won't ever see them cry.

I myself don't think she's correct,

So could you please try this problem to see if you object?

S 9.6.15

a. $H_0 : p = 0.60$
b. $H_a : p < 0.60$
c. Let $P' =$ the proportion of family members who shed tears at a reunion.
d. normal for a single proportion
e. –1.71
f. 0.0438
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the p-value and alpha are quite close, so other tests should be done.

i. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. $(0.3829, 0.6171)$. The"plus-4s" confidence interval (see chapter 8) is $(0.3861, 0.6139)$

Note that here the "large-sample" $1-$ PropZTest provides the approximate p-value of 0.0438. Whenever a p-value based on a normal approximation is close to the level of significance, the exact p-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

Q 9.6.16

"The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,

The catalyst came from the magazine, Time.

On the magazine cover I did find

The realm of angels tickling my mind.

Inside, 69% I found to be

In angels, Americans do believe.

Then, it was time to rise to the task,

Ninety-five high school and college students I did ask.

Viewing all as one group,

Random sampling to get the scoop.

So, I asked each to be true,

"Do you believe in angels?" Tell me, do!

Hypothesizing at the start,

Totally believing in my heart

That the proportion who said yes

Would be equal on this test.

Lo and behold, seventy-three did arrive,

Out of the sample of ninety-five.

Now your job has just begun,

Solve this problem and have some fun.

Q 9.6.17

"Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,

I had to find some sane defense.

Some light and lifting simple play

To float my math anxiety away.

Blowing bubbles lifts me high

Takes my troubles to the sky.

POIK! They're gone, with all my stress

Bubble therapy is the best.

The label said each time I blew

The average number of bubbles would be at least 22.

I blew and blew and this I found

From 64 blows, they all are round!

But the number of bubbles in 64 blows

Varied widely, this I know.

20 per blow became the mean

They deviated by 6, and not 16.

From counting bubbles, I sure did relax

But now I give to you your task.

Was 22 a reasonable guess?

Find the answer and pass this test!

S 9.6.17

a. $H_0 : \mu \geq 22$
b. $H_a : \mu < 22$
c. Let $\bar{X} =$ the mean number of bubbles per blow.
d. Student's *t*-distribution
e. –2.667
f. $p\text{-value} = 0.00486$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.
i. $(18.501, 21.499)$

Q 9.6.18

"Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles

Bred puppies with numerous freckles

The Dalmatians he sought

Possessed spot upon spot

The more spots, he thought, the more shekels.

His competitors did not agree

That freckles would increase the fee.

They said, "Spots are quite nice

But they don't affect price;

One should breed for improved pedigree."

The breeders decided to prove

This strategy was a wrong move.

Breeding only for spots

Would wreak havoc, they thought.

His theory they want to disprove.

They proposed a contest to Spreckles

Comparing dog prices to freckles.

In records they looked up

One hundred one pups:

Dalmatians that fetched the most shekels.

They asked Mr. Spreckles to name

An average spot count he'd claim

To bring in big bucks.

Said Spreckles, “Well, shucks,

It's for one hundred one that I aim.”

Said an amateur statistician

Who wanted to help with this mission.

“Twenty-one for the sample

Standard deviation's ample:

They examined one hundred and one

Dalmatians that fetched a good sum.

They counted each spot,

Mark, freckle and dot

And tallied up every one.

Instead of one hundred one spots

They averaged ninety six dots

Can they muzzle Spreckles’

Obsession with freckles

Based on all the dog data they've got?

Q 9.6.19

"Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese:

- 5 stores @ $2.02

- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

S 9.6.19

a. $H_0 : \mu \leq 1$
b. $H_a : \mu > 1$
c. Let $\bar{X} =$ the mean cost in dollars of macaroni and cheese in a certain town.
d. Student's t-distribution
e. $t = 0.340$
f. p-value $= 0.36756$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05
   iv. Conclusion: The mean cost could be $1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.
i. $(0.8291, 1.241)$

Q 9.6.20

"William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi

THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet's tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

Q 9.6.21

"Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

S 9.6.21

a. $H_0 : p = 0.01$
b. $H_a : p > 0.01$
c. Let $P' =$ the proportion of errors generated
d. Normal for a single proportion
e. 2.13
f. 0.0165
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.
i. Confidence interval: $(0, 0.094)$.

The"plus-4s" confidence interval is $(0.004, 0.144)$.

Q 9.6.22

"Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

Q 9.6.23

"Phillip's Wish," by Suzanne Osorio

My nephew likes to play

Chasing the girls makes his day.

He asked his mother

If it is okay

To get his ear pierced.

She said, "No way!"

To poke a hole through your ear,

Is not what I want for you, dear.

He argued his point quite well,

Says even my macho pal, Mel,

Has gotten this done.

It's all just for fun.

C'mon please, mom, please, what the hell.

Again Phillip complained to his mother,

Saying half his friends (including their brothers)

Are piercing their ears

And they have no fears

He wants to be like the others.

She said, "I think it's much less.

We must do a hypothesis test.

And if you are right,

I won't put up a fight.

But, if not, then my case will rest."

We proceeded to call fifty guys

To see whose prediction would fly.

Nineteen of the fifty

Said piercing was nifty

And earrings they'd occasionally buy.

Then there's the other thirty-one,

Who said they'd never have this done.

So now this poem's finished.

Will his hopes be diminished,

Or will my nephew have his fun?

S 9.6.23

a. $H_0 : p = 0.50$
b. $H_a : p < 0.50$
c. Let $P' =$ the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. p-value $= 0.0448$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05. (However, they are very close.)
   iv. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.
i. Confidence Interval: $(0.245, 0.515)$: The "plus-4s" confidence interval is $(0.259, 0.519)$.

Q 9.6.24

"The Craven," by Mark Salangsang

Once upon a morning dreary

In stats class I was weak and weary.

Pondering over last night's homework

Whose answers were now on the board

This I did and nothing more.

While I nodded nearly napping

Suddenly, there came a tapping.

As someone gently rapping,

Rapping my head as I snore.

Quoth the teacher, "Sleep no more."

"In every class you fall asleep,"

The teacher said, his voice was deep.

“So a tally I’ve begun to keep

Of every class you nap and snore.

The percentage being forty-four.”

“My dear teacher I must confess,

While sleeping is what I do best.

The percentage, I think, must be less,

A percentage less than forty-four.”

This I said and nothing more.

“We’ll see,” he said and walked away,

And fifty classes from that day

He counted till the month of May

The classes in which I napped and snored.

The number he found was twenty-four.

At a significance level of 0.05,

Please tell me am I still alive?

Or did my grade just take a dive

Plunging down beneath the floor?

Upon thee I hereby implore.

Q 9.6.25

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

S 9.6.25

a. $H_0 : p = 0.40$
b. $H_a : p < 0.40$
c. Let $P' =$ the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. p-value $= 0.1563$
g. Check student’s solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.
i. Confidence Interval: $(0.3241, 0.4240)$: The “plus-4s” confidence interval is $(0.3257, 0.4250)$.

Q 9.6.26

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California’s percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time

faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

Q 9.6.27

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

S 9.6.27

a. $H_0 : p = 0.14$
b. $H_a : p < 0.14$
c. Let $P' =$ the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. p-value $= 0.3914$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.
i. Confidence Interval: $(0.0502, 0.2070)$: The "plus-4s" confidence interval (see chapter 8) is $(0.0676, 0.2297)$.

Q 9.6.28

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Q 9.6.29

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

S 9.6.29

a. $H_0 : \mu = 69,110$
b. $H_0 : \mu > 69,110$
c. Let $\bar{X} =$ the mean salary in dollars for California registered nurses.
d. Student's t-distribution
e. $t = 1.719$
f. p-value : 0.0466
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.
i. $(\$68,757, \$73,485)$

Q 9.6.30

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Q 9.6.31

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?

After conducting the test, your decision and conclusion are

a. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

S 9.6.31

c

Q 9.6.32

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Q 9.6.33

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $\alpha = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

S 9.6.33

c

Instructions: For the following ten exercises,

Hypothesis testing: For the following ten exercises, answer each question.

State the null and alternate hypothesis.

State the p-value.

State α.

What is your decision?

Write a conclusion.

Answer any other questions asked in the problem.

Q 9.6.34

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

Q 9.6.35

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

S 9.6.35

a. $H_0 : p = 0.488$ $H_a : p \neq 0.488$
b. p-value $= 0.0114$
c. $\alpha = 0.05$
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

Q 9.6.36

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

Q 9.6.37

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

S 9.6.37

a. $H_0 : p = 0.517$ $H_0 : p \neq 0.517$
b. p-value $= 0.9203$.
c. $\alpha = 0.05$.
d. Do not reject the null hypothesis.
e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

Q 9.6.38

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

Q 9.6.39

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05 level$, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

S 9.6.39

a. $H_0 : \mu \geq 11.52$ $H_a : \mu < 11.52$
b. $p\text{-value} = 0.000002$ which is almost 0.
c. $\alpha = 0.05$.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the $p\text{-value}$ is almost 0.

Q 9.6.40

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Q 9.6.41

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

S 9.6.42

1. $H_0 : \mu \leq 5.8$ $H_a : \mu > 5.8$
2. $p\text{-value} = 0.9987$
3. $\alpha = 0.05$
4. Do not reject the null hypothesis.
5. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

Q 9.6.42

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:

5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

Q 9.6.43

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

S 9.6.43

a. $H_0 : \mu \geq 150$ $H_0 : \mu < 150$
b. $p\text{-value} = 0.0622$
c. $\alpha = 0.01$
d. Do not reject the null hypothesis.
e. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

9.7: Hypothesis Testing of a Single Mean and Single Proportion

This page titled 1.E: Hypothesis Testing with One Sample (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **9.E: Hypothesis Testing with One Sample (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

1.E: Null and Alternative Hypotheses (Optional Exercises)

Exercise 1.E.5

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

Answer

The random variable is the mean Internet speed in Megabits per second.

Exercise 1.E.1

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Exercise 1.E.1

The American family has an average of two children. What is the random variable? Describe in words.

Answer

The random variable is the mean number of children an American family has.

Exercise 1.E.8

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Exercise 1.E.9

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

Answer

The random variable is the proportion of people picked at random in Times Square visiting the city.

Exercise 1.E.10

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Exercise 1.E.11

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Answer

a. $H_0 : p = 0.42$
b. $H_a : p < 0.42$

Exercise 1.E.12

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0:_____
b. H_a:_____

Exercise 1.E.13

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0:_____
b. H_a:_____

Answer

a. $H_0 : \mu = 15$
b. $H_a : \mu \neq 15$

Exercise

The National Institute 9.2.14 of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

1. H_0:_____
2. H_a:_____

1.E: Null and Alternative Hypotheses (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Outcomes and the Type I and Type II Errors (Optional Exercises)

Exercise 1.E.5

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Answer

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

Exercise 1.E.6

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

Exercise 1.E.7

For Exercise 9.12, what are α and β in words?

Answer

$\alpha =$ the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

$\beta =$ the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

Exercise 1.E.8

In words, describe $1 - \beta$ For Exercise Runaway argument for \eatSpaces?

Exercise 1.E.9

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Answer

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Exercise 1.E.10

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

Exercise 1.E.11

The power of a test is 0.981. What is the probability of a Type II error?

Answer

0.019

Exercise 1.E.12

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Exercise 1.E.13

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

Answer

0.998

Exercise 1.E.14

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

1.E: Outcomes and the Type I and Type II Errors (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Rare Events, the Sample, Decision and Conclusion (Optional Exercises)

Exercise 1.E.3

When do you reject the null hypothesis?

Exercise 1.E.4

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

Answer

The outcome of winning is very unlikely.

Exercise 1.E.5

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

Exercise 1.E.6

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p-value is almost zero. State the null and alternative hypotheses and interpret the p-value.

Answer

$H_0 : \mu \geq 73$

$H_a : \mu \leq 73$

The p-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

Exercise 1.E.7

The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

Exercise 1.E.8

Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?

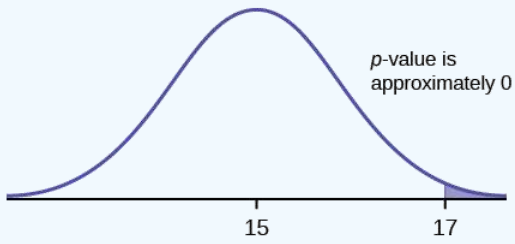


Figure 1.E.3.

Answer

The shaded region shows a low p-value.

Exercise 1.E.9

What should you do when $\alpha > p\text{-value}$?

Exercise 1.E.10

What should you do if $\alpha = p\text{-value}$?

Answer

Do not reject H_0.

Exercise 1.E.11

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Exercise 1.E.12

Is this a test of means or proportions?

Answer

means

Exercise 1.E.13

What symbol represents the random variable for this test?

Exercise 1.E.14

In words, define the random variable for this test.

Answer

the mean time spent in jail for 26 first time convicted burglars

Exercise 1.E.15

Is the population standard deviation known and, if so, what is it?

Exercise 1.E.16

Calculate the following:

a. $\bar{x}$ _____

b. σ _____

c. s_x _____

d. n _____

Answer

a. 3
b. 1.5
c. 1.8
d. 26

Exercise 1.E.17

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

Exercise 1.E.18

State the distribution to use for the hypothesis test.

Answer

$\bar{X} - N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

Exercise 1.E.19

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   $H0$: _____ H_a: _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:
   i. $\bar{x}$ = _____
   ii. s = _____
   iii. n = _____
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

1.E: Rare Events, the Sample, Decision and Conclusion (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

9: Inferences with Two Samples

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded. To compare two means or two proportions, you work with two groups. The groups are classified either as independent or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

9.1: Prelude to Hypothesis Testing with Two Samples

9.2: Inferences for Two Population Means- Large, Independent Samples

9.3: Inferences for Two Population Means - Unknown Standard Deviations

9.4: Inferences for Two Population Means - Paired Samples

9.5: Inferences for Two Population Proportions

9.6: Which Analysis Should You Conduct?

9.E: Hypothesis Testing with Two Samples (Optional Exercises)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9: Inferences with Two Samples is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Prelude to Hypothesis Testing with Two Samples

Learning Objectives

By the end of this chapter, the student should be able to:

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.



Figure **9.1.1**. If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and *p*-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Test of the two population proportions by testing one population mean of differences.

This page titled 9.1: Prelude to Hypothesis Testing with Two Samples is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.2: Inferences for Two Population Means- Large, Independent Samples

Learning Objectives

- To understand the logical framework for estimating the difference between the means of two distinct populations and performing tests of hypotheses concerning those means.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using large, independent samples.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using large, independent samples.

Suppose we wish to compare the means of two distinct populations. Figure **9.2.1** illustrates the conceptual framework of our investigation in this and the next section. Each population has a mean and a standard deviation. We arbitrarily label one population as Population **1** and the other as Population **2**, and subscript the parameters with the numbers **1** and **2** to tell them apart. We draw a random sample from Population **1** and label the sample statistics it yields with the subscript **1**. Without reference to the first sample we draw a sample from Population **2** and label its sample statistics with the subscript **2**.

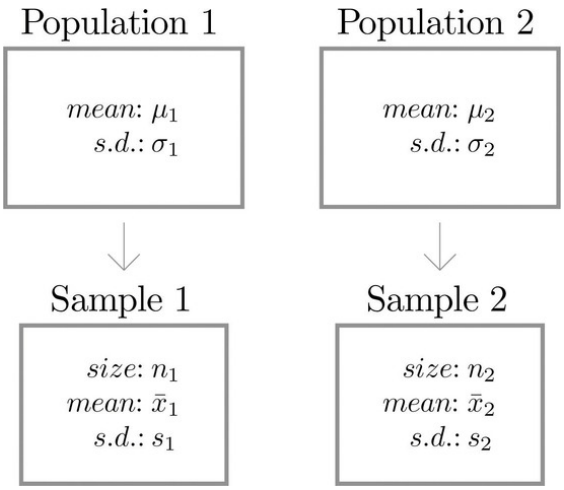


Figure **9.2.1**: Independent Sampling from Two Populations

Definition: Independence

Samples from two distinct populations are *independent* if each one is drawn without reference to the other, and has no connection with the other.

Our goal is to use the information in the samples to estimate the difference $\mu_1 - \mu_2$ in the means of the two populations and to make statistically valid inferences about it.

Confidence Intervals

Since the mean $x - 1$ of the sample drawn from Population **1** is a good estimator of μ_1 and the mean $x - 2$ of the sample drawn from Population **2** is a good estimator of μ_2, a reasonable point estimate of the difference $\mu_1 - \mu_2$ is $\bar{x_1} - \bar{x_2}$. In order to widen this point estimate into a confidence interval, we first suppose that both samples are large, that is, that both $n_1 \geq 30$ and $n_2 \geq 30$. If so, then the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid. The symbols s_1^2 and s_2^2 denote the squares of s_1 and s_2. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Large, Independent Samples

The samples must be independent, and *each* sample must be large:

Example 9.2.1

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$x-1 = 3.51$	$x-2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

Construct a point estimate and a 99% confidence interval for $\mu_1 - \mu_2$, the difference in average satisfaction levels of customers of the two companies as measured on this five-point scale.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x_1} - \bar{x_2} = 3.51 - 3.24 = 0.27$$

In words, we estimate that the average customer satisfaction level for Company 1 is 0.27 points higher on this five-point scale than it is for Company 2.

To apply the formula for the confidence interval, proceed exactly as was done in Chapter 7. The 99% confidence level means that $\alpha = 1 - 0.99 = 0.01$ so that $z_{\alpha/2} = z_{0.005}$. From Figure 7.1.6 "Critical Values of " we read directly that $z_{0.005} = 2.576$. Thus

$$(\bar{x_1} - \bar{x_2}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.27 \pm 2.576\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}} = 0.27 \pm 0.12$$

We are 99% confident that the difference in the population means lies in the interval $[0.15, 0.39]$, in the sense that in repeated sampling 99% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. In the context of the problem we say we are 99% confident that the average level of customer satisfaction for Company 1 is between 0.15 and 0.39 points higher, on this five-point scale, than that for Company 2.

Hypothesis Testing

Hypotheses concerning the relative sizes of the means of two populations are tested using the same critical value and p-value procedures that were used in the case of a single population. All that is needed is to know how to express the null and alternative hypotheses and to know the formula for the standardized test statistic and the distribution that it follows.

The null and alternative hypotheses will always be expressed in terms of the difference of the two population means. Thus the null hypothesis will always be written

$$H_0 : \mu_1 - \mu_2 = D_0$$

where D_0 is a number that is deduced from the statement of the situation. As was the case with a single population the alternative hypothesis can take one of the three forms, with the same terminology:

Form of H*a*	Terminology
$H_a : \mu_1 - \mu_2 < D_0$	Left-tailed
$H_a : \mu_1 - \mu_2 > D_0$	Right-tailed
$H_a : \mu_1 - \mu_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Large, Independent Samples

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: $n_1 \geq 30$ and $n_2 \geq 30$.

Example 9.2.2

Refer to Example 9.2.1 concerning the mean satisfaction levels of customers of two competing cable television companies. Test at the 1% level of significance whether the data provide sufficient evidence to conclude that Company 1 has a higher mean satisfaction rating than does Company 2. Use the critical value approach.

Solution:

- **Step 1**. If the mean satisfaction levels μ_1 and μ_2 are the same then $\mu_1 = \mu_2$, but we always express the null hypothesis in terms of the difference between μ_1 and μ_2, hence H_0 is $\mu_1 - \mu_2 = 0$. To say that the mean customer satisfaction for Company 1 is higher than that for Company 2 means that $\mu_1 > \mu_2$, which in terms of their difference is $\mu_1 - \mu_2 > 0$. The test is therefore

$$H_0 : \mu_1 - \mu_2 = 0$$

$$vs.$$

$$H_a : \mu_1 - \mu_2 > 0 \ @ \ \alpha = 0.01$$

- **Step 2**. Since the samples are independent and both are large the test statistic is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.51 - 3.24) - 0}{\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}} = 5.684$$

- **Step 4**. Since the symbol in H_a is “>” this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.01}$, which from the last line in Figure 7.1.6 "Critical Values of " we read off as 2.326. The rejection region is $[2.326, \infty)$.

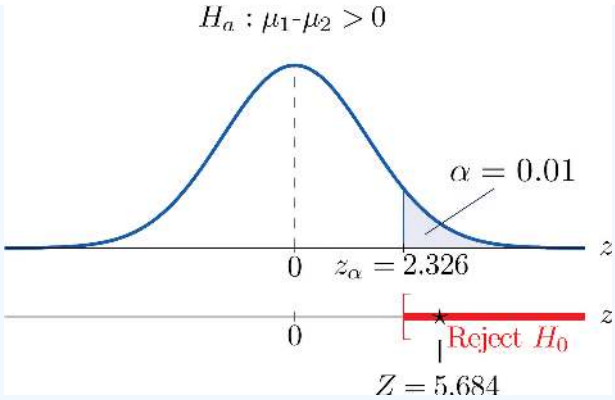


***Figure* 9.2.2**: *Rejection Region and Test Statistic for Example* **9.2.2**

- **Step 5**. As shown in Figure **9.2.2** the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the **1%** level of significance, to conclude that the mean customer satisfaction for Company **1** is higher than that for Company **2**.

Example 9.2.3

Perform the test of Example **9.2.2** using the p-value approach.

Solution:

The first three steps are identical to those in Example **9.2.2**

- **Step 4**. The observed significance or p-value of the test is the area of the right tail of the standard normal distribution that is cut off by the test statistic $Z = 5.684$. The number **5.684** is too large to appear in Figure 7.1.5, which means that the area of the left tail that it cuts off is **1.0000** to four decimal places. The area that we seek, the area of the right tail, is therefore $1 - 1.0000 = 0.0000$ to four decimal places. See Figure **9.2.3**. That is, p-value=**0.0000** to four decimal places. (The actual value is approximately **0.000000007**.)

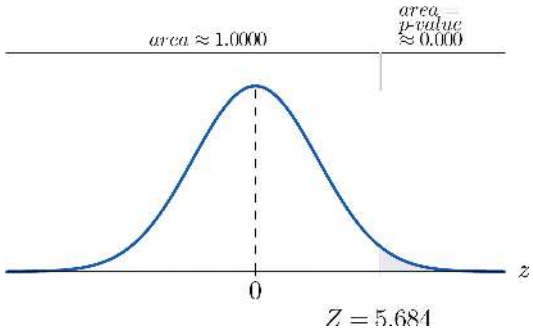


Figure **9.2.3**: ***P*-Value for Example 9.2.3**

- **Step 5**. Since $0.0000 < 0.01$, $p - value < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the **1%** level of significance, to conclude that the mean customer satisfaction for Company **1** is higher than that for Company **2**.

Key Takeaway

- A point estimate for the difference in two population means is simply the difference in the corresponding sample means.
- In the context of estimating or testing hypotheses concerning two population means, “large” samples means that both samples are large.

- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means. The only difference is in the formula for the standardized test statistic.

9.2: Inferences for Two Population Means- Large, Independent Samples is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

- **9.1: Comparison of Two Population Means- Large, Independent Samples** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.3: Inferences for Two Population Means - Unknown Standard Deviations

Skills to Develop

- To learn how to construct a confidence interval for the difference in the means of two distinct populations using independent samples with unknown population standard deviations.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using independent samples with unknown population standard deviations.

Often, the population standard deviation is unknown. In this case, a t distribution is used to draw inferences from these samples, when the samples are independent and drawn from two approximately normally distributed populations. We will assume that the variances for the populations are not necessarily equal.

Confidence Intervals

When the two populations are normally distributed, the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{9.3.1}$$

where the number of degrees of freedom is the smaller value of $n_1 - 1$ or $n_2 - 1$., or

$$\min(n_1 - 1, n_2 - 1). \tag{9.3.2}$$

The samples must be independent, the samples are random samples, and when the sample sizes are less than 30, the populations must be normally distributed or approximately normally distributed.

Example 9.3.1

A software company markets a new computer game with two experimental packaging designs. Design 1 is sent to 11 stores; their average sales the first month is 52 units with sample standard deviation 12 units. Design 2 is sent to 6 stores; their average sales the first month is 46 units with sample standard deviation 10 units. Construct a point estimate and a 95% confidence interval for the difference in average monthly sales between the two package designs. Assume the populations are approximately normal.

Solution:

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 52 - 46 - 6 \tag{9.3.3}$$

In words, we estimate that the average monthly sales for Design 1 is 6 units more per month than the average monthly sales for Design 2.

To apply the formula for the confidence interval (Equation 9.3.1), we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = \min(n_1 - 1, n_2 - 1) = \min(11 - 1, 6 - 1) = \min(10, 5) = 5$ we read that
$t_{0.025} = 2.571$. From the formula we compute

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha}/2\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 6 \pm (2.571)\sqrt{\frac{12^2}{11} + \frac{10^2}{6}} \approx 6 \pm 14.0 \tag{9.3.4}$$

We are 95% confident that the difference in the population means lies in the interval $[-8.0, 20.0]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. Because the interval contains

both positive and negative values the statement in the context of the problem is that we are 95% confident that the average monthly sales for Design 1 is between 20 units higher and 8 units lower than the average monthly sales for Design 2.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means for unknown standard deviations is done precisely as it is done for large samples, using the following standardized test statistic. The same conditions on the populations that were required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: when the standard deviations are unknown

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{9.3.5}$$

The test statistic has Student's t-distribution with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom.

The samples are random, the sample data must be independent, and when the sample sizes are less than 30, the populations must be normal or approximately normal.

Example 9.3.2

Refer to Example 9.3.1 concerning the mean sales per month for the same computer game but sold with two package designs. Test at the 1% level of significance whether the data provide sufficient evidence to conclude that the mean sales per month of the two designs are different. Use the critical value approach.

Solution:

- **Step 1**. The relevant test is

$$H_0 : \mu_1 - \mu_2 = 0 \tag{9.3.6}$$

vs.

$$H_a : \mu_1 - \mu_2 \neq 0 \ @ \ \alpha = 0.01 \tag{9.3.7}$$

- **Step 2**. Since the samples are independent and at least one is less than 30 the test statistic is

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{9.3.8}$$

which has Student's t-distribution with $df = \min(n_1 - 1, n_2 - 1) = \min(11 - 1, 6 - 1) = \min(10, 5) = 5$ degrees of freedom.

- **Step 3**. Inserting the data and the value $D_0 = 0$ into the formula for the test statistic gives

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(52 - 46) - 0}{\sqrt{\frac{12^2}{11} + \frac{10^2}{6}}} = 1.100 \tag{9.3.9}$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2} = \pm t_{0.005}$. From the row in Figure 7.1.6 with the heading $df = 5$ we read off $t_{0.005} = 4.032$. The rejection region is $(-\infty, -4.032] \cup [4.032, \infty)$.

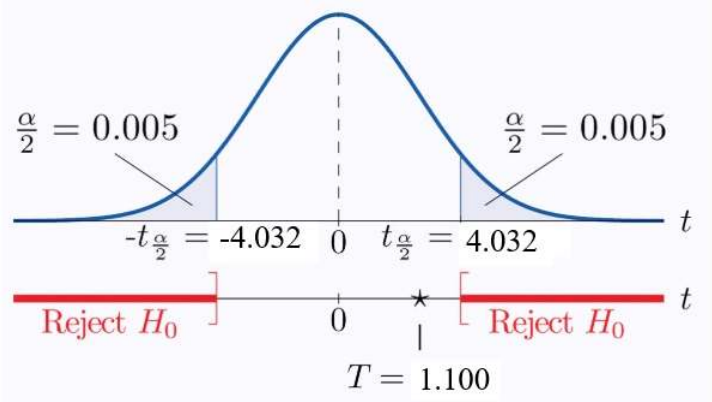


***Figure* 9.3.1**: *Rejection Region and Test Statistic for "Example* **9.3.2**"

- **Step 5**. As shown in Figure **9.3.1** the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the **1%** level of significance, to conclude that the mean sales per month of the two designs are different.

Example 9.3.3

Perform the test of Example **9.3.2** using the p-value approach.

Solution:

The first three steps are identical to those in Example **9.3.2**.

- **Step 4**. Because the test is two-tailed the observed significance or p-value of the test is the double of the area of the right tail of Student'st-distribution, with **5** degrees of freedom, that is cut off by the test statistic $T = 1.100$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 5$, the number **1.100** is less than the number **1.476**, corresponding to $t_{0.100}$. The area cut off by $t = 1.476$ is **0.100**. Since **1.100** is smaller than **1.476**, the area it cuts off must be larger than **0.100**. Thus the p-value (since the area must be doubled) is larger than **0.200**.
- **Step 5**. Since $p > 0.200 > 0.01,\ \ p > \alpha$, so the decision is not to reject the null hypothesis:

  The data do not provide sufficient evidence, at the **1%** level of significance, to conclude that the mean sales per month of the two designs are different.

Notice the conclusion for the critical value method and the p-value method are the same.

Additional Notes

The degrees of freedom given here is a rough estimate. The formula to calculate actual degrees of freedom is used in technology, so if you try these problems on your TI-84 or in Excel, you will get a different degree of freedom for the example problem.

The formula for degrees of freedom for populations with unequal variances is

$$d.f. = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)}. \tag{9.3.10}$$

For hypothesis tests for normally or approximately normally distributed populations where variances are assumed to be equal, the following formula is used for the test statistic.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{9.3.11}$$

This test statistic has Student's t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

Key Takeaway

- In the context of estimating or testing hypotheses concerning two population means when the standard deviations are unknown, we must use the t-distribution unless the samples are very large. When the samples are very large (both are greater than 30), the normal distribution can be used as an approximation to the t-distribution as the differences between them are minimal.
- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.

9.3: Inferences for Two Population Means - Unknown Standard Deviations is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

9.4: Inferences for Two Population Means - Paired Samples

Learning Objectives

- To learn the distinction between independent samples and paired samples.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using paired samples.
- To learn how to perform a test of hypotheses concerning the difference in the means of two distinct populations using paired samples

Suppose chemical engineers wish to compare the fuel economy obtained by two different formulations of gasoline. Since fuel economy varies widely from car to car, if the mean fuel economy of two independent samples of vehicles run on the two types of fuel were compared, even if one formulation were better than the other the large variability from vehicle to vehicle might make any difference arising from difference in fuel difficult to detect. Just imagine one random sample having many more large vehicles than the other. Instead of independent random samples, it would make more sense to select pairs of cars of the same make and model and driven under similar circumstances, and compare the fuel economy of the two cars in each pair. Thus the data would look something like Table **9.4.1**, where the first car in each pair is operated on one formulation of the fuel (call it Type **1** gasoline) and the second car is operated on the second (call it Type **2** gasoline).

Table **9.4.1**: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2
Buick LaCrosse	17.0	17.0
Dodge Viper	13.2	12.9
Honda CR-Z	35.3	35.4
Hummer H 3	13.6	13.2
Lexus RX	32.7	32.5
Mazda CX-9	18.4	18.1
Saab 9-3	22.5	22.5
Toyota Corolla	26.8	26.7
Volvo XC 90	15.1	15.0

The first column of numbers form a sample from Population **1**, the population of all cars operated on Type **1** gasoline; the second column of numbers form a sample from Population **2**, the population of all cars operated on Type **2** gasoline. It would be incorrect to analyze the data using the formulas from the previous section, however, since the samples were not drawn independently. What is correct is to compute the difference in the numbers in each pair (subtracting in the same order each time) to obtain the third column of numbers as shown in Table **9.4.2** and treat the differences as the data. At this point, the new sample of differences $d_1 = 0.0, \cdots, d_9 = 0.1$ in the third column of Table **9.4.2** may be considered as a random sample of size $n = 9$ selected from a population with mean $\mu_d = \mu_1 - \mu_2$. This approach essentially transforms the paired two-sample problem into a one-sample problem as discussed in the previous two chapters.

Table **9.4.2**: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2	Difference
Buick LaCrosse	17.0	17.0	0.0
Dodge Viper	13.2	12.9	0.3
Honda CR-Z	35.3	35.4	−0.1
Hummer H 3	13.6	13.2	0.4
Lexus RX	32.7	32.5	0.2
Mazda CX-9	18.4	18.1	0.3
Saab 9-3	22.5	22.5	0.0
Toyota Corolla	26.8	26.7	0.1
Volvo XC 90	15.1	15.0	0.1

Note carefully that although it does not matter what order the subtraction is done, it must be done in the same order for all pairs. This is why there are both positive and negative quantities in the third column of numbers in Table **9.4.2**.

Confidence Intervals

When the population of differences is normally distributed the following formula for a confidence interval for $\mu_d = \mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Paired Difference Samples

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The number of degrees of freedom is

$$df = n - 1.$$

The population of differences must be *normally distributed*.

Example 9.4.1

Using the data in Table **9.4.1** construct a point estimate and a **95%** confidence interval for the difference in average fuel economy between cars operated on Type **1** gasoline and cars operated on Type **2** gasoline.

Solution

We have referred to the data in Table **9.4.1** because that is the way that the data are typically presented, but we emphasize that with paired sampling one immediately computes the differences, as given in Table **9.4.2**, and uses the differences as the data.

The mean and standard deviation of the differences are

$$\bar{d} = \frac{\sum d}{n} = \frac{1.3}{9} = 0.1\bar{4}$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{1}{n}(\sum d)^2}{n-1}} = \sqrt{\frac{0.41 - \frac{1}{9}(1.3)^2}{8}} = 0.1\bar{6}$$

The point estimate of $\mu_1 - \mu_2 = \mu_d$ is

$$\bar{d} = 0.14$$

In words, we estimate that the average fuel economy of cars using Type 1 gasoline is 0.14 mpg greater than the average fuel economy of cars using Type 2 gasoline.

To apply the formula for the confidence interval, we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = 9 - 1 = 8$ we read that $t_{0.025} = 2.306$. Thus

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}} = 0.14 \pm 2.306\left(\frac{0.1\bar{6}}{\sqrt{9}}\right) \approx 0.14 \pm 0.13$$

We are 95% confident that the difference in the population means lies in the interval $[0.01, 0.27]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_d = \mu_1 - \mu_2$. Stated differently, we are 95% confident that mean fuel economy is between 0.01 and 0.27 mpg greater with Type 1 gasoline than with Type 2 gasoline.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means using paired difference samples is done precisely as it is done for independent samples, although now the null and alternative hypotheses are expressed in terms of μ_d instead of $\mu_1 - \mu_2$. Thus the null hypothesis will always be written

$$H_0 : \mu_d = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \mu_d < D_0$	Left-tailed
$H_a : \mu_d > D_0$	Right-tailed
$H_a : \mu_d \neq D_0$	Two-tailed

The same conditions on the population of differences that was required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested. Here is the standardized test statistic that is used in the test.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Paired Difference Samples

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The test statistic has Student's t-distribution with $df = n - 1$ degrees of freedom.

The population of differences must be normally distributed.

Example 9.4.2: using the critical value approach

Using the data of Table 9.4.2 test the hypothesis that mean fuel economy for Type 1 gasoline is greater than that for Type 2 gasoline against the null hypothesis that the two formulations of gasoline yield the same mean fuel economy. Test at the 5% level of significance using the critical value approach.

Solution

The only part of the table that we use is the third column, the differences.

- **Step 1**. Since the differences were computed in the order $\textbf{Type 1 mpg} - \textbf{Type 2 mpg}$, better fuel economy with Type 1 fuel corresponds to $\mu_d = \mu_1 - \mu_2 > 0$. Thus the test is

$$H_0 : \mu_d = 0$$
vs.
$$H_a : \mu_d > 0 \ @ \ \alpha = 0.05$$

(If the differences had been computed in the opposite order then the alternative hypotheses would have been $H_a : \mu_d < 0$.)

- **Step 2**. Since the sampling is in pairs the test statistic is

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

- **Step 3**. We have already computed $\bar{d}$ and s_d in the previous example. Inserting their values and $D_0 = 0$ into the formula for the test statistic gives

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{0.1\bar{4}}{0.1\bar{6}/\sqrt{3}} = 2.600$$

- **Step 4**. Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $t_\alpha = t_{0.05}$ with 8 degrees of freedom, which from the row labeled $df = 8$ in Figure 7.1.6 we read off as 1.860. The rejection region is $[1.860, \infty)$.
- **Step 5**. As shown in Figure 9.4.1 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

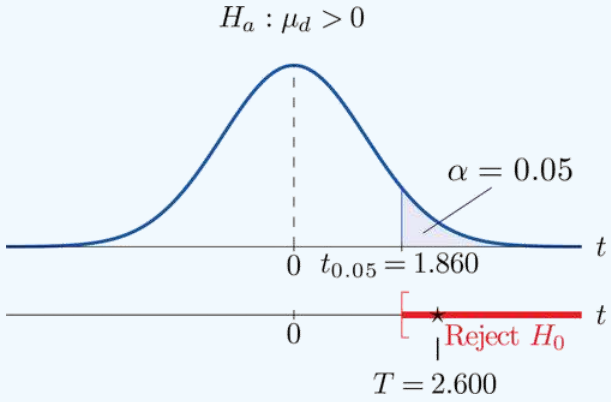


Figure 9.4.1: Rejection Region and Test Statistic for "Example 9.4.2"

The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

✔ Example 9.4.3: using the p-value approach

Perform the test in Example 9.4.2 using the p-value approach.

Solution

The first three steps are identical to those 9.4.2.

- **Step 4**. Because the test is one-tailed the observed significance or p-value of the test is just the area of the right tail of Student's t-distribution, with 8 degrees of freedom, that is cut off by the test statistic $T = 2.600$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 8$, the number 2.600 is between the numbers 2.306 and 2.896, corresponding to $t_{0.025}$ and $t_{0.010}$. The area cut off by $t = 2.306$ is 0.025 and the area cut off by $t = 2.896$ is 0.010. Since 2.600 is between 2.306 and 2.896 the area it cuts off is between 0.025 and 0.010. Thus the p-value is between 0.025 and 0.010. In particular it is less than 0.025. See Figure 9.4.2.

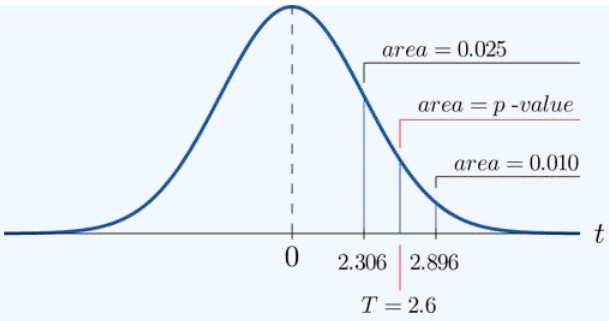


Figure 9.4.2: P-Value for "Example 9.4.3"

- **Step 5**. Since $0.025 < 0.05$, $p < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

The paired two-sample experiment is a very powerful study design. It bypasses many unwanted sources of "statistical noise" that might otherwise influence the outcome of the experiment, and focuses on the possible difference that might arise from the one factor of interest.

If the sample is large (meaning that $n \geq 30$) then in the formula for the confidence interval we may replace $t_{\alpha/2}$ by $z_{\alpha/2}$. For hypothesis testing when the number of pairs is at least 30, we may use the same statistic as for small samples for hypothesis testing, except now it follows a standard normal distribution, so we use the last line of Figure 7.1.6 to compute critical values, and p-values can be computed exactly with Figure 7.1.5, not merely estimated using Figure 7.1.6.

Key Takeaway

- When the data are collected in pairs, the differences computed for each pair are the data that are used in the formulas.
- A confidence interval for the difference in two population means using paired sampling is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means using pair sampling. The only difference is in the formula for the standardized test statistic.

9.4: Inferences for Two Population Means - Paired Samples is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

- **9.3: Comparison of Two Population Means - Paired Samples** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.5: Inferences for Two Population Proportions

Learning Objectives

- To learn how to construct a confidence interval for the difference in the proportions of two distinct populations that have a particular characteristic of interest.
- To learn how to perform a test of hypotheses concerning the difference in the proportions of two distinct populations that have a particular characteristic of interest.

Suppose we wish to compare the proportions of two populations that have a specific characteristic, such as the proportion of men who are left-handed compared to the proportion of women who are left-handed. Figure **9.5.1** illustrates the conceptual framework of our investigation. Each population is divided into two groups, the group of elements that have the characteristic of interest (for example, being left-handed) and the group of elements that do not. We arbitrarily label one population as Population **1** and the other as Population **2**, and subscript the proportion of each population that possesses the characteristic with the number **1** or **2** to tell them apart. We draw a random sample from Population **1** and label the sample statistic it yields with the subscript **1**. Without reference to the first sample we draw a sample from Population **2** and label its sample statistic with the subscript **2**.

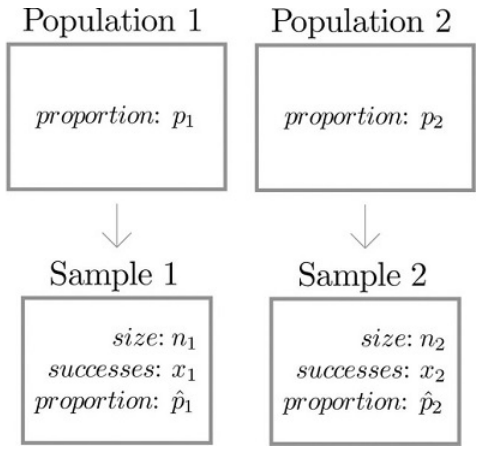


Figure **9.5.1**: Independent Sampling from Two Populations In Order to Compare Proportions

Our goal is to use the information in the samples to estimate the difference $p_1 - p_2$ in the two population proportions and to make statistically valid inferences about it.

Confidence Intervals

Since the sample proportion $\hat{p}_1$ computed using the sample drawn from Population **1** is a good estimator of population proportion p_1 of Population **1** and the sample proportion $\hat{p}_2$ computed using the sample drawn from Population **2** is a good estimator of population proportion p_2 of Population **2**, a reasonable point estimate of the difference $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. In order to widen this point estimate into a confidence interval we suppose that both samples are large, as described in Section 7.3 and repeated below. If so, then the following formula for a confidence interval for $p_1 - p_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The samples must be independent, and *each* sample must be large: each of the intervals

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]$$

and

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.5.1

The department of code enforcement of a county government issues permits to general contractors to work on residential projects. For each permit issued, the department inspects the result of the project and gives a "pass" or "fail" rating. A failed project must be re-inspected until it receives a pass rating. The department had been frustrated by the high cost of re-inspection and decided to publish the inspection records of all contractors on the web. It was hoped that public access to the records would lower the re-inspection rate. A year after the web access was made public, two samples of records were randomly selected. One sample was selected from the pool of records before the web publication and one after. The proportion of projects that passed on the first inspection was noted for each sample. The results are summarized below. Construct a point estimate and a 90% confidence interval for the difference in the passing rate on first inspection between the two time periods.

No public web access	$n_1 = 500 \;\; \hat{p}_1 = 0.67$
Public web access	$n_2 = 100 \;\; \hat{p}_2 = 0.80$

Solution

The point estimate of $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 = 0.67 - 0.80 = -0.13$$

Because the "No public web access" population was labeled as Population 1 and the "Public web access" population was labeled as Population 2, in words this means that we estimate that the proportion of projects that passed on the first inspection increased by 13 percentage points after records were posted on the web.

The sample sizes are sufficiently large for constructing a confidence interval since for sample 1:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.67)(0.33)}{500}} = 0.06$$

so that

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right] = [0.67 - 0.06, 0.67 + 0.06] = [0.61, 0.73] \subset [0, 1]$$

and for sample 2:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.8)(0.2)}{100}} = 0.12$$

so that

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right] = [0.8 - 0.12, 0.8 + 0.12] = [0.68, 0.92] \subset [0, 1]$$

To apply the formula for the confidence interval, we first observe that the 90% confidence level means that $\alpha = 1 - 0.90 = 0.10$ so that $z_{\alpha/2} = z_{0.05}$. From Figure 7.1.6 we read directly that $z_{0.05} = 1.645$. Thus the desired confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{9.5.1}$$

$$= 0.13 \pm 1.645\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}} \tag{9.5.2}$$

$$= -0.13 \pm 0.07 \tag{9.5.3}$$

The 90% confidence interval is $[-0.20, -0.06]$. We are 90% confident that the difference in the population proportions lies in the interval $[-0.20, -0.06]$, in the sense that in repeated sampling 90% of all intervals constructed from the sample data in this manner will contain $p_1 - p_2$. Taking into account the labeling of the two populations, this means that we are 90% confident that the proportion of projects that pass on the first inspection is between 6 and 20 percentage points higher after public access to the records than before.

Hypothesis Testing

In hypothesis tests concerning the relative sizes of the proportions p_1 and p_2 of two populations that possess a particular characteristic, the null and alternative hypotheses will always be expressed in terms of the difference of the two population proportions. Hence the null hypothesis is always written

$$H_0 : p_1 - p_2 = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : p_1 - p_2 < D_0$	Left-tailed
$H_a : p_1 - p_2 > D_0$	Right-tailed
$H_a : p_1 - p_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Proportions

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: each of the intervals

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]$$

and

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.5.2

Using the data of Example 9.5.1, test whether there is sufficient evidence to conclude that public web access to the inspection records has increased the proportion of projects that passed on the first inspection by more than 5 percentage points. Use the critical value approach at the 10% level of significance.

Solution

- **Step 1**. Taking into account the labeling of the populations an increase in passing rate at the first inspection by more than 5 percentage points after public access on the web may be expressed as $p_2 > p_1 + 0.05$, which by algebra is the same as $p_1 - p_2 < -0.05$. This is the alternative hypothesis. Since the null hypothesis is always expressed as an equality, with the same

number on the right as is in the alternative hypothesis, the test is

$$H_0 : p_1 - p_2 = -0.05$$
vs.
$$H_a : p_1 - p_2 < -0.05 \ @ \ \alpha = 0.10$$

- **Step 2**. Since the test is with respect to a difference in population proportions the test statistic is

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

- **Step 3**. Inserting the values given in Example 9.5.1 and the value $D_0 = -0.05$into the formula for the test statistic gives

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} = \frac{(-0.13) - (-0.05)}{\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}}} = -1.770$$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $z_\alpha = -z_{0.10}$. From the last row in Figure 7.1.6 $z_{0.10} = 1.282$, so $-z_{0.10} = -1.282$. The rejection region is $(-\infty, -1.282]$.
- **Step 5**. As shown in Figure 9.5.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 10% level of significance, to conclude that the rate of passing on the first inspection has increased by more than 5 percentage points since records were publicly posted on the web.

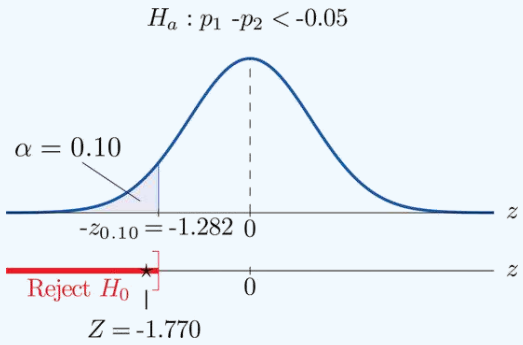


Figure 9.5.2: Rejection Region and Test Statistic for "Example 9.5.2"

Example 9.5.3

Perform the test of Example 9.5.2 using the p-value approach.

Solution

The first three steps are identical to those in Example 9.5.2

- **Step 4**. Because the test is left-tailed the observed significance or p-value of the test is just the area of the left tail of the standard normal distribution that is cut off by the test statistic $Z = -1.770$. From Figure 7.1.5 the area of the left tail determined by -1.77 is 0.0384. The p-value is 0.0384.
- **Step 5**. Since the p-value 0.0384 is less than $\alpha = 0.10$, the decision is to reject the null hypothesis: The data provide sufficient evidence, at the 10% level of significance, to conclude that the rate of passing on the first inspection has increased by more than 5 percentage points since records were publicly posted on the web.

Finally a common misuse of the formulas given in this section must be mentioned. Suppose a large pre-election survey of potential voters is conducted. Each person surveyed is asked to express a preference between, say, Candidate A and Candidate B. (Perhaps "no preference" or "other" are also choices, but that is not important.) In such a survey, estimators $\hat{p}_A$ and $\hat{p}B$ of p_A and p_B can be calculated. It is important to realize, however, that these two estimators were not calculated from two independent samples. While $\hat{p}A - \hat{p}_B$ may be a reasonable estimator of $p_A - p_B$, the formulas for confidence intervals and for the standardized test statistic given in this section are not valid for data obtained in this manner.

Key Takeaway

- A confidence interval for the difference in two population proportions is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population proportion is used to test hypotheses concerning the difference between two population proportions. The only difference is in the formula for the standardized test statistic.

9.5: Inferences for Two Population Proportions is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

- **9.4: Comparison of Two Population Proportions** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

9.6: Which Analysis Should You Conduct?

One of the most important concept that you need to understand is deciding which analysis you should conduct for a particular situation. To help you to figure out the analysis to conduct, there are a series of questions you should ask yourself.

1. Does the problem deal with mean or proportion?
   Sometimes the problem states explicitly the words mean or proportion, but other times you have to figure it out based on the information you are given. If you counted number of individuals that responded in the affirmative to a question, then you are dealing with proportion. If you measured something, then you are dealing with mean.
2. Does the problem have one or two samples?
   So look to see if one group was measured or if two groups were measured. If you have the data sets, then it is usually easy to figure out if there is one or two samples, then there is either one data set or two data sets. If you don't have the data, then you need to decide if the problem describes collecting data from one group or from two groups.
3. If you have two samples, then you need to determine if the samples are independent or dependent.
   If the individuals are different for both samples, then most likely the samples are independent. If you can't tell, then determine if a data value from the first sample influences the data value in the second sample. In other words, can you pair data values together so you can find the difference, and that difference has meaning. If the answer is yes, then the samples are paired. Otherwise, the samples are independent.
4. Does the situation involve a hypothesis test or a confidence interval?
   If the problem talks about "do the data show", "is there evidence of", "test to see", then you are doing a hypothesis test. If the problem talks about "find the value", "estimate the" or "find the interval", then you are doing a confidence interval.

So if you have a situation that has two samples, independent samples, involving the mean, and is a hypothesis test, then you have a two-sample independent t-test. Now you look up the assumptions and the formula or technology process for doing this test. Every hypothesis test involves the same six steps, and you just have to use the correct assumptions and calculations. Every confidence interval has the same five steps, and again you just need to use the correct assumptions and calculations. So this is why it is so important to figure out what analysis you should conduct.

Data Sources:

AP exam scores. (2013, November 20). Retrieved from wiki.stat.ucla.edu/socr/index...08_APExamScore s

Buy sushi grade fish online. (2013, November 20). Retrieved from http://www.catalinaop.com/

Center for Disease Control and Prevention, Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network. (2008). *Autism and developmental disabilities monitoring network-2012*. Retrieved from website: www.cdc.gov/ncbddd/autism/doc...nityReport.pdf

Cholesterol levels after heart attack. (2013, September 25). Retrieved from http://www.statsci.org/data/general/cholest.html

Flanagan, R., Rooney, C., & Griffiths, C. (2005). Fatal poisoning in childhood, england & wales 1968-2000. *Forensic Science International*, 148:121-129, Retrieved from http://www.cdc.gov/nchs/data/ice/fat...ning_child.pdf

Friday the 13th datafile. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Datafil...aythe13th.html

Gettler, L. T., McDade, T. W., Feranil, A. B., & Kuzawa, C. W. (2011). Longitudinal evidence that fatherhood decreases testosterone in human males. *The Proceedings of the National Academy of Sciences, PNAS 2011*, doi: 10.1073/pnas.1105403108 Length of NZ rivers. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/nzrivers.html

Lim, L. L. United Nations, International Labour Office. (2002). *Female labour-force participation*. Retrieved from website: www.un.org/esa/population/pub...ty/RevisedLIMp aper.PDF

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=137

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, 1(2), 108-13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9765732

Population reference bureau. (2013, October 8). Retrieved from http://www.prb.org/DataFinder/Topic/...gs.aspx?ind=25

Seafood online. (2013, November 20). Retrieved from http://www.allfreshseafood.com/

SOCR 012708 id data hotdogs. (2013, November 13). Retrieved from http://wiki.stat.ucla.edu/socr/index...D_Data_HotDogs

SOCR data nips infantvitK shotdata. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...tVitK_ShotData

SOCR data Oct2009 id ni. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index..._Oct2009_ID_NI

Statistics brain. (2013, November 30). Retrieved from http://www.statisticbrain.com/infidelity-statistics/

Student t-distribution. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Stories/student.html

---

This page titled 9.6: Which Analysis Should You Conduct? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

- **9.4: Which Analysis Should You Conduct?** by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

9.E: Hypothesis Testing with Two Samples (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

10.1: Introduction

10.2: Two Population Means with Unknown Standard Deviations

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

a. independent group means, population standard deviations, and/or variances known
b. independent group means, population standard deviations, and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

Exercise 10.2.3

It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

Answer

two proportions

Exercise 10.2.4

A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

Exercise 10.2.5

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

Answer

matched or paired samples

Exercise 10.2.6

The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

Exercise 10.2.7

The average worker in Germany gets eight weeks of paid vacation.

Answer

single mean

Exercise 10.2.8

According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

Exercise 10.2.9

It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

Answer

independent group means, population standard deviations and/or variances unknown

Exercise 10.2.10

The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

Exercise 10.2.11

In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

Answer

two proportions

Exercise 10.2.12

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.

Exercise 10.2.13

It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

Answer

independent group means, population standard deviations and/or variances unknown

Exercise 10.2.14

Varsity athletes practice five times a week, on average.

Exercise 10.2.15

A sample of 12 in-state graduate school programs at school A has a mean tuition of $64,000 with a standard deviation of $8,000. At school B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

Answer

independent group means, population standard deviations and/or variances unknown

Exercise 10.2.16

A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

Exercise 10.2.17

A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

Answer

two proportions

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

Exercise 10.2.18

Are standard deviations known or unknown?

Exercise 10.2.19

What is the random variable?

Answer

The random variable is the difference between the mean amounts of sugar in the two soft drinks.

Exercise 10.2.20

Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

Exercise 10.2.21

Is this a test of means or proportions?

Answer

means

Exercise 10.2.22

State the null and alternative hypotheses.

a. H_0: _____
b. H_a: _____

Exercise 10.2.23

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

two-tailed

Exercise 10.2.24

In symbols, what is the random variable of interest for this test?

Exercise 10.2.25

In words, define the random variable of interest for this test.

Answer

the difference between the mean life spans of whites and nonwhites

Exercise 10.2.26

Which distribution (normal or Student's t) would you use for this hypothesis test?

Exercise 10.2.27

Explain why you chose the distribution you did for Exercise.

Answer

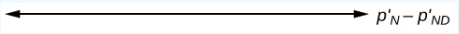
This is a comparison of two population means with unknown population standard deviations.

Exercise 10.2.28

Calculate the test statistic and p-value.

Exercise 10.2.29

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.



Answer

Check student's solution.

Exercise 10.2.30

Find the p-value.

Exercise 10.2.31

At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

Answer

a. Reject the null hypothesis

b. p-value < 0.05
c. There is not enough evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

Exercise 10.2.32

Does it appear that the means are the same? Why or why not?

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Q 10.2.1

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

Q 10.2.2

A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

S 10.2.2

Subscripts: 1: two-year colleges; 2: four-year colleges

a. $H_0 : \mu_1 \geq \mu_2$
b. $H_a : \mu_1 < \mu_2$
c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.
d. Student's-*t*
e. test statistic: -0.2480
f. p-value : 0.4019
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject
   iii. Reason for Decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

Q 10.2.3

At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Q 10.2.4

Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were \$46,100 and \$46,700, respectively. Their standard deviations were \$3,450 and \$4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

S 10.2.4

Subscripts: 1: mechanical engineering; 2: electrical engineering

a. $H_0 : \mu_1 \geq \mu_2$
b. $H_a : \mu_1 < \mu_2$
c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean entry level salaries of mechanical engineers and electrical engineers.
d. t_{108}
e. test statistic: $t = -0.82$
f. p-value : 0.2061
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for Decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

Q 10.2.5

Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from [link] to answer the next four exercises.

Q 10.2.6

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

S 10.2.6

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 \neq \mu_2$
c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean times for completing a lap in races and in practices.
d. $t_{20.32}$
e. test statistic: –4.70
f. p-value : 0.0001
g. Check student's solution.
h.
   i. α : 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Q 10.2.7

Repeat the test in Exercise 10.83, but use Lap 5 data this time.

Q 10.2.8

Repeat the test in Exercise 10.83, but this time combine the data from Laps 1 and 5.

S 10.2.8

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 \neq \mu_2$
c. is the difference between the mean times for completing a lap in races and in practices.
d. $t_{40.94}$
e. test statistic: –5.08
f. p-value : 0
g. Check student's solution.
h.
   i. α : 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Q 10.2.9

In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question. "Does Terri Vogel drive faster in races than she does in practices?"

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to answer the next two exercises.

Q 10.2.10

The **exact** distribution for the hypothesis test is:

a. the normal distribution
b. the Student's *t*-distribution
c. the uniform distribution
d. the exponential distribution

Q 10.2.11

If the level of significance is 0.05, the conclusion is:

a. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
b. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
c. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
d. Unable to determine

Q 10.2.12

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. A concluding statement is:

a. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
b. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

Q 10.2.13

Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

S 10.2.13

Test: two independent sample means, population standard deviations unknown.

Random variable:

$$\bar{X}_1 - \bar{X}_2 \tag{9.E.1}$$

Distribution: $H_0 : \mu_1 = \mu_2 H_a : \mu_1 < \mu_2$ The mean age of entering prostitution in Canada is lower than the mean age in the United States.

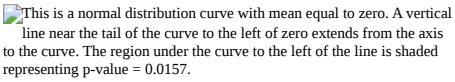


Figure 10.2.1.

Graph: left-tailed

p-value : 0.0151

Decision: Do not reject H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

Q 10.2.14

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Q 10.2.15

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

a. $\mu_{day} > \mu_{night}$
b. $\mu_{day} < \mu_{night}$
c. $\mu_{day} = \mu_{night}$
d. $\mu_{day} \neq \mu_{night}$

S 10.2.15

d

10.3: Two Population Means with Known Standard Deviations

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. Table shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

Exercise 10.3.2

What is the random variable?

Answer

The difference in mean speeds of the fastball pitches of the two pitchers

Exercise 10.3.3

State the null and alternative hypotheses.

Exercise 10.3.4

What is the test statistic?

Answer

-2.46

Exercise 10.3.5

What is the p-value?

Exercise 10.3.6

At the 1% significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez's fastball is faster than Wesley's.

Use the following information to answer the next five exercises. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

Exercise 10.3.7

Is the population standard deviation known or unknown?

Exercise 10.3.8

State the null and alternative hypotheses.

Answer

Subscripts: 1 = Food, 2 = No Food

$H_0 : \mu_1 \leq \mu_2$

$H_a : \mu_1 > \mu_2$

Exercise 10.3.9

What is the p-value?

Exercise 10.3.10

Draw the graph of the ***p*-value**.

Answer

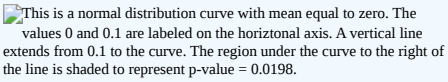


Figure 10.3.3.

Exercise 10.3.11

At the 1% significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. 15 pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample Mean Melting Temperatures (°F)	Population Standard Deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

Exercise 10.3.12

State the null and alternative hypotheses.

Answer

Subscripts: 1 = Gamma, 2 = Zeta

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

Exercise 10.3.13

Is this a right-, left-, or two-tailed test?

Exercise 10.3.14

What is the ***p*-value**?

Answer

0.0062

Exercise 10.3.15

Draw the graph of the ***p*-value**.

Exercise 10.3.16

At the 1% significance level, what is your conclusion?

Answer

There is sufficient evidence to reject the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Q 10.3.1

A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

Q 10.3.2

Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether or not you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

S 10.3.3

Subscripts: 1 = boys, 2 = girls

a. $H_0 : \mu_1 \leq \mu_2$
b. $H_a : \mu_1 > \mu_2$
c. The random variable is the difference in the mean auto insurance costs for boys and girls.
d. normal
e. test statistic: $z = 2.50$
f. p-value : 0.0062
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

Q 10.3.4

A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1,011, respectively. The population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

Q 10.3.5

Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirty-one non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that the population standard deviations are known to be six and three, respectively. Conduct a hypothesis test to evaluate the manufacturers claim.

S 10.3.5

Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

a. $H_0 : \mu_1 \geq \mu_2$
b. $H_a : \mu_1 < \mu_2$
c. The random variable is the difference in the mean miles per gallon of non-hybrid sedans and hybrid sedans.
d. normal
e. test statistic: 6.36
f. p-value : 0
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $< \alpha$
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

Q 10.3.6

A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

Q 10.3.7

One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). Table contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

S 10.3.7

a. $H_0 : \mu_d = 0$
b. $H_a : \mu_d < 0$
c. The random variable X_d is the average difference between husband's and wife's satisfaction level.
d. t_9
e. test statistic: $t = -1.86$
f. p-value : 0.0479
g. Check student's solution
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis, but run another test.
   iii. Reason for Decision: p-value $< \alpha$
   iv. Conclusion: This is a weak test because alpha and the p-value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

10.4: Comparing Two Independent Population Proportions

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

Exercise 10.4.2

Is this a test of means or proportions?

Exercise 10.4.3

What is the random variable?

Answer

$P'_{OS_1} - P'_{OS_2} =$ difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

Exercise 10.4.4

State the null and alternative hypotheses.

Exercise 10.4.5

What is the p-value?

Answer

0.1018

Exercise 10.4.6

What can you conclude about the two operating systems?

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Exercise 10.4.7

Is this a test of means or proportions?

Answer

proportions

Exercise 10.4.8

State the null and alternative hypotheses.

a. H_0: _____
b. H_a: _____

https://stats.libretexts.org/@go/page/10989

Exercise 10.4.9

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Answer

right-tailed

Exercise 10.4.10

What is the random variable of interest for this test?

Exercise 10.4.11

In words, define the random variable for this test.

Answer

The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

Exercise 10.4.12

Which distribution (normal or Student's *t*) would you use for this hypothesis test?

Exercise 10.4.13

Explain why you chose the distribution you did for the Exercise 10.56.

Answer

Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

Exercise 10.4.14

Calculate the test statistic.

Exercise 10.4.15

Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.

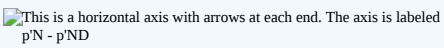


Figure 10.4.5.

Answer

Check student's solution.

Exercise 10.4.16

Find the p-value.

Exercise 10.4.17

At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

Answer

a. Reject the null hypothesis.
b. p-value $< \alpha$
c. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

Exercise 10.4.18

Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Q 10.4.1

A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

Q 10.4.2

We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the whites and the blacks races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24. We will let female suicide victims be our population.

S 10.4.2

a. $H_0 : P_W = P_B$
b. $H_a : P_W \neq P_B$
c. The random variable is the difference in the proportions of white and black suicide victims, aged 15 to 24.
d. normal for two proportions
e. test statistic: –0.1944
f. p-value : 0.8458
g. Check student's solution.
h.
   i. α : 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of white and black female suicide victims, aged 15 to 24, are different.

Q 10.4.3

Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{(larger+smallerdimension}{largerdimension}\right)$ was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

Q 10.4.4

A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

S 10.4.4

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

a. $H_0 : p_1 = p_2$
b. $H_a : p_1 \neq p_2$
c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
d. normal for two proportions
e. test statistic: 4.29
f. p-value : 0.00002
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

Use the following information to answer the next three exercises. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system . It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

- "2011" subscript: 2011 group.
- "2010" subscript: 2010 group

Q 10.4.5

This is:

a. a test of two proportions
b. a test of two independent means
c. a test of a single mean
d. a test of matched pairs.

Q 10.4.6

An appropriate null hypothesis is:

a. $p_{2011} \leq p_{2010}$
b. $p_{2011} \geq p_{2010}$
c. $\mu_{2011} \leq \mu_{2010}$
d. $p_{2011} > p_{2010}$

S 10.4.6

a

Q 10.4.7

The p-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.

Q 10.4.8

Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders.

S 10.4.9

Test: two independent sample proportions.

Random variable: $p'_1 - p'_2$

Distribution:

$H_0 : p_1 = p_2$

$H_a : p_1 \neq p_2$

The proportion of eReader users is different for the 16- to 29-year-old users from that of the 30 and older users.

Graph: two-tailed

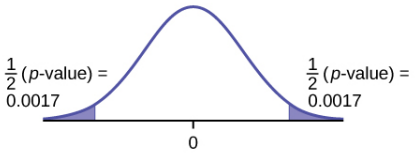


Figure 10.4.1.

p-value : 0.0033

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of eReader users 16 to 29 years old is different from the proportion of eReader users 30 and older.

Q 10.4.10

are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in Table. Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

Q 10.4.11

Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. Table details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 year olds	30 years old and older
Own a Tablet	69	231
Sample Size	628	2,309

S 10.4.11

Test: two independent sample proportions

Random variable: $p'_1 - p'_2$

Distribution:

$H_0 : p_1 = p_2$

$H_a : p_1 > p_2$

A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Graph: right-tailed

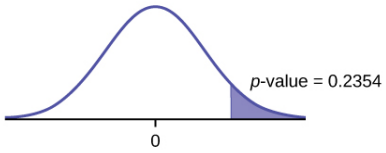


Figure 10.4.2.

p-value : 0.2354

Decision: Do not reject the H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Q 10.4.12

A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

Q 10.4.13

While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

S 10.4.13

Subscripts: 1: men; 2: women

a. $H_0 : p_1 \leq p_2$
b. $H_a : p_1 > p_2$
c. $P'_1 - P\_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
d. normal for two proportions
e. test statistic: 0.22
f. **p-value : 0.4133**

https://stats.libretexts.org/@go/page/10989

g. Check student's solution.

h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for Decision: $p\text{-value} > \alpha$
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

Q 10.4.14

We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

Q 10.4.15

Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

S 10.4.15

a. $H_0 : p_1 = p_2$
b. $H_a : p_1 \neq p_2$
c. $P_1' - P\_2$ is the difference between the proportions of men and women that have at least one pierced ear.
d. normal for two proportions
e. test statistic: –4.82
f. $p\text{-value} : 0$
g. Check student's solution.

h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for Decision: $p\text{-value} < \alpha$
iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

Q 10.4.16

Use the data sets found in [link] to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

Q 10.4.17

"To Breakfast or Not to Breakfast?" by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in Table, solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

S 10.4.17

a. $H_0 : \mu_d = 0$
b. $H_a : \mu_d > 0$
c. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
d. t_9
e. test statistic: 4.8963
f. p-value : 0.0004
g. Check student's solution.
h.
   i. α : 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision:p-value $< \alpha$
   iv. Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

10.5: Matched or Paired Samples

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in Table. The "before" value is matched to an "after" value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

Exercise 10.5.4

What is the random variable?

Answer

the mean difference of the system failures

Exercise 10.5.5

State the null and alternative hypotheses.

Exercise 10.5.6

What is the p-value?

Answer

0.0067

Exercise 10.5.7

Draw the graph of the p-value.

Exercise 10.5.8

What conclusion can you draw about the software patch?

Answer

With a p-value0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

Use the following information to answer next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

Exercise 10.5.9

State the null and alternative hypotheses.

Exercise 10.5.10

What is the p-value?

Answer

0.0021

Exercise 10.5.11

What is the sample mean difference?

Exercise 10.5.12

Draw the graph of the ***p*-value**.

Answer

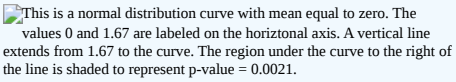


Figure 10.5.4.

Exercise 10.5.13

What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

Exercise 10.5.14

State the null and alternative hypotheses.

Answer

$H_0 : \mu_d \geq 0$

$H_a : \mu_d < 0$

Exercise 10.5.15

What is the test statistic?

Exercise 10.5.16

What is the ***p*-value**?

Answer

0.0699

Exercise 10.5.17

What is the sample mean difference?

Exercise 10.5.18

What is the conclusion?

Answer

We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

Bringing It Together

Use the following information to answer the next ten exercises. indicate which of the following choices best identifies the hypothesis test.

a. independent group means, population standard deviations and/or variances known
b. independent group means, population standard deviations and/or variances unknown
c. matched or paired samples
d. single mean
e. two proportions
f. single proportion

Exercise 10.5.19

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

Exercise 10.5.20

A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

Answer

e

Exercise 10.5.21

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from nine males and 16 females.

Exercise 10.5.22

A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

Answer

d

Exercise 10.5.23

A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and one year, respectively.

Exercise 10.5.24

According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

Answer

f

Exercise 10.5.25

According to a recent study, U.S. companies have a mean maternity-leave of six weeks.

Exercise 10.5.26

A recent drug survey showed an increase in use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.

Answer

e

Exercise 10.5.27

A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

Exercise 10.5.28

University of Michigan researchers reported in the *Journal of the National Cancer Institute* that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.

Answer

f

Exercise 10.5.29

Lesley E. Tan investigated the relationship between left-handedness vs. right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The

experiment produced the means and standard deviations shown Table. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

a. Two independent means, normal distribution
b. Two independent means, Student's-t distribution
c. Matched or paired samples, Student's-t distribution
d. Two population proportions, normal distribution

Exercise 10.5.30

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as Table.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

This is:

a. a test of two independent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

Answer

a

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Q 10.5.1

Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded in Table. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

S 10.5.1

p-value $= 0.1494$

At the 5% significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

Use the following information to answer the next two exercises. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript $t =$ treated patient and $ut =$ untreated patient.

Q 10.5.2

The appropriate hypotheses are:

a. $H_0 : p_t < p_{ut}$ and $H_a : p_t \geq p_{ut}$
b. $H_0 : p_t \leq p_{ut}$ and $H_a : p_t > p_{ut}$
c. $H_0 : p_t = p_{ut}$ and $H_a : p_t \neq p_{ut}$
d. $H_0 : p_t = p_{ut}$ and $H_a : p_t < p_{ut}$

Q 10.5.3

If the p-value is 0.0062 what is the conclusion (use $\alpha = 0.05$)?

a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.
c. There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients who develop AIDS after four years.
d. There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.

S 10.5.3

b

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a “biofeedback exercise program.” Six subjects were randomly selected and blood

pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after - before) producing the following results: $\bar{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

Q 10.5.4

The distribution for the test is:

a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. $N\left(-10.2, \frac{8.4}{\sqrt{6}}\right)$

Q 10.5.5

If $\alpha = 0.05$, the p-value and the conclusion are

a. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
b. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
c. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
d. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.

S 10.5.5

c

Q 10.5.6

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

The correct decision is:

a. Reject H_0.
b. Do not reject the H_0.

Q 10.5.7

A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are in Table.

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

S 10.5.7

Test: two matched pairs or paired samples (*t*-test)

Random variable: $\bar{X}_d$

Distribution: t_{12}

$H_0 : \mu_d = 0 H_a : \mu_d > 0$

The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.

Graph: right-tailed

p-value : 0.0004

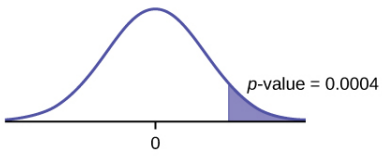


Figure 10.5.1.

Decision: Reject H_0

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

Q 10.5.8

A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in Table. Test at the 1% level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

Q 10.5.9

A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are in Table.

Northeastern States	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

S 10.5.9

Test: matched or paired samples (*t*-test)

Difference data: $\{-0.9, -3.7, -3.2, -0.5, 0.6, -1.9, -0.5, 0.2, 0.6, 0.4, 1.7, -2.4, 1.8\}$

Random Variable: $\bar{X}_d$

Distribution: $H_0 : \mu_d = 0 H_a : \mu_d < 0$

https://stats.libretexts.org/@go/page/10989

The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.

Graph: left-tailed.

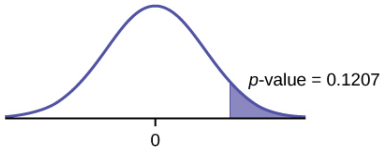


Figure 10.5.2.

p-value : 0.1207

Decision: Do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

10.6: Hypothesis Testing for Two Means and Two Proportions

This page titled 9.E: Hypothesis Testing with Two Samples (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **10.E: Hypothesis Testing with Two Samples (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

10: Correlation and Regression

Regression analysis is a statistical process for estimating the relationships among variables and includes many techniques for modeling and analyzing several variables. When the focus is on the relationship between a dependent variable and one or more independent variables.

10.1: Prelude to Linear Regression and Correlation

10.1.1: Review- Linear Equations

10.1.2: Scatter Plots

10.2: Testing the Significance of the Correlation Coefficient

10.3: The Regression Equation

10.3.1: Prediction

10.4: Outliers

10.E: Linear Regression and Correlation (Optional Exercises)

1.E: Linear Equations (Optional Exercises)

1.E: Outliers (Optional Exercises)

1.E: Prediction (Optional Exercises)

1.E: Scatter Plots (Optional Exercises)

1.E: Testing the Significance of the Correlation Coefficient (Optional Exercises)

1.E: The Regression Equation (Optional Exercise)

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 10: Correlation and Regression is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Prelude to Linear Regression and Correlation

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it? In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.



Figure 10.1.1: Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables. In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

This page titled 10.1: Prelude to Linear Regression and Correlation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1.1: Review- Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx$$

where a and b are constant numbers. The variable x is the *independent variable*, and y is the *dependent variable*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 10.1.1.1

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

Exercise 10.1.1.1

Is the following an example of a linear equation?

$$y = -0.125 - 3.5x$$

Answer

yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example 10.1.1.2

Graph the equation $y = -1 + 2x$.

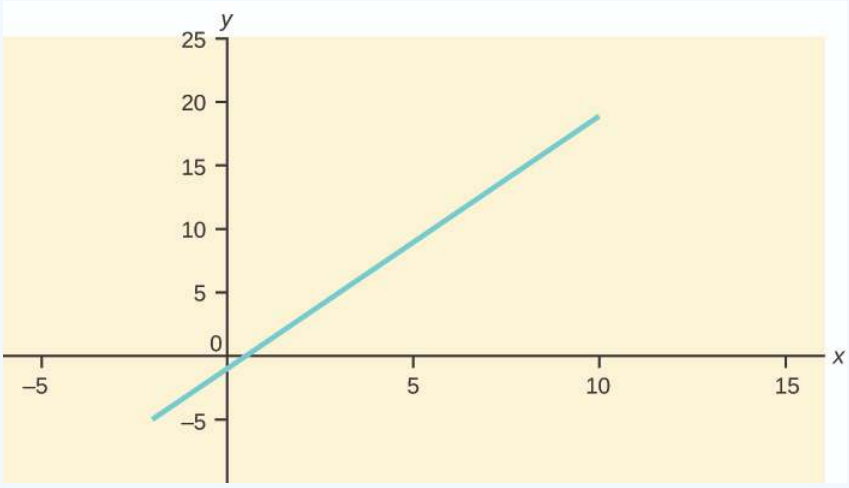


Figure 10.1.1.1.

Exercise 10.1.1.2

Is the following an example of a linear equation? Why or why not?

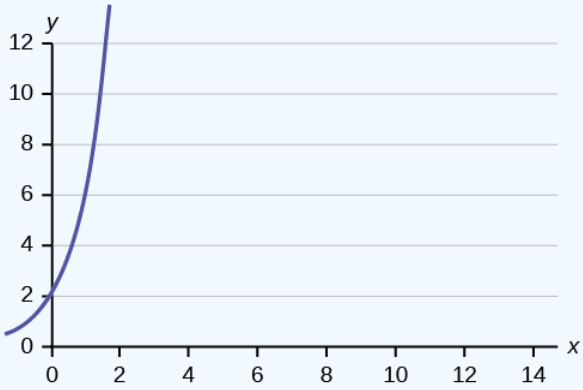


Figure **10.1.1.2**.

Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

Example 10.1.1.3

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

Let $x =$ the number of hours it takes to get the job done.

Let $y =$ the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Exercise 10.1.1.3

Emma’s Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer

$y = 50 + 20x$

Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, $b =$ slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

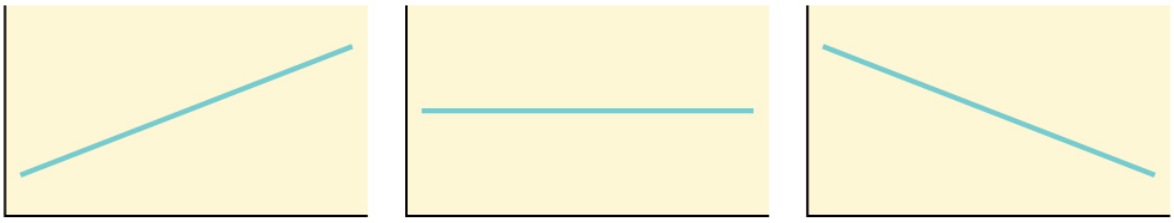


Figure 10.1.1.3:. Three possible graphs of $y = a + bx$ (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 10.1.1.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Exercise 10.1.1.4

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

Summary

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a coefficient) is called the **slope**. The constant a is called the y-intercept.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The **y-intercept** is used to describe the dependent variable when the independent variable equals zero.

Formula Review

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

This page titled 10.1.1: Review- Linear Equations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1.2: Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a *scatter plot*. The following example illustrates a scatter plot.

Example 10.1.2.1

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let $x =$ the year and let $y =$ the number of m-commerce users, in millions.

Table **10.1.2.1**: Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

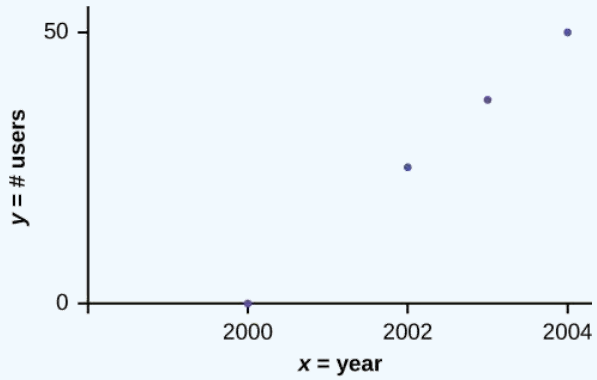


Figure **10.1.2.1**: Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

a. Enter your X data into list L1 and your Y data into list L2.
b. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
c. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
d. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
e. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
f. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
g. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Exercise 10.1.2.1

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Answer

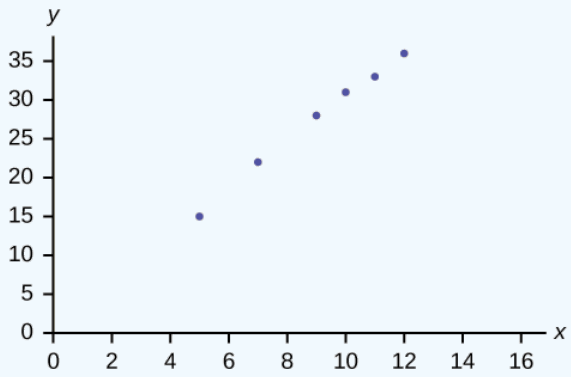


***Figure* 10.1.2.2**

Yes, Amelia's assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the *direction of a relationship* between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the *strength of the relationship* by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatter plot, you want to notice the *overall pattern* and any *deviations* from the pattern. The following scatterplot examples illustrate these concepts.

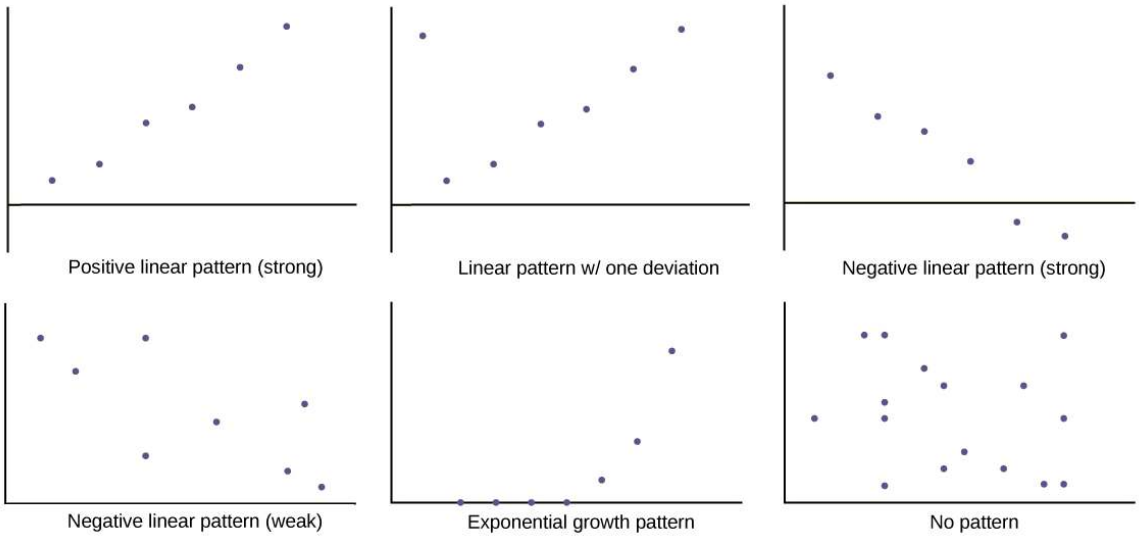


Figure **10.1.2.3**:

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called linear regression. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

Summary

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

This page titled 10.1.2: Scatter Plots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.2: Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together. We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

NOTE

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- **Null Hypothesis:** $H_0 : \rho = 0$
- **Alternate Hypothesis:** $H_a : \rho \neq 0$

WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis H_0:** The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis H_a:** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:There are two methods of making the decision. The two methods are equivalent and give the same result.

- **Method 1: Using the p-value**

- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

NOTE

Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

Using the TI83, 83+, 84, 84+ CALCULATOR

To calculate the p-value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq 0$"

The output screen shows the p-value on the line that reads "$p =$".

(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p-value. The following describes the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.

An alternative way to calculate the p-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.

- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 10.2.1

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

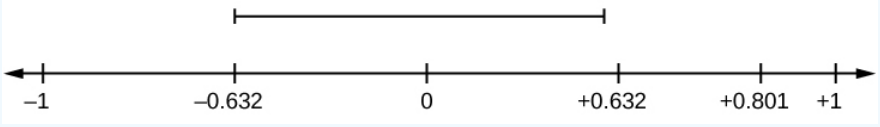


Figure 10.2.1. r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Exercise 10.2.1

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576. Can the line be used for prediction? Why or why not?

Answer

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r >$ the positive critical value.

Example 10.2.2

Suppose you computed $r = –0.624$ with 14 data points. $df = 14–2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line can be used for prediction

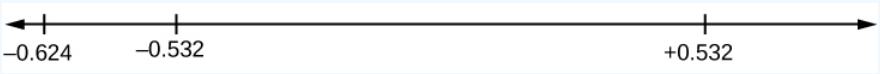


Figure 10.2.2. $r = -0.624 - 0.532$. Therefore, r is significant.

Exercise 10.2.2

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example 10.2.3

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.



Figure 10.2.3. $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

? Exercise 10.2.3

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is $= 0.707$. Can the line be used for prediction? Why or why not?

Answer

Yes, the line can be used for prediction, because $r <$ the negative critical value.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the third exam/final exam example. The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$
- $\alpha = 0.05$
- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are -0.602 and $+0.602$
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

✓ Example 10.2.4

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

a. $r = -0.567$ and the sample size, n, is 19. The $df = n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
b. $r = 0.708$ and the sample size, n, is 9. The $df = n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
c. $r = 0.134$ and the sample size, n, is 14. The $df = 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
d. $r = 0$ and the sample size, n, is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

? Exercise 10.2.4

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Answer

No, the line cannot be used for prediction no matter what the sample size is.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatter plot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

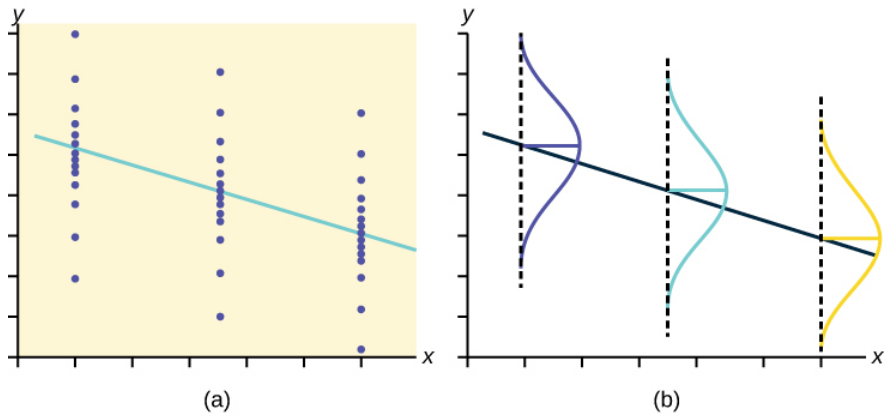


Figure **10.2.4**. The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Summary

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis $H_0 : \rho =$ *hypothesized value*, use a linear regression t-test. The

most common null hypothesis is $H_0 : \rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx \tag{10.2.1}$$

where

$$a = y\text{-intercept} \tag{10.2.2}$$

$$b = \text{slope} \tag{10.2.3}$$

Standard deviation of the residuals:

$$s = \sqrt{\frac{SSE}{n-2}} \tag{10.2.4}$$

where

$$SSE = \text{sum of squared errors} \tag{10.2.5}$$

$$n = \text{the number of data points} \tag{10.2.6}$$

This page titled 10.2: Testing the Significance of the Correlation Coefficient is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.3: The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a Line of Best Fit **or** Least-Squares Line.

COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y -intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 10.3.1

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

1a: Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

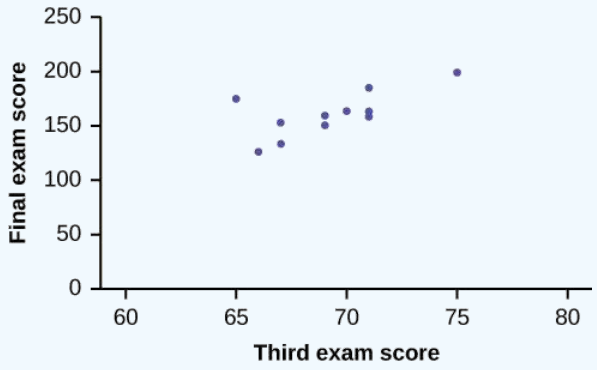


Figure **10.3.1**: Scatter plot showing the scores on the final exam based on scores from the third exam.

Exercise 10.3.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Answer

$\hat{y} = 127.24–1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read "y **hat**" and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

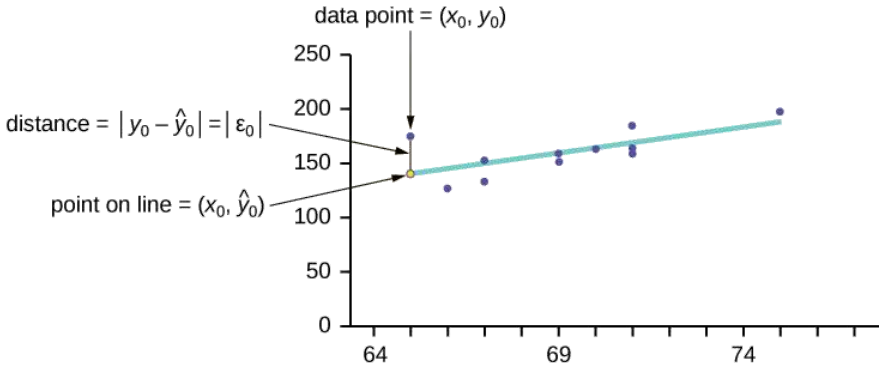


Figure **10.3.2**

The term $y_0–\hat{y}_0 = \varepsilon_0$ is called the **"error" or** residual. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in Figure, $y_0–\hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\varepsilon =$ the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + ... + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \quad (10.3.1)$$

Equation10.3.1 is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx \quad (10.3.2)$$

where

- $a = \bar{y} - b\bar{x}$ and
- $b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$.

The sample means of the x values and the x values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where $s_y =$ the standard deviation of the y values and $s_x =$ the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

Least Square Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line** .

> **Note**
>
> Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third-exam/final-exam example is as follows:

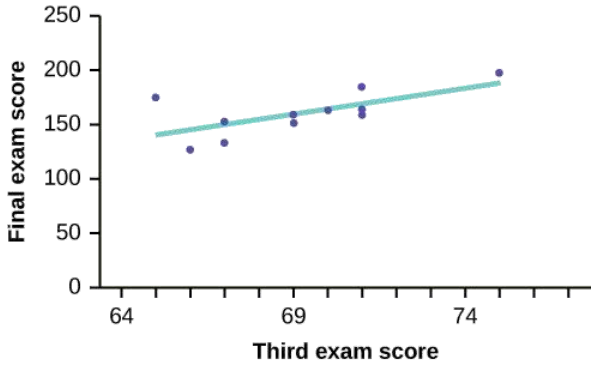


Figure **10.3.3**

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{10.3.3}$$

REMINDER

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

Understanding Slope

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Using the Linear Regression T Test: LinRegTTest

a. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x, y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
b. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
c. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
d. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER
e. Leave the line for "RegEq:" blank
f. Highlight Calculate and press ENTER.

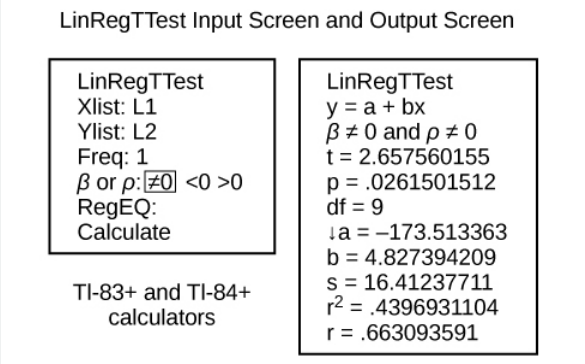


Figure 10.3.4

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
8. To graph the best-fit line, press the "$Y =$" key and type the equation $-173.5 + 4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

Note

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

a. Make sure you have done the scatter plot. Check it on your screen.
b. Go to LinRegTTest and enter the lists.
c. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
d. Press $Y =$ (you will see the regression equation).
e. Press GRAPH. The line will be drawn."

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y. The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{10.3.4}$$

where $n =$ the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

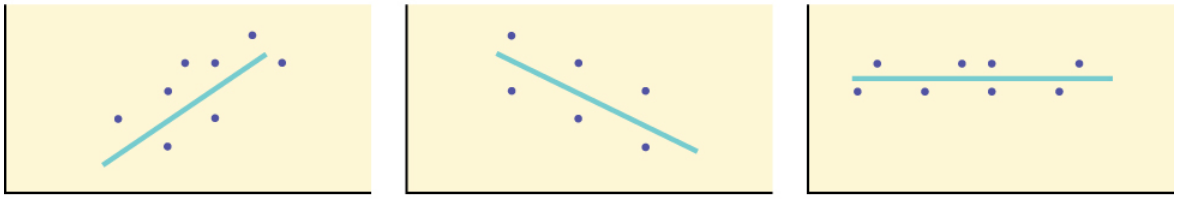
What the VALUE of r tells us:

- The value of r is always between –1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.
- If $r = 0$ there is absolutely no linear relationship between x and y **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen.

What the SIGN of r tells us:

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

> **Note**
>
> Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**



(a) Positive correlation (b) Negative correlation (c) Zero correlation

Figure 10.3.5: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called *the coefficient of determination* and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation $(1 - 0.44 = 0.56)$ in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Summary

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Glossary

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{10.3.5}$$

where n is the number of data points. The coefficient cannot be more than 1 or less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

This page titled 10.3: The Regression Equation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.3.1: Prediction

Recall the third exam/final exam example. We examined the scatter plot and showed that the correlation coefficient is significant. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores (x-values) range from 65 to 75. Since 73 is between the x-values 65 and 75, substitute $x = 73$ into the equation. Then:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 10.3.1.1

Recall the third exam/final exam example.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?
b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Answer

a. 145.27

b. The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.)

To understand really how unreliable the prediction can be outside of the observed x-values observed in the data, make the substitution $x = 90$ into the equation.

$$\hat{y} = -173.51 + 4.83(90) = 261.19$$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

The process of predicting inside of the observed x values observed in the data is called *interpolation*. The process of predicting outside of the observed x-values observed in the data is called *extrapolation*.

Exercise 10.3.1.1

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Answer

86.5

Summary

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

If your hypothesis test does not show a significant correlation (if r is not strong), then you cannot use the line of best fit to predict anything. Instead, the best predicted value for a specific x value is the mean of the y values of the original data set.

References

1. Data from the Centers for Disease Control and Prevention.
2. Data from the National Center for HIV, STD, and TB Prevention.
3. Data from the United States Census Bureau. Available online at www.census.gov/compendia/stat...atalities.html
4. Data from the National Center for Health Statistics.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 10.3.1: Prediction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.4: Outliers

In some data sets, there are values (*observed data points*) called outliers. *Outliers* are observed data points that are far from the least squares line. They have large "errors", where the "error" or residual is the vertical distance from the line to the point. Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called influential points. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatter plot and best fit-line. However, we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier.** The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

Example 10.4.1

In the third exam/final exam example, you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Answer

Graphical Identification of Outliers

With the TI-83, 83+, 84+ graphing calculators, it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to 2*s* or more, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find $s = 16.412$.

Line $Y2 = -173.5 + 4.83x - 2(16.4)$ and line $Y3 = -173.5 + 4.83x + 2(16.4)$

where $\hat{y} = -173.5 + 4.83x$ is the line of best fit. $Y2$ and $Y3$ have the same slope as the line of best fit.

Graph the scatterplot with the best fit line in equation $Y1$, then enter the two extra lines as $Y2$ and $Y3$ in the "$Y =$" equation editor and press ZOOM 9. You will find that the only data point that is not between lines $Y2$ and $Y3$ is the point $x = 65$, $y = 175$. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than two standard deviations away from the best-fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

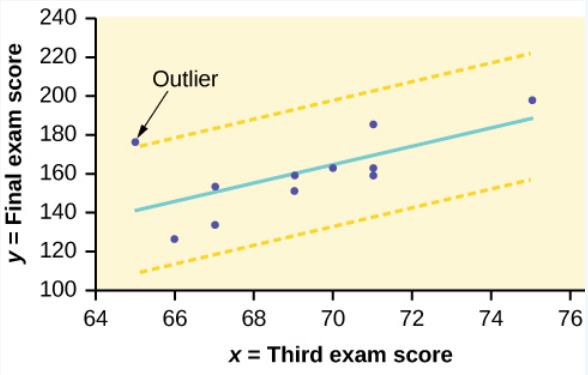


Figure 12.7.1.

? Exercise 10.4.1

Identify the potential outlier in the scatter plot. The standard deviation of the residuals or errors is approximately 8.6.

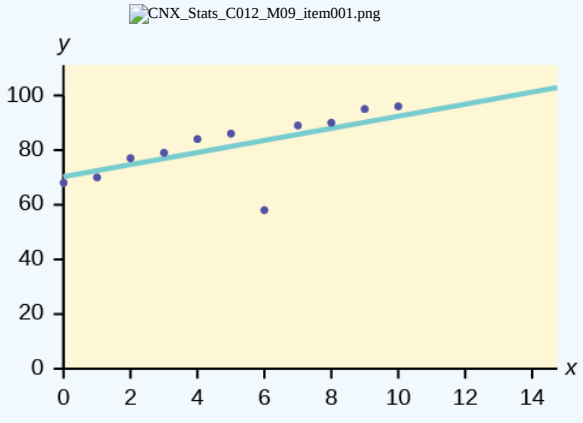


Figure 12.7.2.

Answer

The outlier appears to be at (6, 58). The expected y value on the line for the point (6, 58) is approximately 82. Fifty-eight is 24 units from 82. Twenty-four is more than two standard deviations ($2s = (2)(8.6) = 17.2$). So 82 is more than two standard deviations from 58, which makes $(6, 58)$ a potential outlier.

Numerical Identification of Outliers

In the table below, the first two columns are the third-exam and final-exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, have been calculated in the fourth column of the table: observed y value−predicted y value $= y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values where $n =$ the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{SSE}{n-2}}$$

NOTE

We divide by $(n–2)$ because the regression model involves two estimates.

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, the calculator function LinRegTTest found $s = 16.4$ as the standard deviation of the residuals 35; –17; 16; –6; –19; 9; 3; –1; –10; –9; –1 .

x	y	$\hat{y}$	$y–\hat{y}$
65	175	140	175 – 140 = 35
67	133	150	133 – 150= –17
71	185	169	185 – 169 = 16
71	163	169	163 – 169 = –6
66	126	145	126 – 145 = –19
75	198	189	198 – 189 = 9
67	153	150	153 – 150 = 3
70	163	164	163 – 164 = –1
71	159	169	159 – 169 = –10
69	151	160	151 – 160 = –9
69	159	160	159 – 160 = –1

We are looking for all data points for which the residual is greater than $2s = 2(16.4) = 32.8$ or less than $–32.8$. Compare these values to the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line?

Numerically and graphically, we have identified the point (65, 175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.

Compute a new best-fit line and correlation coefficient using the ten remaining points

On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$$\hat{y} = –355.19 + 7.39x$$

and

$$r = 0.9121$$

The new line with $r = 0.9121$ is a stronger correlation than the original ($r = 0.6631$) because $r = 0.9121$ is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating s and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, square each $|y–\hat{y}|$

The squares are 35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2

Then, add (sum) all the $|y–\hat{y}|$ squared terms using the formula

$$\sum_{i=11}^{11} (|y_i - \hat{y}_i|)^2 = \sum_{i-1}^{11} \varepsilon_i^2$$

Recall that

$$\begin{aligned} y_i - \hat{y}_i &= \varepsilon_i \\ &= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2 \\ &= 2440 = SSE. \end{aligned}$$

The result, SSE is the Sum of Squared Errors.

Next, calculate s, the standard deviation of all the $y - \hat{y} = \varepsilon$ values where $n =$ the total number of data points.

The calculation is

$$s = \sqrt{\frac{SSE}{n-2}}.$$

For the third exam/final exam problem:

$$s = \sqrt{\frac{2440}{11-2}} = 16.47.$$

Next, multiply s by 2:

$$(2)(16.47) = 32.94$$

32.94 is 2 standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $2s$, then we would consider the data point to be "too far" from the line of best fit. We call that point a potential outlier.

For the example, if any of the $|y–\hat{y}|$ values are **at least** 32.94, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam problem, all the $|y–\hat{y}|$'s are less than 31.29 except for the first one which is 35.

$35 > 31.29$ That is, $|y–\hat{y}| \geq (2)(s)$

The point which corresponds to $|y–\hat{y}| = 35$ is $(65, 175)$. **Therefore, the data point $(65, 175)$ is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

NOTE

When outliers are deleted, the researcher should either record that data was deleted, and why, or the researcher should provide results both with and without the deleted data. If data is erroneous and the correct values are known (e.g., student one actually scored a 70 instead of a 65), then this correction can be made to the data.

The next step is to compute a new best-fit line using the ten remaining points. The new line of best fit and the correlation coefficient are:

$$\hat{y} = -355.19 + 7.39x$$

and

$$r = 0.9121$$

Example 10.4.2

Using this new line of best fit (based on the remaining ten data points in the third exam/final exam example), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Answer

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Exercise 10.4.2

The data points for a study that was done are as follows: (1, 5), (2, 7), (2, 6), (3, 9), (4, 12), (4, 13), (5, 18), (6, 19), (7, 12), and (7, 21). Remove the outlier and recalculate the line of best fit. Find the value of $\hat{y}$ when $x = 10$.

Answer

$\hat{y} = 1.04 + 2.96x; 30.64$

Example 10.4.3: The Consumer Price Index

The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Data

x	y	x	y
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

a. Draw a scatterplot of the data.
b. Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
c. Draw the line on the scatterplot.
d. Find the correlation coefficient. Is it significant?
e. What is the average CPI for the year 1990?

Answer

a. See Figure.
b. $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.

c. $r = 0.8694$

d. The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since 0.8694 > 0.532, r is significant.

$$\hat{y} = -3204 + 1.662(1990) = 103.4\text{CPI}$$

e. Using the calculator LinRegTTest, we find that $s = 25.4$; graphing the lines $Y2 = -3204 + 1.662X–2(25.4)$ and $Y3 = -3204 + 1.662X + 2(25.4)$ shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)

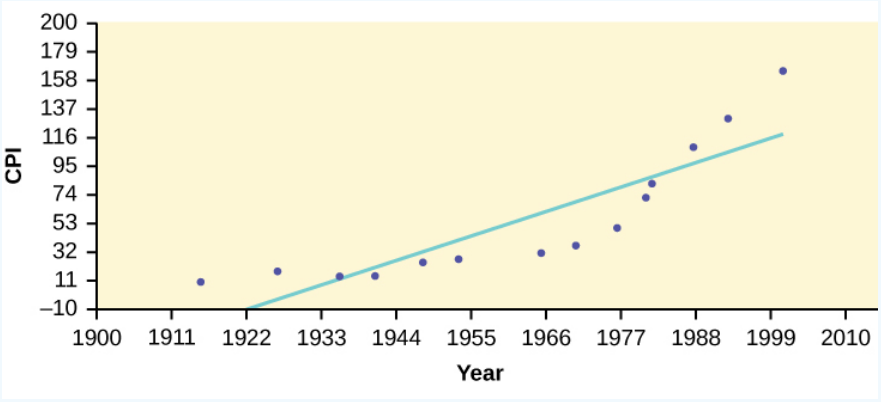


Figure 12.7.3.

NOTE

In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years: 2004: $\text{CPI} = 188.9$; 2008: $\text{CPI} = 215.3$; 2011: $\text{CPI} = 224.9$. See how it affects the model. (Check: $\hat{y} = -4436 + 2.295x$; $r = 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

Exercise 10.4.3

The following table shows economic development measured in per capita income PCINC.

Year	PCINC	Year	PCINC
1870	340	1920	1050
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

a. What are the independent and dependent variables?

b. Draw a scatter plot.

c. Use regression to find the line of best fit and the correlation coefficient.
d. Interpret the significance of the correlation coefficient.
e. Is there a linear relationship between the variables?
f. Find the coefficient of determination and interpret it.
g. What is the slope of the regression equation? What does it mean?
h. Use the line of best fit to estimate PCINC for 1900, for 2000.
i. Determine if there are any outliers.

Answer a

The independent variable (x) is the year and the dependent variable (y) is the per capita income.

Answer b

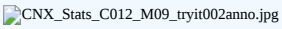


Figure 12.7.4.

Answer c

$\hat{y} = 18.61x - 34574$; $r = 0.9732$

Answer d

At $df = 8$, the critical value is 0.632. The r value is significant because it is greater than the critical value.

Answer e

There does appear to be a linear relationship between the variables.

Answer f

The coefficient of determination is 0.947, which means that 94.7% of the variation in PCINC is explained by the variation in the years.

Answer g and h

The slope of the regression equation is 18.61, and it means that per capita income increases by $18.61 for each passing year. $\hat{y} = 785$ when the year is 1900, and $\hat{y} = 2,646$ when the year is 2000.

Answer i

There do not appear to be any outliers.

95% Critical Values of the Sample Correlation Coefficient Table

Degrees of Freedom: n	Critical Values: (+ and –)
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273

Degrees of Freedom: n–2	Critical Values: (+ and –)
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Summary

To determine if a point is an outlier, do one of the following:

1. Input the following equations into the TI 83, 83+,84, 84+:

$$y_1 = a + bx$$

$$y_2 = a + bx + 2s$$

$$y_3 = a + bx - 2s$$

   where s is the standard deviation of the residuals

   If any point is above y_2 or below y_3 then the point is considered to be an outlier.
2. Use the residuals and compare their absolute values to $2s$ where s is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to $2s$, then the corresponding point is an outlier.

Note: The calculator function LinRegTTest (STATS TESTS LinRegTTest) calculates s.

References

1. Data from the House Ways and Means Committee, the Health and Human Services Department.
2. Data from Microsoft Bookshelf.
3. Data from the United States Department of Labor, the Bureau of Labor Statistics.
4. Data from the Physician's Handbook, 1990.
5. Data from the United States Department of Labor, the Bureau of Labor Statistics.

Glossary

Outlier

an observation that does not fit the rest of the data

This page titled 10.4: Outliers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.E: Linear Regression and Correlation (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

12.1: Introduction

12.2: Linear Equations

Q 12.2.1

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

S 12.2.1

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Q 12.2.2

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

12.3: Scatter Plots

Q 12.3.1

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. Table shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

S 12.3.1

Check student's solution.

Q 12.3.2

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Q 12.3.3

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

S 12.3.3

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Q 12.3.4

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

Q 12.3.5

If there are 15 data points in a set of data, what is the number of degree of freedom?

S 12.3.5

13

12.4: The Regression Equation

Q 12.4.1

What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Q 12.4.2

Explain what it means when a correlation has an r^2 of 0.72.

S 12.4.2

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Q 12.4.3

Can a coefficient of determination be negative? Why or why not?

12.5: Testing the Significance of the Correlation Coefficient

Q 12.5.1

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

S 12.5.1

We do not reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

Q 12.5.2

If there are 15 data points in a set of data, what is the number of degree of freedom?

12.6: Prediction

Q 12.6.1

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

a. For each age group, pick the midpoint of the interval for the *x* value. (For the 75+ group, use 80.)

b. Using “ages” as the independent variable and “Number of driver deaths per 100,000” as the dependent variable, make a scatter plot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 and 60.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least squares (best-fit) line? Interpret the slope.

S 12.6.1

a.

Age	Number of Driver Deaths per 100,000
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

b. Check student’s solution.
c. $\hat{y} = 35.5818045 - 0.19182491x$
d. $r = -0.57874$

For four df and alpha $= 0.05$, the LinRegTTest gives $p\text{-value} = 0.2288$ so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.

Using the table of critical values for the correlation coefficient, with four *df*, the critical value is 0.811. The correlation coefficient $r = -0.57874$ is not less than –0.811, so we do not reject the null hypothesis.

e. if age = 40, $\hat{y}\ (\text{deaths}) = 35.5818045–0.19182491(40) = 27.9$

if age = 60, $\hat{y}\ (\text{deaths}) = 35.5818045–0.19182491(60) = 24.1$

f. For entire dataset, there is a linear relationship for the ages up to age 74. The oldest age group shows an increase in deaths from the prior group, which is not consistent with the younger ages.
g. $\text{slope} = -0.19182491$

Q 12.6.2

Table shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in Table that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Q 12.6.3

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given in Table

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.

c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page ten and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
g. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

S 12.6.3

a. We wonder if the better discounts appear earlier in the book so we select page as X and discount as Y.
b. Check student's solution.
c. $\hat{y} = 17.21757 - 0.01412x$
d. $r = -0.2752$

For seven df and $\alpha = 0.05$, using LinRegTTest $p\text{-value} = 0.4736$ so we do not reject; there is a not a significant linear relationship between page and discount.

Using the table of critical values for the correlation coefficient, with seven df, the critical value is 0.666. The correlation coefficient $xi = -0.2752$ is not less than 0.666 so we do not reject.

e. page 10: 17.08 page 70: 16.23
f. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount.
g. page 200: 14.39
h. No, using the regression equation to predict for page 200 is extrapolation.
i. $\text{slope} = -0.01412$

As the page number increases by one page, the discount decreases by $0.01412

Q 12.6.4

Table gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

a. Decide which variable should be the independent variable and which should be the dependent variable.

b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from part f different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

Q 12.6.5

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

S 12.6.5

a. Year is the independent or x variable; the number of letters is the dependent or y variable.
b. Check student's solution.
c. no
d. $\hat{y} = 47.03 - 0.0216x$
e. -0.4280
f. 6; 5
g. No, the relationship does not appear to be linear; the correlation is not significant.

h. current year: 2013: 3.55 or four letters; this is not an appropriate use of the least squares line. It is extrapolation.

12.7: Outliers

Q 12.7.1

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

a. Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data in Table, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Q 12.7.2

Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return: 74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38

Percent new: 5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator’s regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

S 12.7.2

a. Check student's solution.
b. Check student's solution.
c. The slope of the regression line is -0.3179 with a y-intercept of 32.966. In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 31% new sparrow hawks, which doesn't make sense since if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 0.3179%.
d. If we examine $r2$, we see that only 50.238% of the variation in the percent of new birds is explained by the model and the correlation coefficient, $r = 0.71$ only indicates a somewhat strong correlation between returning and new percentages.
e. The ordered pair $(66, 6)$ generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.98%. If we remove this data pair, we see only an adjusted slope of -0.2723 and an adjusted intercept of 30.606. In other words, even though this data generates the largest residual, it is not an outlier, nor is the data pair an influential point.
f. If there are 70% returning birds, we would expect to see $y = -0.2723(70) + 30.606 = 0.115$ or 11.5% new birds in the colony.

Q 12.7.3

The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Yearly wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of alcohol consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.4

The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

S 12.7.4

a. Check student's solution.
b. Check student's solution.
c. We have a slope of -1.4946 with a y-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the y-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim 2,000 meters, the less effort the heart puts out), the y-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
d. Since only 1.5% of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
e. The point $(34.72, 124)$ generates the largest residual of -11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes 1.6914 with the y-intercept changing to 83.694. While the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the y-intercept becomes more meaningful.

Q 12.7.5

A researcher is investigating whether non-white minorities commit a disproportionate number of homicides. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are white males.

White Males	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

a. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context.
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation r
   iv. The coefficient of determination $r2$.
d. Do the data provide convincing evidence that there is a linear relationship between the number of white males in the population and the homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.6

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

S 12.7.6

If we remove the two service academies (the tuition is $0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

12.8: Regression (Distance from School)

12.9: Regression (Textbook Cost)

12.10: Regression (Fuel Efficiency)

This page titled 10.E: Linear Regression and Correlation (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **12.E: Linear Regression and Correlation (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

1.E: Linear Equations (Optional Exercises)

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

Exercise 12.2.5

What are the dependent and independent variables?

Answer

dependent variable: fee amount; independent variable: time

Exercise 12.2.6

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise 12.2.7

Graph the equation from Exercise.

Answer

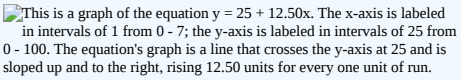


Figure **1.*E*.4**.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise 12.2.8

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise 12.2.9

Graph the equation from Exercise.

Answer

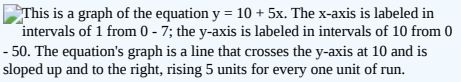


Figure **1.*E*.5**.

Exercise 12.2.10

Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise 12.2.11

Which of the following equations are linear?

a. $y = 6x + 8$
b. $y + 7 = 3x$
c. $y - x = 8x^2$
d. $4y = 8$

Answer

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Exercise 12.2.12

Does the graph show a linear equation? Why or why not?

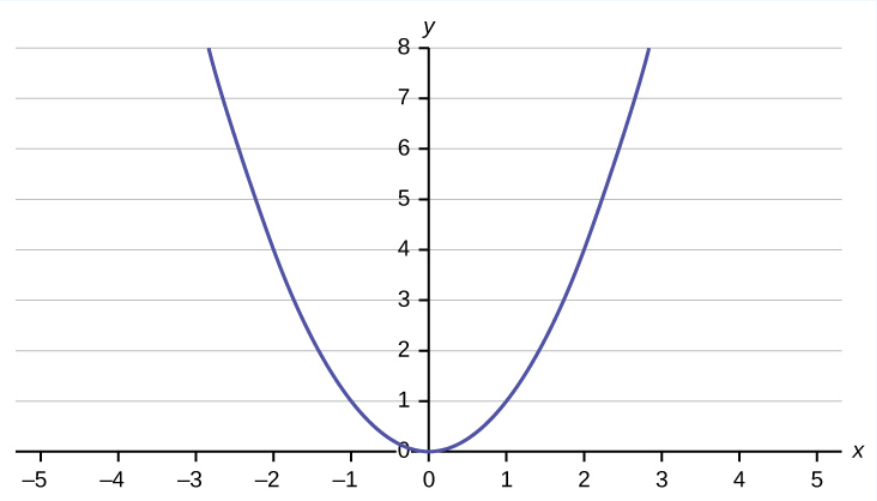


Figure 1.E.6.

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

https://stats.libretexts.org/@go/page/10993

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Exercise 12.2.13

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Answer

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise 12.2.14

What are the independent and dependent variables?

Exercise 12.2.15

What is the *y*-intercept and what is the slope? Interpret them using complete sentences.

Answer

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.

Exercise 12.2.16

What are the independent and dependent variables?

Exercise 12.2.17

How many pounds of soil does the shoreline lose in a year?

Answer

12,000 pounds of soil

Exercise 12.2.18

What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15–1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise 12.2.19

What are the slope and *y*-intercept? Interpret their meaning.

Answer

The slope is -1.5 ($b = -1.5$). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

Exercise 12.2.19

If you owned this stock, would you want a positive or negative slope? Why?

1.E: Linear Equations (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Outliers (Optional Exercises)

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

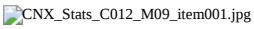


Figure $1.E.5$.

Exercise 12.7.4

Do there appear to be any outliers?

Answer

Yes, there appears to be an outlier at $(6, 58)$.

Exercise 12.7.5

A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

Exercise 12.7.6

What effect did the potential outlier have on the line of best fit?

Answer

The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

Exercise 12.7.7

Are you more or less confident in the predictive ability of the new line of best fit?

Exercise 12.7.8

The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?

Answer

$s = 1.75$

Exercise 12.7.9

The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

Bring It Together

Exercise 12.7.10

The average number of people in a family that received welfare for various years is given in Table.

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

a. Using “year” as the independent variable and “welfare family size” as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
e. Based on the data in Table, is there a linear relationship between the year and the average number of people in a welfare family?
f. Using the least-squares line, estimate the welfare family sizes for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
g. Are there any outliers in the data?
h. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least squares (best-fit) line? Interpret the slope.

? Exercise 12.7.11

The percent of female wage and salary workers who are paid hourly rates is given in Table for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

a. Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?

c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated percents for 1991 and 1988.
f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
g. Are there any outliers in the data?
h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a. Check student's solution.
b. yes
c. $\hat{y} = -266.8863 + 0.1656x$
d. 0.9448; Yes
e. $62.8233; 62.3265$
f. yes
g. yes; $(1987, 62.7)$
h. 72.5937; no
i. $slope = 0.1656$.

As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in Table.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Exercise 12.7.12

a. Using “size” as the independent variable and “cost” as the dependent variable, draw a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. If the laundry detergent were sold in a 40-ounce size, find the estimated cost.
f. If the laundry detergent were sold in a 90-ounce size, find the estimated cost.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the given data?
i. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would you cost? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.13

a. Complete Table for the cost per ounce of the different sizes.
b. Using “size” as the independent variable and “cost per ounce” as the dependent variable, draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
g. If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the the data?
j. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost per ounce? Why or why not?
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a.

Size (ounces)	Cost ($)	cents/oz
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

b. Check student’s solution.
c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
d. $\hat{y} = 20.2368 - 0.0819x$
e. $r = -0.8086$
f. 40-oz: 16.96 cents/oz
g. 90-oz: 12.87 cents/oz
h. The relationship is not linear; the least squares line is not appropriate.
i. no outliers
j. No, you would be extrapolating. The 300-oz size is outside the range of x.
k. $slope = -0.08194$; for each additional ounce in size, the cost per ounce decreases by 0.082 cents.

Exercise 12.7.14

According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a next taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.15

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a. Size is x, the independent variable, price is y, the dependent variable.
b. Check student's solution.
c. The relationship does not appear to be linear.
d. $\hat{y} = -745.252 + 54.75569x$
e. $r = 0.8944$, yes it is significant
f. 32-inch: \$1006.93, 50-inch: \$1992.53
g. No, the relationship does not appear to be linear. However, r is significant.
h. yes, the 60-inch TV
i. For each additional inch, the price increases by \$54.76

Exercise 12.7.16

Table shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.17

State	# letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. What are the independent and dependent variables?
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in part f to plot the least-squares line on your graph from part b.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
k. Delete “Hawaii” and substitute “Alaska” for it. Alaska is the forty-ninth, state with an area of 656,424 square miles.
l. Calculate the new least-squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that’s the case?
n. Do you think that, in general, newer states are larger than the original states?

Answer

a. Let rank be the independent variable and area be the dependent variable.
b. Check student’s solution.
c. There appears to be a linear relationship, with one outlier.
d. $\hat{y}\ (\text{area}) = 24177.06 + 1010.478x$
e. $r = 0.50047$, r is not significant so there is no relationship between the variables.
f. Alabama: 46407.576 Colorado: 62575.224
g. Alabama estimate is closer than Colorado estimate.
h. If the outlier is removed, there is a linear relationship.
i. There is one outlier (Hawaii).

j. rank 51: 75711.4; no

k.

Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Alaska	6	1959	51	656,424
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

l. $\hat{y} = -87065.3 + 7828.532x$

m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.

n. yes, with the exception of Hawaii

1.E: Outliers (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Prediction (Optional Exercises)

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$$\hat{y} = 101.32 + 2.48x \tag{1.E.1}$$

where $\hat{y}$ is in thousands of dollars.

Exercise 12.6.2

What would you predict the sales to be on day 60?

Answer

$250,120

Exercise 12.6.3

What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$$\hat{y} = 1350 - 1.2x \tag{1.E.2}$$

where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

Exercise 12.6.4

How many acres will be left to mow after 20 hours of work?

Answer

1,326 acres

Exercise 12.6.5

How many acres will be left to mow after 100 hours of work?

Exercise 12.6.7

How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

Answer

1,125 hours, or when $x = 1,125$

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

? Exercise 12.6.8

Graph “year” versus “# AIDS cases diagnosed” (plot the scatter plot). Do not include pre-1981 data.

? Exercise 12.6.9

Perform linear regression. What is the linear equation? Round to the nearest whole number.

Answer

Check student’s solution.

Exercise 12.6.10

Write the equations:

a. Linear equation: _____
b. $a =$ _____
c. $b =$ _____
d. $r =$ _____
e. $n =$ _____

Exercise 12.6.11

Solve.

a. When $x = 1985$, $\hat{y} =$ _____
b. When $x = 1990$, $\hat{y} =$_____
c. When $x = 1970$, $\hat{y} =$_____ Why doesn't this answer make sense?

Answer

a. When $x = 1985$, $\hat{y} = 25,52$
b. When $x = 1990$, $\hat{y} = 34,275$
c. When $x = 1970$, $\hat{y} = -725$ Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, where we are predicting AIDS cases diagnosed.

Exercise 12.6.11

Does the line seem to fit the data? Why or why not?

Exercise 12.6.12

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Also, the correlation $r = 0.4526$. If r is compared to the value in the 95% Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But the scatter plot indicates otherwise.

Exercise 12.6.13

Plot the two given points on the following graph. Then, connect the two points to form the regression line.

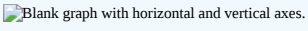


Figure 1.E.1.

Obtain the graph on your calculator or computer.

Exercise 12.6.14

Write the equation: $\hat{y} =$ _____

Answer

$\hat{y} = 3,448,225 + 1750x$

Exercise 12.6.15

Hand draw a smooth curve on the graph that shows the flow of the data.

Exercise 12.6.16

Does the line seem to fit the data? Why or why not?

Answer

There was an increase in AIDS cases diagnosed until 1993. From 1993 through 2002, the number of AIDS cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

Exercise 12.6.17

Do you think a linear fit is best? Why or why not?

Exercise 12.6.18

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Since there is no linear association between year and # of AIDS cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable “causes” the other variable.

Exercise 12.6.19

Graph “year” vs. “# AIDS cases diagnosed.” Do not include pre-1981. Label both axes with words. Scale both axes.

Exercise 12.6.20

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding to four decimal places:

Answer

We don’t know if the pre-1981 data was collected from a single year. So we don’t have an accurate x value for this figure.

Regression equation: Misplaced '#'

	Coefficients
Intercept	–3,448,225
x Variable 1	1,749.777

Exercise 12.6.21

Calculate the following:

a. $a =$ _____
b. $b =$ _____
c. correlation = _____

d. $n =$ _____

1.E: Prediction (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Scatter Plots (Optional Exercises)

Exercise 1.E.1

Does the scatter plot appear linear? Strong or weak? Positive or negative?

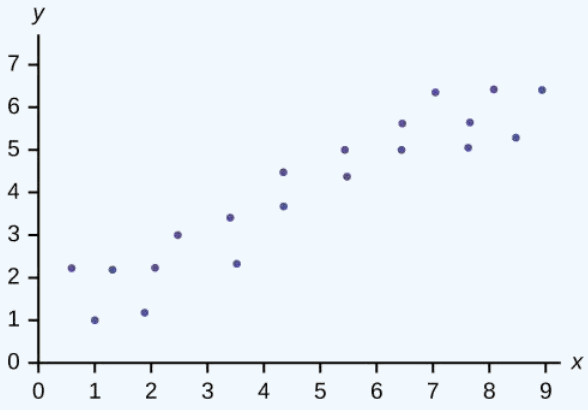


Figure 1.E.4

Answer

The data appear to be linear with a strong, positive correlation.

Exercise 1.E.3

Does the scatter plot appear linear? Strong or weak? Positive or negative?

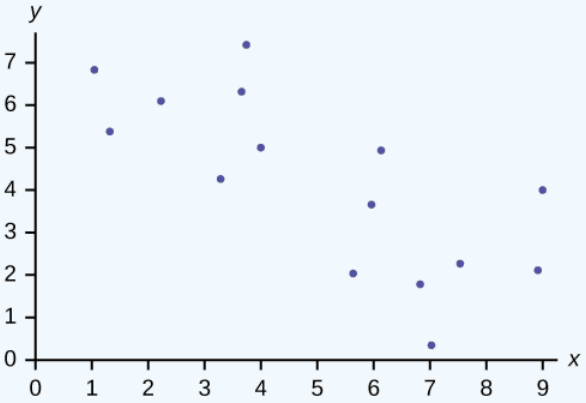


Figure 1.E.5

Exercise 1.E.4

Does the scatter plot appear linear? Strong or weak? Positive or negative?

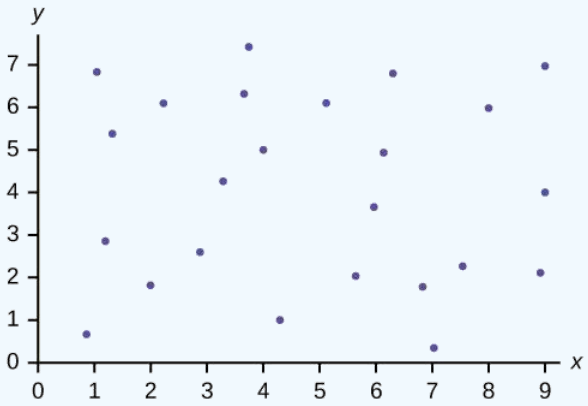


Figure **1.*E*.6**

Answer

The data appear to have no correlation.

1.E: Scatter Plots (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: Testing the Significance of the Correlation Coefficient (Optional Exercises)

Exercise 1.E.5

When testing the significance of the correlation coefficient, what is the null hypothesis?

Exercise 1.E.6

When testing the significance of the correlation coefficient, what is the alternative hypothesis?

Answer

$H_a : \rho \neq 0$

Exercise 1.E.7

If the level of significance is 0.05 and the p-value is 0.04, what conclusion can you draw?

1.E: Testing the Significance of the Correlation Coefficient (Optional Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

1.E: The Regression Equation (Optional Exercise)

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Exercise 12.4.2

Draw a scatter plot of the data.

Exercise 12.4.3

Use regression to find the equation for the line of best fit.

Answer

$\hat{y} = 2.23 + 1.99x$

Exercise 12.4.4

Draw the line of best fit on the scatter plot.

Exercise 12.4.5

What is the slope of the line of best fit? What does it represent?

Answer

The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

Exercise 12.4.6

What is the y-intercept of the line of best fit? What does it represent?

Exercise 12.4.7

What does an r value of zero mean?

Answer

It means that there is no correlation between the data sets.

Exercise 12.4.8

When $n = 2$ and $r = 1$, are the data significant? Explain.

Exercise 12.4.9

When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

1.E: The Regression Equation (Optional Exercise) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

11: Chi-Square and Analysis of Variance (ANOVA)

A chi-squared test is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true.

11.1: Prelude to The Chi-Square Distribution

11.1.1: Facts About the Chi-Square Distribution

11.2: Goodness-of-Fit Test

11.3: Tests Using Contingency tables

11.3.1: Test of Independence

11.3.2: Test for Homogeneity

11.3.3: Comparison of the Chi-Square Tests

11.4: Prelude to F Distribution and One-Way ANOVA

11.4.1: One-Way ANOVA

11.4.2: The F Distribution and the F-Ratio

11.4.3: Facts About the F Distribution

11.4.4: How to Use Microsoft Excel® for Regression Analysis

11.E: F Distribution and One-Way ANOVA (Optional Exercises)

11.E: The Chi-Square Distribution (Optional Exercises)

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11: Chi-Square and Analysis of Variance (ANOVA) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.1: Prelude to The Chi-Square Distribution

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests.

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.



Figure **11.1.1**: The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

In this chapter, you will learn the three major applications of the chi-square distribution:

a. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
b. the test of independence, which determines if events are independent, such as in the movie example
c. the test of a single variance, which tests variability, such as in the coffee example

Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see [link]). TI-83+ and TI-84 calculator instructions are included in the text.

COLLABORATIVE CLASSROOM EXERCISE

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

This page titled 11.1: Prelude to The Chi-Square Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.1.1: Facts About the Chi-Square Distribution

The notation for the chi-square distribution is:

$$\chi \sim \chi^2_{df} \tag{11.1.1.1}$$

where df = degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is

$$\sigma = \sqrt{2(df)}. \tag{11.1.1.2}$$

The random variable is shown as χ^2, but may be any upper case letter. The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + ... + (Z_k)^2 \tag{11.1.1.3}$$

a. The curve is nonsymmetrical and skewed to the right.
b. There is a different chi-square curve for each df.

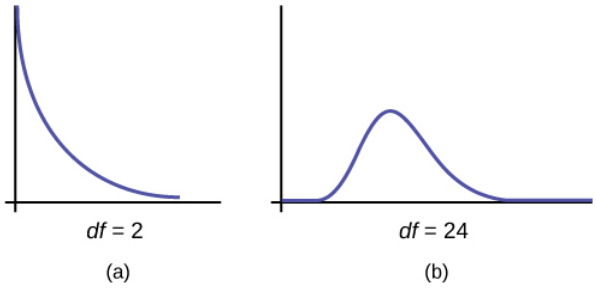


Figure 11.1.1.1

c. The test statistic for any test is always greater than or equal to zero.
d. When $df > 90$, the chi-square curve approximates the normal distribution. For $\chi \sim \chi^2_{1,000}$ the mean, $\mu = df = 1,000$ and the standard deviation, $\mu = \sqrt{2(1,000)}$. Therefore, $X \sim N(1,000, 44.7)$, approximately.
e. The mean, μ, is located just to the right of the peak.

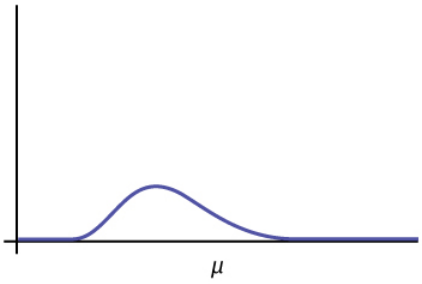


Figure 11.1.1.2

References

1. Data from *Parade Magazine*.
2. “HIV/AIDS Epidemiology Santa Clara County.”Santa Clara County Public Health Department, May 2011.

Review

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom df in a given problem. The random variable in the chi-square distribution is the sum of squares of df standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom df. For $df > 90$, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero. Such application tests are almost always right-tailed tests.

Formula Review

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_{df})^2 \quad (11.1.1.4)$$

chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2(df)}$ Chi-Square distribution population standard deviation

? Exercise 11.1.1.1

If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

Answer

mean $= 25$ and standard deviation $= 7.0711$

? Exercise 11.1.1.2

If $df > 90$, the distribution is _____. If $df = 15$, the distribution is _____.

? Exercise 11.1.1.3

When does the chi-square curve approximate a normal distribution?

Answer

when the number of degrees of freedom is greater than 90

? Exercise 11.1.1.4

Where is μ located on a chi-square curve?

? Exercise 11.1.1.5

Is it more likely the df is 90, 20, or two in the graph?

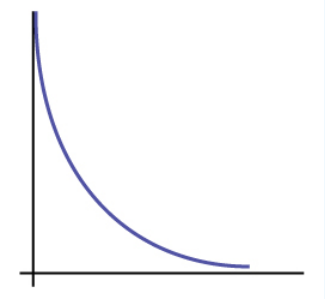


Figure 11.1.1.3.

Answer

$df = 2$

This page titled 11.1.1: Facts About the Chi-Square Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.2: Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data "fit" a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is:

$$\sum_{k} \frac{(O-E)^2}{E} \tag{11.2.1}$$

where:

- $O =$ observed values (data)
- $E =$ expected values (from theory)
- $k =$ the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is $df = (\text{number of categories} - 1)$.

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

> The expected value for each cell needs to be at least five in order for you to use this test.

Example 11.3.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to the table below.

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number **(observed)** of absences in a course. The chart in the table below displays the results of that survey.

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

- H_0: Student absenteeism **fits** faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

- H_a: Student absenteeism **does not fit** faculty perception.

? Exercise 11.2.1.1

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Answer

a. **No.** Notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in the table below.

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

? Exercise 11.2.1.2

b. What is the number of degrees of freedom (df)?

Answer

b. There are four "cells" or categories in each of the new tables.

$df = \text{number of cells} - 1 = 4 - 1 = 3$

? Exercise 11.2.1

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in the table below.

Number produced	Number defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

A random sample was taken to determine the actual number of defects. The table below shows the results of the survey.

Number produced	Number defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Answer

H_0:The number of defects fits expectations.

H_a:The number of defects does not fit expectations.

$df = 4$

Example 11.3.2

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the table below. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

Day of the Week Employees were Most Absent

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Answer

The null and alternative hypotheses are:

- H_0: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: $15 + 12 + 9 + 9 + 15 = 60$), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values $(12, 12, 12, 12, 12)$
- Observed (O) values $(15, 12, 9, 9, 15)$
- $(O–E)$
- $(O–E)^2$
- $\frac{(O–E)^2}{E}$

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get $p\text{-value} = 0.5578$.)

The dfs are the $\text{number of cells} - 1 = 5 - 1 = 4$

Press `2nd DISTR` . Arrow down to χ^2cdf. Press `ENTER` . Enter `(3,10^99,4)` . Rounded to four decimal places, you should see 0.5578, which is the $p\text{-value}$.

Next, complete a graph like the following one with the proper labeling and shading. (You should shade the right tail.)

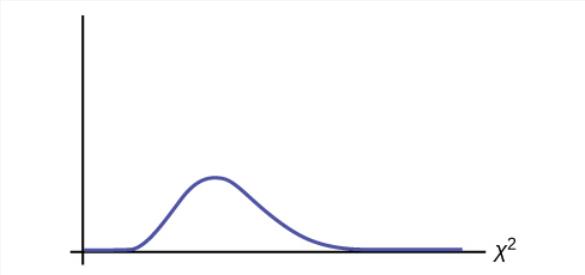


Figure **11.2.1**.

The decision is not to reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example Example has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF` . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF` . Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw` . Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Alternatively, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER` .

? Exercise 11.2.2

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students were asked on which night of the week they did the most homework. The results were distributed as in the table below.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Answer

$df = 6$

p-value $= 0.6093$

We decline to reject the null hypothesis. There is not enough evidence to support that students do not do the majority of their homework equally throughout the week.

Example 11.3.3

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as in the table below.

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the data in the table below.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

Exercise 11.2.3.1

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Answer

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are shown in the table below.

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
over 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter $0.10 * 600$.

H_0: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $df = (\text{the number of cells}) - 1 = 5 - 1 = 4$.

Note 11.3.3.1

$df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:

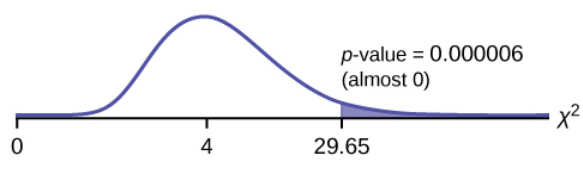


Figure 11.2.2.

Probability statement: $p\text{-value} = P(\chi^2 > 29.65) = 0.000006$

Compare α and the p-value:

$\alpha = 0.01$

$p\text{-value} = 0.000006$

So, $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600, .16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`.

Press `2nd DISTR` . Press `7` or Arrow down to `7:χ2cdf` and press `ENTER` . Enter `(29.65,1E99,4)` . Rounded to four places, you should see `5.77E-6 = .000006` (rounded to six decimal places), which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF` . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF` . Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw` . Make sure you clear any lists before you start.

Exercise 11.2.3

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in the table below.

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in the table below.

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, does it appear that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the p-value?

Answer

p-value $= 0.0036$

We reject the null hypothesis that the distributions are the same. There is sufficient evidence to conclude that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole.

Example 11.3.4

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5% significance level.

Answer

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is HH, HT, TH, TT. Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins ($20HH, 27HT, 30TH, 23TT$) fit the expected distribution?"

Random Variable: Let $X =$ the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the **number of cells is three**. Since $X =$ the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

H_0: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:

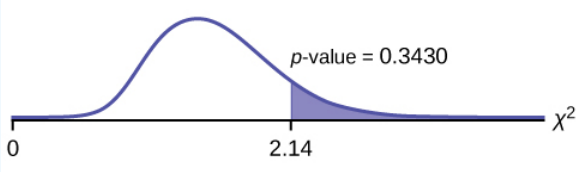


Figure **11.2.3**.

Probability statement: $p\text{-value} = P(\chi^2 > 2.14) = 0.3430$

Compare α and the p-value:

$\alpha = 0.05$

$p\text{-value} = 0.3430$

$\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`. `Enter L3`. Rounded to two decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to four places, you should see `.3430`, which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

? Exercise 11.2.4

Students in a social studies class hypothesize that the literacy rates across the world for every region are 82%. The table below shows the actual literacy rates across the world broken down by region. What are the test statistic and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed Regions	99.0
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91.0
Eastern Asia	93.8
Southern Asia	61.9
South-Eastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Answer

$df = 9$

χ^2 test statistic $= 26.38$

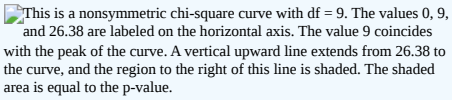


Figure **11.2.4**.

Press `STAT` and `ENTER`. Make sure you clear lists `L1, L2,` and `L3` if they have data in them. Into L1, put the observed frequencies `99, 99.5, 67.3, 62.5, 91, 93.8, 61.9, 91.9, 84.5, 66.4`. Into `L2`, put the expected frequencies `82, 82, 82, 82, 82, 82, 82, 82, 82, 82`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`. `Enter L3`. Rounded to two decimal places, you should see `26.38`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `26.38,1E99,9)`. Rounded to four places, you should see `.0018`, which is the *p*-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

References

1. Data from the U.S. Census Bureau
2. Data from the College Board. Available online at http://www.collegeboard.com.
3. Data from the U.S. Census Bureau, Current Population Reports.
4. Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I.S. Ockene, “Association between Eating Patterns and Obesity in a Free-living US Adult Population.” *American Journal of Epidemiology* volume 158, no. 1, pages 85-92.
5. Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, “Prevalence of Obesity in the United States, 2009–2010.” NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).
6. Stevens, Barbara J., “Multi-family and Commercial Solid Waste and Recycling Survey.” Arlington Count, VA. Available online at www.arlingtonva.us/department.../file84429.pdf (accessed May 24,2013).

Review

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_k \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where:

O: observed values

E: expected value

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

Determine the appropriate test to be used in the next three exercises.

? Exercise 11.2.5

An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

? Exercise 11.2.6

An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

Answer

a goodness-of-fit test

? Exercise 11.2.7

A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises: A teacher predicts that the distribution of grades on the final exam will be and they are recorded in the table below.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 20 is in the table below.

Grade	Frequency
A	7
B	7
C	5
D	1

Exercise 11.2.8

$df =$ _____

Answer

3

Exercise 11.2.9

State the null and alternative hypotheses.

Exercise 11.2.10

χ^2test statistic = _____

Answer

2.04

Exercise 11.2.11

p-value = _____

Exercise 11.2.12

At the 5% significance level, what can you conclude?

Answer

We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in the table below.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is as in the table below.

Ethnicity	Percentage of total county population	Number expected (round to two decimal places)
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Exercise 11.2.13

If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.

Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.2.14

H_0: _____

Answer

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.2.15

H_a: _____

Exercise 11.2.16

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

right-tailed

Exercise 11.2.17

degrees of freedom = _____

Exercise 11.2.18

χ^2test statistic = _____

Answer

88,621

Exercise 11.2.19

p-value = _____

Exercise 11.2.20

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



Figure 11.2.5.

Let $\alpha = 0.05$

Decision: _____

Reason for the Decision: _____

Conclusion (write out in complete sentences): _____

Answer

Graph: Check student's solution.

Decision: Reject the null hypothesis.

Reason for the Decision: $p\text{-value} < \alpha$

Conclusion (write out in complete sentences): The make-up of AIDS cases does not fit the ethnicities of the general population of Santa Clara County.

Exercise 11.2.21

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

This page titled 11.2: Goodness-of-Fit Test is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **11.3: Goodness-of-Fit Test** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

11.3: Tests Using Contingency tables

11.3: Tests Using Contingency tables is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.3.1: Test of Independence

Tests of independence involve using a contingency table of observed (data) values.

The test statistic for a *test of independence* is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{11.3.1.1}$$

where:

- $O =$ observed values
- $E =$ expected values
- $i =$ the number of rows in the table
- $j =$ the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

The expected value for each cell needs to be at least five in order for you to use this test.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Probability Topics. As a review, consider the following example.

Example 11.3.1.1

Suppose $A =$ a speeding violation in the last year and $B =$ a cell phone user while driving. If A and B are independent then $P(A \text{ AND } B) = P(A)P(B)$. $A \text{ AND } B$ is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell phone while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let $y =$ expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right)$$

Solve for y:

$$y = \frac{(70)(305)}{755} = 28.3$$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternative hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example, then the null hypothesis is:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

$$df = (\text{number of columns} - 1)(\text{number of rows} - 1)$$

The following formula calculates the **expected number** (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

Exercise 11.3.1.1

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Answer

About 16 students are expected to be music students and on the honor roll.

Example 11.3.1.2

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In Table **11.3.1.1** is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table **11.3.1.1**: Number of Hours Worked Per Week by Volunteer Type (Observed). The table contains **observed (O)** values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Is the number of hours volunteered **independent** of the type of volunteer?

Answer

The **observed table** and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

- H_0: The number of hours volunteered is **independent** of the type of volunteer.
- H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected results are in Table **11.3.1.2**.

Table **11.3.1.2**: Number of Hours Worked Per Week by Volunteer Type (Expected). The table contains **expected**(E) values (data).

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

For example, the calculation for the expected frequency for the top left cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$$df = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$$

Graph:

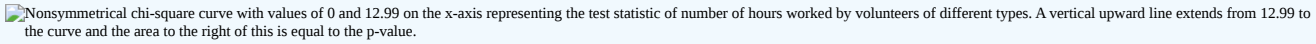


Figure **11.3.1.1**.

Probability statement: $p\text{-value} = P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0113$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the example in Table, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Table. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. If necessary, use the arrow keys to move the cursor after `Observed:` and press `2nd MATRX`. Press `1:[A]` to select matrix A. It is not necessary to enter expected values. The matrix listed after `Expected:` can be blank. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the p-value = 0.0113. Do the procedure a second time, but arrow down to `Draw` instead of `calculate`.

Exercise 11.3.1.2

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table **11.3.1.3** shows the results:

Table **11.3.1.3**

Industry Sector	2000	2010	2020	Total
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Answer

- H_0: The number of jobs is independent of the year.
- H_a: The number of jobs is dependent on the year.

$df = 12$



Figure **11.3.1.2**.

Press the `MATRX` key and arrow over to `EDIT` . Press `1:[A]` . Press `3 ENTER 3 ENTER` . Enter the table values by row. Press `ENTER` after each. Press `2nd QUIT` . Press `STAT` and arrow over to `TESTS` . Arrow down to c:χ^2-TEST. Press `ENTER` . You should see `Observed:[A] and Expected:[B]` . Arrow down to `Calculate` . Press `ENTER` . The test statistic is 227.73 and the p-value $= 5.90E-42 = 0$. Do the procedure a second time but arrow down to `Draw` instead of `calculate` .

Example 11.3.1.3

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. Table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

a. How many high anxiety level students are expected to have a high need to succeed in school?
b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?
c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} =$ _____
d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about _____.

Solution

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09 \quad (11.3.1.2)$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

d. 8

? Exercise 11.3.1.3

Refer back to the information in Note. How many service providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

Answer

12,727, 14,965

References

1. DiCamilo, Mark, Mervin Field, “Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs.” The Field Poll, released Feb. 14, 2013. Available online at field.com/fieldpollonline/sub...rs/Rls2436.pdf (accessed May 24, 2013).
2. Harris Interactive, “Favorite Flavor of Ice Cream.” Available online at http://www.statisticbrain.com/favori...r-of-ice-cream (accessed May 24, 2013)
3. “Youngest Online Entrepreneurs List.” Available online at http://www.statisticbrain.com/younge...repreneur-list (accessed May 24, 2013).

Review

To assess whether two factors are independent or not, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values.

The test is right-tailed. Each observation or cell category must have an expected value of at least 5.

Formula Review

Test of Independence

- The number of degrees of freedom is equal to $(\text{number of columns - 1})(\text{number of rows - 1})$.
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where $O =$ observed values, $E =$ expected values, $i =$ the number of rows in the table, and $j =$ the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Determine the appropriate test to be used in the next three exercises.

? Exercise 11.3.1.4

A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

Answer

a test of independence

? Exercise 11.3.1.5

The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

? Exercise 11.3.1.6

A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.

Answer

a test of independence

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. Table **11.3.1.4** shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance they must travel.

Table **11.3.1.4**

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

? Exercise 11.3.1.7

State the hypotheses.

- H_0: _____

- H_a: _____

Exercise 11.3.1.8

$df =$ _____

Answer

8

Exercise 11.3.1.9

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

Exercise 11.3.1.10

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

Answer

6.6

Exercise 11.3.1.11

What is the test statistic?

Exercise 11.3.1.12

What is the p-value?

Answer

0.0435

Exercise 11.3.1.13

What can you conclude at the 5% level of significance?

Use the following information to answer the next eight exercises: An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650 whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

Exercise 11.3.1.14

Complete the table.

Table **11.3.1.5**: Smoking Levels by Ethnicity (Observed)

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Answer

Table **11.3.1.5*B***

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1-10	9,886	2,745	12,831	8,378	7,650	41,490
11-20	6,514	3,062	4,932	10,680	9,877	35,065
21-30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

? Exercise 11.3.1.15

State the hypotheses.

- H_0: _____
- H_a: _____

? Exercise 11.3.1.16

Enter expected values in Table. Round to two decimal places.

Calculate the following values:

Answer

Table **11.3.1.6**

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7777.57	3310.11	8248.02	10771.29	11383.01
11-20	6573.16	2797.52	6970.76	9103.29	9620.27
21-30	2863.02	1218.49	3036.20	3965.05	4190.23
31+	1616.25	687.87	1714.01	2238.37	2365.49

Exercise 11.3.1.17

$df =$ _____

Exercise 11.3.1.18

χ^2test statistic = _____

Answer

10,301.8

Exercise 11.3.1.19

p-value = _____

Exercise 11.3.1.20

Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Answer

right

Exercise 11.3.1.21

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



Figure **11.3.1.3**.

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

Exercise 11.3.1.22

$\alpha = 0.05$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

Answer

a. Reject the null hypothesis.
b. p-value $< \alpha$
c. There is sufficient evidence to conclude that smoking level is dependent on ethnic group.

Exercise 11.3.1.23

$\alpha = 0.05$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

Glossary

Contingency Table

a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

This page titled 11.3.1: Test of Independence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **11.4: Test of Independence** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

11.3.2: Test for Homogeneity

The goodness–of–fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the test for homogeneity, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

> The expected value for each cell needs to be at least five in order for you to use this test.

Hypotheses

- H_0: The distributions of the two populations are the same.
- H_a: The distributions of the two populations are not the same.

Test Statistic

- Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of Freedom (df)

- $df = \text{number of columns} - 1$

Requirements

- All values in the table must be greater than or equal to five.

Common Uses

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Example 11.3.2.1

Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other. The results are shown in Table **11.3.2.1**. Do male and female college students have the same distribution of living arrangements?

Table **11.3.2.1**: Distribution of Living Arragements for College Males and College Females

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Answer

- H_0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.
- H_a: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of Freedom (df):

$df = \text{number of columns} - 1 = 4 - 1 = 3$

Distribution for the test: χ^2_3

Calculate the test statistic: $\chi^2 = 10.1287$ (calculator or computer)

Probability statement: $p\text{-value} = P(\chi^2 > 10.1287) = 0.0175$

Press the

```
MATRX
```

key and arrow over to

```
EDIT
```

. Press

```
1:[A]
```

. Press

```
2 ENTER 4 ENTER
```

. Enter the table values by row. Press

```
ENTER
```

after each. Press

```
2nd QUIT
```

. Press

```
STAT
```

and arrow over to

```
TESTS
```

. Arrow down to

```
C:χ2-TEST
```

. Press

```
ENTER
```

. You should see

```
Observed:[A] and Expected:[B]
```

. Arrow down to

```
Calculate
```

. Press

```
ENTER
```

. The test statistic is 10.1287 and the p-value $= 0.0175$. Do the procedure a second time but arrow down to

```
Draw
```

instead of

```
calculate
```

.

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value $= 0.0175$. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0. This means that the distributions are not the same.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Exercise 11.3.2.1

Do families and singles have the same distribution of cars? Use a level of significance of 0.05. Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in Table **11.3.2.2**. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

Table **11.3.2.1**

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Answer

With a p-value of almost zero, we reject the null hypothesis. The data show that the distribution of cars is not the same for families and singles.

Example 11.5.2

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. Table shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Answer

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of Freedom (df):

$df = \text{number of columns} - 1 = 3 - 1 = 2$

Distribution for the test: χ^2_2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator or computer)

Probability statement: $p\text{-value} = P(\chi^2 > 3.2603) = 0.1959$

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 3.2603 and the p-value = 0.1959. Do the procedure a second time but arrow down to Draw instead of calculate.

Compare α and the p-value: $\alpha = 0.05$ and the $p\text{-value} = 0.1959$. $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At a 5% level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

Exercise 11.3.2.2

Ivy League schools receive many applications, but only some can be accepted. At the schools listed in Table, two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the p-value, and draw a conclusion about the test of homogeneity.

Answer

H_0: The distribution of regular applications accepted is the same as the distribution of early applications accepted.

H_a: The distribution of regular applications accepted is not the same as the distribution of early applications accepted.

$df = 5$

$\chi^2\text{test statistic} = 430.06$

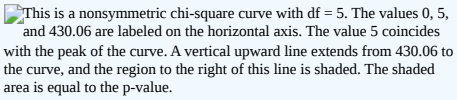


Figure 11.3.2.1.

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C :χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 430.06 and the $p\text{-value} = 9.80E-91$. Do the procedure a second time but arrow down to Draw instead of calculate.

References

1. Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).
2. “Energy use (kg of oil equivalent per capita).” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...G.OE/countries (accessed May 24, 2013).
3. “Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/pubsinf...?pubid=2009030 (accessed May 24, 2013).
4. “Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

Review

To assess whether two data sets are derived from the same distribution—which need not be known, you can apply the test for homogeneity that uses the chi-square distribution. The null hypothesis for this test states that the populations of the two data sets come from the same distribution. The test compares the observed values against the expected values if the two populations followed the same distribution. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_{i \cdot j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where: O = observed values

E = expected values

i = number of rows in data contingency table

j = number of columns in data contingency table

$df = (i-1)(j-1)$ Degrees of freedom

? Exercise 11.3.2.3

A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

Answer

test for homogeneity

? Exercise 11.3.2.4

What are the null and alternative hypotheses for Exercise?

? Exercise 11.3.2.5

A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

Answer

test for homogeneity

? Exercise 11.3.2.6

A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

Exercise 11.3.2.7

What condition must be met to use the test for homogeneity?

Answer

All values in the table must be greater than or equal to five.

Use the following information to answer the next five exercises: Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in Table.

	20–30	30–40	40–50	50–60
Private Practice	16	40	38	6
Hospital	8	44	59	39

Exercise 11.3.2.8

State the null and alternative hypotheses.

Exercise 11.3.2.9

$df =$ _____

Answer

3

Exercise 11.3.2.10

What is the test statistic?

Exercise 11.3.2.11

What is the p-value?

Answer

0.00005

Exercise 11.3.2.12

What can you conclude at the 5% significance level?

This page titled 11.3.2: Test for Homogeneity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.3.3: Comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:
  - H_0: The population fits the given distribution.
  - H_a: The population does not fit the given distribution.
- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:
  - H_0: The two variables (factors) are independent.
  - H_a: The two variables (factors) are dependent.
- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:
  - H_0: The two populations follow the same distribution.
  - H_a: The two populations have different distributions.

Review

The goodness-of-fit test is typically used to determine if data fits a particular distribution. The test of independence makes use of a contingency table to determine the independence of two factors. The test for homogeneity determines whether two populations come from the same distribution, even if this distribution is unknown.

? Exercise 11.3.3.1

Which test do you use to decide whether an observed distribution is the same as an expected distribution?

Answer

a goodness-of-fit test

? Exercise 11.3.3.2

What is the null hypothesis for the type of test from Exercise?

? Exercise 11.3.3.3

Which test would you use to decide whether two factors have a relationship?

Answer

a test for independence

Exercise 11.3.3.4

Which test would you use to decide if two populations have the same distribution?

Exercise 11.3.3.5

How are tests of independence similar to tests for homogeneity?

Answer

Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{i \cdot j} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

Exercise 11.3.3.6

How are tests of independence different from tests for homogeneity?

Bringing It Together

Exercise 11.3.3.7

a. Explain why a goodness-of-fit test and a test of independence are generally right-tailed tests.
b. If you did a left-tailed test, what would you be testing?

Answer a

The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.

Answer b

Testing to see if the data fits the distribution “too well” or is too perfect.

This page titled 11.3.3: Comparison of the Chi-Square Tests is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.4: Prelude to F Distribution and One-Way ANOVA

CHAPTER OBJECTIVES

By the end of this chapter, the student should be able to:

- Interpret the F probability distribution as the number of groups and the sample size change.
- Discuss two uses for the F distribution: one-way ANOVA and the test of two variances.
- Conduct and interpret one-way ANOVA.
- Conduct and interpret hypothesis tests of two variances

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.



Figure **11.4.1**: One-way ANOVA is used to measure information from several groups.

For hypothesis tests comparing averages between more than two groups, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the ***F*** distribution, used for one-way ANOVA, and the test of two variances. This is just a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.4: Prelude to F Distribution and One-Way ANOVA is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.4.1: One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses variances to help determine if the means are equal or not. To perform a one-way ANOVA test, there are several basic assumptions to be fulfilled:

Five basic assumptions of one-way ANOVA to be fulfilled

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

- $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$
- H_a : At least two of the group means$\mu_2 = \mu_3 = ... = \mu_k$are not equal

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H_0 : \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

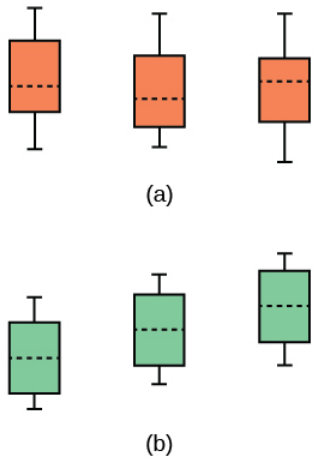


Figure 11.4.1.1: (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

Review

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with two different degrees of freedom.

Assumptions:

a. Each population from which a sample is taken is assumed to be normal.
b. All samples are randomly selected and independent.
c. The populations are assumed to have equal standard deviations (or variances).

Glossary

Analysis of Variance

also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

One-WayANOVA

a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

Variance

mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.4.1: One-Way ANOVA is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.4.2: The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the F distribution, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

The F distribution is derived from the Student's t-distribution. The values of the F distribution are squares of the corresponding values of the t-distribution. One-Way ANOVA expands the t-test for comparing more than two groups. The scope of that derivation is beyond the level of this course.

To calculate the F ratio, two estimates of the variance are made.

a. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
b. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- SS_{between} = the sum of squares that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in discussed previously.

MS means "mean square." MS_{between} is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the jth group}
- s_j = the sum of the values in the jth group
- n = total number of all the values combined (total sample size):

$$n = \sum n_j \tag{11.4.2.1}$$

- x = one value:

$$\sum x = \sum s_j \tag{11.4.2.2}$$

- Sum of squares of all values from every group combined:

$$\sum x^2 \tag{11.4.2.3}$$

- Between group variability:

$$SS_{\text{total}} = \sum x^2 - \frac{(\sum x^2)}{n} \tag{11.4.2.4}$$

- Total sum of squares:

$$\sum x^2 - \frac{(\sum x)^2}{n} \tag{11.4.2.5}$$

- Explained variation: sum of squares representing variation among the different samples:

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n} \quad (11.4.2.6)$$

- Unexplained variation: sum of squares representing variation within samples due to chance:

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}} \quad (11.4.2.7)$$

- df's for different groups (df's for the numerator):

$$df = k - 1 \quad (11.4.2.8)$$

- Equation for errors within samples (df's for the denominator):

$$df_{\text{within}} = n - k \quad (11.4.2.9)$$

- Mean square (variance estimate) explained by the different groups:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} \quad (11.4.2.10)$$

- Mean square (variance estimate) that is due to chance (unexplained):

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} \quad (11.4.2.11)$$

MS_{between} and MS_{within} can be written as follows:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1} \quad (11.4.2.12)$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{SS_{\text{within}}}{n-k} \quad (11.4.2.13)$$

The one-way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \quad (11.4.2.14)$$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}.Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2} \tag{11.4.2.15}$$

where ...

- $n =$ the sample size
- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- $s_{\text{pooled}}^2 =$ the mean of the sample variances (pooled variance)
- $s_{\bar{x}^2} =$ the variance of the sample means

Data are typically put into a table for easy viewing. One-Way ANOVA results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	$SS(\text{Factor})$	$k-1$	$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)}$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})}$
Error (Within)	$SS(\text{Error})$	$n-k$	$MS(\text{Error}) = \frac{SS(\text{Error})}{(n-k)}$	
Total	$SS(\text{Total})$	$n-1$		

Example 11.4.2.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in Table.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

$$s_1 = 16.5, s_2 = 15, s_3 = 15.7 \tag{11.4.2.16}$$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(\text{between}) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n} \tag{11.4.2.17}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} + \frac{(s_1 + s_2 + s_3)^2}{10} \tag{11.4.2.18}$$

where $n_1 = 4, n_2 = 3, n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$ so

$$SS(\text{between}) = \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} = \frac{(16.5 + 15 + 15.5)^2}{10} \tag{11.4.2.19}$$

$$= 2.2458 \tag{11.4.2.20}$$

$$S(\text{total}) = \sum x^2 - \frac{(\sum x)^2}{n} \tag{11.4.2.21}$$
$$= (5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2) \tag{11.4.2.22}$$
$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10} \tag{11.4.2.23}$$
$$= 244 - \frac{47^2}{10} = 244 - 220.9 \tag{11.4.2.24}$$
$$= 23.1 \tag{11.4.2.25}$$

$$SS(\text{within}) = SS(\text{total}) - SS(\text{between}) \tag{11.4.2.26}$$
$$= 23.1 - 2.2458 \tag{11.4.2.27}$$
$$= 20.8542 \tag{11.4.2.28}$$

One-Way ANOVA Table: The formulas for $SS(\text{Total})$, $SS(\text{Factor}) = SS(\text{Between})$ and $SS(\text{Error}) = SS(\text{Within})$ as shown previously. The same information is provided by the TI calculator hypothesis test function ANOVA in STAT TESTS (syntax is $ANOVA(L1, L2, L3)$ where $L1, L2, L3$ have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	$SS(\text{Factor}) = SS(\text{Between}) = 2.2458$	$k - 1 = 3$ groups $- 1 = 2$	$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)} = \frac{2.2458}{2} = 1.1229$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})} = \frac{1.1229}{2.9792} = 0.3769$
Error (Within)	$SS(\text{Error}) = SS(\text{Within}) = 20.8542$	$n–k = 10$ total data $- 3$ groups $= 7$	$MS(\text{Error})) = \frac{SS(\text{Error})}{(n–k)} = \frac{20.8542}{7} = 2.9792$	
Total	$SS(\text{Total}) = 2.2458 + 20.8542 = 23.1$	$n - 1 = 10$ total data $- 1 = 9$		

Exercise 11.4.2.1

As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments

- bare soil
- a commercial ground cover
- black plastic
- straw
- compost

All plants grew under the same conditions and were the same variety. Students recorded the weight (in grams) of tomatoes produced by each of the $n = 15$ plants:

Bare: $n_1 = 3$	Ground Cover: $n_2 = 3$	Plastic: $n_3 = 3$	Straw: $n_4 = 3$	Compost: $n_5 = 3$
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Create the one-way ANOVA table.

Answer

Enter the data into lists L1, L2, L3, L4 and L5. Press STAT and arrow over to TESTS. Arrow down to ANOVA. Press ENTER and enter L1, L2, L3, L4, L5). Press ENTER. The table was filled in with the results from the calculator.

One-Way ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	36,648,561	$5-1=4$	$\frac{36,648,561}{4}=9,162,140$	$\frac{9,162,140}{2,044,672.6}=4.4810$
Error (Within)	20,446,726	$15-5=10$	$\frac{20,446,726}{10}=2,044,672.6$	
Total	57,095,287	$15-1=14$		

The one-way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim Fdf(\text{num}), df(\text{denom})$

where $df(\text{num}) = df_{between} and df(\text{denom}) = df_{within}$

The mean for the F distribution is $\mu = \frac{df(\text{num})}{df(\text{denom})-1}$

References

1. Tomato Data, Marist College School of Science (unpublished student research)

Review

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group to the variation of the mean of each group. The ratio of these two is the F statistic from an F distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

Formula Review

$$SS_{between} = \sum\left[\frac{(s_j)^2}{n_j}\right] - \frac{(\sum s_j)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$$df_{\text{between}} = df(\text{num}) = k-1$$

$$df_{\text{within}} = df(\text{denom}) = n-k$$

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

F ratio when the groups are the same size: $F = \frac{ns_{\bar{x}}^2}{s_{\text{pooled}}^2}$

Mean of the F distribution: $\mu = \frac{df(\text{num})}{df(\text{denom})-1}$

where:

- $k =$ the number of groups
- $n_j =$ the size of the jth group
- $s_j =$ the sum of the values in the jth group
- $n =$ the total number of all values (observations) combined
- $x =$ one value (one observation) from the data
- $s_x^2 =$ the variance of the sample means
- $s_{\text{pooled}}^2 =$ the mean of the sample variances (pooled variance)

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.4.2: The F Distribution and the F-Ratio is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **13.3: The F Distribution and the F-Ratio** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

11.4.3: Facts About the F Distribution

Here are some facts about the F distribution:

a. The curve is not symmetrical but skewed to the right.
b. There is a different curve for each set of dfs.
c. The F statistic is greater than or equal to zero.
d. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
e. Other uses for the F distribution include comparing two variances and two-way Analysis of Variance. Two-Way Analysis is beyond the scope of this chapter.

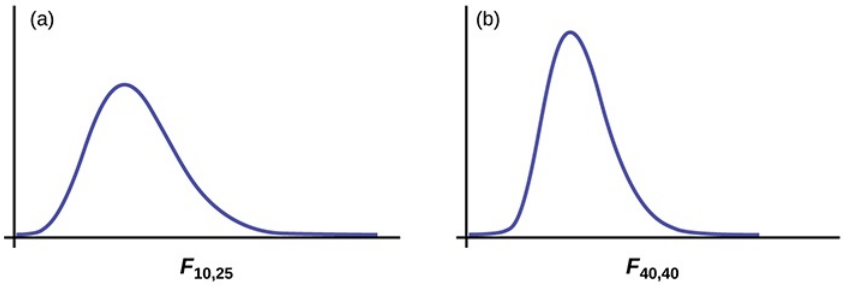


Figure 11.4.3.1

Example 11.4.3.1

Let's return to the slicing tomato exercise. The means of the tomato yields under the five mulching conditions are represented by $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5%, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Answer

The null and alternative hypotheses are:

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- $H_a : \mu_i \neq \mu_j$ some $i \neq j$

The one-way ANOVA results are shown in Table

one-way ANOVA results

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	36,648,561	$5 - 1 = 4$	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	$15 - 5 = 10$	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	$15 - 1 = 14$		

Distribution for the test: $F_{4,10}$

$$df(\text{num}) = 5 - 1 = 4 \quad (11.4.3.1)$$

$$df(\text{denom}) = 15 - 5 = 10 \quad (11.4.3.2)$$

Test statistic: $F = 4.4810$

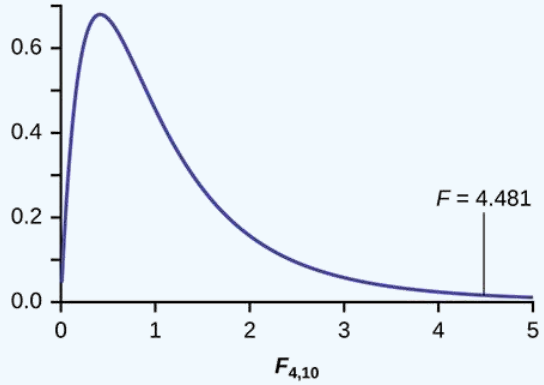


Figure 11.4.3.2

Probability Statement: $p\text{-value} = P(F > 4.481) = 0.0248$.

Compare α and the p-value: $\alpha = 0.05, p\text{-value} = 0.0248$

Make a decision: Since $\alpha > p\text{-value}$, we reject H_0.

Conclusion: At the 5% significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of mulches led to different mean yields.

To find these results on the calculator:

Press STAT. Press 1:EDIT. Put the data into the lists L_1, L_2, L_3, L_4, L_5.

Press STAT, and arrow over to TESTS, and arrow down to ANOVA. Press ENTER, and then enter L_1, L_2, L_3, L_4, L_5). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the p-value of the test.

The calculator displays:

- $F = 4.4810$
- $p = 0.0248$ (p-value)

Factor

- $df = 4$
- $SS = 36648560.9$
- $MS = 9162140.23$

Error

- $df = 10$
- $SS = 20446726$
- $MS = 2044672.6$

Exercise 11.4.3.1

MRSA, or *Staphylococcus aureus*, can cause a serious bacterial infections in hospital patients. Table shows various colony counts from different patients who may or may not have MRSA.

Conc = 0.6	Conc = 0.8	Conc = 1.0	Conc = 1.2	Conc = 1.4
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Plot of the data for the different concentrations:

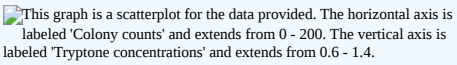


Figure 11.4.3.3

Test whether the mean number of colonies are the same or are different. Construct the ANOVA table (by hand or by using a TI-83, 83+, or 84+ calculator), find the *p*-value, and state your conclusion. Use a 5% significance level.

Answer

While there are differences in the spreads between the groups (Figure 11.4.3.1), the differences do not appear to be big enough to cause concern.

We test for the equality of mean number of colonies:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H_a : \mu_i \neq \mu_j$ some $i \neq j$

The one-way ANOVA table results are shown in Table.

Table 11.4.3.1

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	10,233	$5 - 1 = 4$	$\frac{10,233}{4} = 2,558.25$	$\frac{2,558.25}{4,194.9} = 0.6099$
Error (Within)	41,949	$15 - 5 = 10$		
Total	52,182	$15 - 1 = 14$	$\frac{41,949}{10} = 4,194.9$	

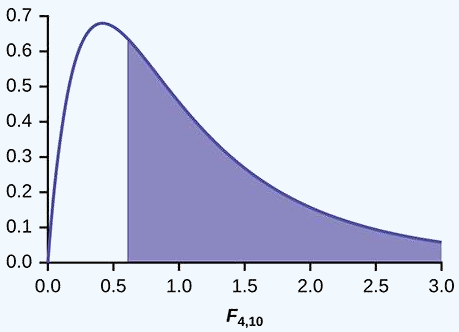


***Figure* 11.4.3.2**

Distribution for the test: $F_{4,10}$

Probability Statement: $p\text{-value} = P(F > 0.6099) = 0.6649$.

Compare α and the p-value: $\alpha = 0.05, p\text{-value} = 0.669, \alpha > p\text{-value}$

Make a decision: Since $\alpha > p\text{-value}$, we do not reject H_0.

Conclusion: At the 5% significance level, there is insufficient evidence from these data that different levels of tryptone will cause a significant difference in the mean number of bacterial colonies formed.

Example 11.4.3.2

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in Table.

Figure 11.4.3.1: MEAN GRADES FOR FOUR SORORITIES

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Using a significance level of 1%, is there a difference in mean grades among the sororities?

Answer

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

This is an example of a balanced design, because each factor (i.e., sorority) has the same number of observations.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means $\mu_1, \mu_2, \mu_3, \mu_4$ are equal.

Distribution for the test: $F_{3,16}$

where $k = 4$ groups and $n = 20$ samples in total

$df(\text{num}) = k - 1 = 4 - 1 = 3$

$df(\text{denom}) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph:

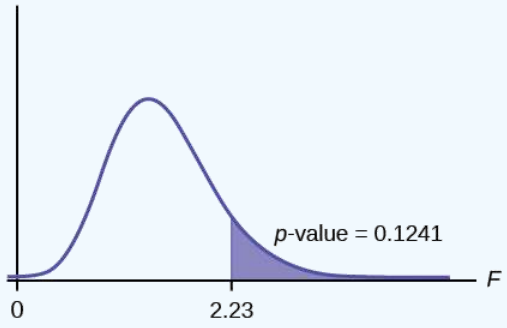


Figure **11.4.3.5**

Probability statement: $p\text{-value} = P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$

$p\text{-value} = 0.1241$

$\alpha < p\text{-value}$

Make a decision: Since $\alpha < p\text{-value}$, you cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Put the data into lists L_1, L_2, L_3, and L_4. Press `STAT` and arrow over to `TESTS`. Arrow down to `F:ANOVA`. Press `ENTER` and Enter (`L1,L2,L3,L4`).

The calculator displays the F statistic, the p-value and the values for the one-way ANOVA table:

$F = 2.2303$

$p = 0.1241$ (p-value)

Factor

$df = 3$

$SS = 2.88732$

$MS = 0.96244$

Error

$df = 1$

$SS = 6.9044$

$MS = 0.431525$

Exercise 11.4.3.2

Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown in Table.

GPAs FOR FOUR SPORTS TEAMS

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Use a significance level of 5%, and determine if there is a difference in GPA among the teams.

Answer

With a p-value of 0.9271, we decline to reject the null hypothesis. There is not sufficient evidence to conclude that there is a difference among the GPAs for the sports teams.

Example 11.4.3.3

A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in Table 11.4.3.3.

Table 11.4.3.3

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 3% level of significance.

Answer

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants, so we will use the formula

$$F' = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2}. \tag{11.4.3.3}$$

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Next, calculate the variance of the three group means (Calculate the variance of 24.2, 25.4, and 24.4). **Variance of the group means** $= 0.413 = s_{\bar{x}}^2$

Then $MS_{\text{between}} = ns_{\bar{x}}^2 = (5)(0.413)$ where $n = 5$ is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (Calculate the mean of 11.7, 18.3, and 16.3). **Mean of the sample variances** $= 15.433 = s_{\text{pooled}}^2$

Then $MS_{\text{within}} = s_{\text{pooled}}^2 = 15.433$.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{ns_{\bar{x}}^2}{s_{\text{pooled}}^2} = \frac{(5)(0.413)}{15.433} = 0.134$

The dfs for the numerator $=$ the number of groups $- 1 = 3 - 1 = 2$.

The dfs for the denominator $=$ the total number of samples $-$ the number of groups $= 15 - 3 = 12$

The distribution for the test is $F_{2,12}$ and the F statistic is $F = 0.134$

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value $= 0.8759$, do not reject H_0. (Why?)

Conclusion: With a 3% level of significance, from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

To calculate the p-value:

*Press `2nd DISTR`

*Arrow down to `Fcdf` (and press `ENTER` .

*Enter 0.134, `E99` , 2, 12)

*Press ENTER

The p-value is 0.8759.

? Exercise 11.4.3.3

Another fourth grader also grew bean plants, but this time in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32. Do a one-way ANOVA test on the four groups. Are the heights of the bean plants different? Use the same method as shown in Example 11.4.3.3.

Answer

- $F = 0.9496$
- p-value $= 0.4402$

From the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Collaborative Exercise

From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1% level of significance. Use one of the solution sheets in [link].

References

1. Data from a fourth grade classroom in 1994 in a private K – 12 school in San Jose, CA.
2. Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets: Data for Fruitfly Fecundity.* London: Chapman & Hall, 1994.
3. Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets.*London: Chapman & Hall, 1994, pg. 50.
4. Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. A Handbook of Small Datasets. London: Chapman & Hall, 1994, pg. 118.
5. “MLB Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012.
6. Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

Review

The graph of the F distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The F statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small F statistic will result, and the area under the F curve to the right will be large, representing a large p-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large F statistic and a small area (small p-value) to the right of the statistic under the F curve.

When the data have unequal group sizes (unbalanced data), then techniques discussed earlier need to be used for hand calculations. In the case of balanced data (the groups are the same size) however, simplified calculations based on group means and variances may be used. In practice, of course, software is usually employed in the analysis. As in any analysis, graphs of various sorts should be used in conjunction with numerical techniques. Always look of your data!

This page titled 11.4.3: Facts About the F Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **13.4: Facts About the F Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

11.4.4: How to Use Microsoft Excel® for Regression Analysis

13.11: How to Use Microsoft Excel for Regression Analysis

11.4.4: How to Use Microsoft Excel® for Regression Analysis is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.E: F Distribution and One-Way ANOVA (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

13.1: Introduction

13.2: One-Way ANOVA

Q 13.2.1

Three different traffic routes are tested for mean driving time. The entries in the table are the driving times in minutes on the three different routes. The one-way $ANOVA$ results are shown in Table.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

State SS_{between}, SS_{within}, and the F statistic.

S 13.2.1

$SS_{\text{between}} = 26$

$SS_{\text{within}} = 441$

$F = 0.2653$

Q 13.2.2

Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

State the hypotheses.

H_0: _____

H_a: _____

13.3: The F-Distribution and the F-Ratio

Use the following information to answer the next five exercises. There are five basic assumptions that must be fulfilled in order to perform a one-way $ANOVA$ test. What are they?

Exercise 13.2.1

Write one assumption.

Answer

Each population from which a sample is taken is assumed to be normal.

Exercise 13.2.2

Write another assumption.

Exercise 13.2.3

Write a third assumption.

Answer

The populations are assumed to have equal standard deviations (or variances).

Exercise 13.2.4

Write a fourth assumption.

Exercise 13.2.5

Write the final assumption.

Answer

The response is a numerical value.

Exercise 13.2.6

State the null hypothesis for a one-way $ANOVA$ test if there are four groups.

Exercise 13.2.7

State the alternative hypothesis for a one-way $ANOVA$ test if there are three groups.

Answer

H_a : At least two of the group means μ_1, μ_2, μ_3 are not equal.

Exercise 13.2.8

When do you use an $ANOVA$ test?

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
s^2	_____	_____	_____	_____	_____

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: At least any two of the group means $\mu_1, \mu_2, \ldots, \mu_5$ are not equal.

Q 13.3.1

degrees of freedom – numerator: $df(\text{num}) =$ _____

Q 13.3.2

degrees of freedom – denominator: $df(\text{denom}) =$ _____

S 13.3.2

$df(\text{denom}) = 15$

Q 13.3.3

F statistic = _____

13.4: Facts About the F Distribution

Exercise 13.4.4

An F statistic can have what values?

Exercise 13.4.5

What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Answer

The curves approximate the normal distribution.

Use the following information to answer the next seven exercise. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in Table.

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Exercise 13.4.6

What is the $df(\text{num})$?

Exercise 13.4.7

What is the $df(\text{denom})$?

Answer

ten

Exercise 13.4.8

What are the Sum of Squares and Mean Squares Factors?

Exercise 13.4.9

What are the Sum of Squares and Mean Squares Errors?

Answer

$SS = 237.33; MS = 23.73$

Exercise 13.4.10

What is the F statistic?

Exercise 13.4.11

What is the p-value?

Answer

0.1614

Exercise 13.4.12

At the 5% significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next seven exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in Table

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

Exercise 13.4.13

What is the $df(\text{num})$?

Answer

two

Exercise 13.4.14

What is the $df(\text{denom})$?

Exercise 13.4.15

What are the SS_{between} and MS_{between}?

Answer

$SS_{\text{between}} = 5,700.4$;

$MS_{\text{between}} = 2,850.2$

Exercise 13.4.16

What are the SS_{within} and MS_{within}?

Exercise 13.4.17

What is the F Statistic?

Answer

3.6101

Exercise 13.4.18

What is the p-value?

Exercise 13.4.19

At the 10% significance level, are the scores among the different groups different?

Answer

Yes, there is enough evidence to show that the scores among the groups are statistically significant at the 10% level.

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Enter the data into your calculator or computer.

Exercise 13.4.20

p-value = _____

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

Exercise 13.4.21

$\alpha = 0.05$

a. Decision: _____
b. Conclusion: _____

Exercise 13.4.22

$\alpha = 0.01$

a. Decision: _____
b. Conclusion: _____

Use the following information to answer the next eight exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in the table are the weights for the different groups. The one-way $ANOVA$ results are shown in Table.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Exercise 13.3.2

What is the Sum of Squares Factor?

Answer

4,939.2

Exercise 13.3.3

What is the Sum of Squares Error?

Exercise 13.3.4

What is the df for the numerator?

Answer

2

Exercise 13.3.5

What is the df for the denominator?

Exercise 13.3.6

What is the Mean Square Factor?

Answer

2,469.6

Exercise 13.3.7

What is the Mean Square Error?

Exercise 13.3.8

What is the F statistic?

Answer

3.7416

Use the following information to answer the next eight exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in the table are the goals per game for the different teams. The one-way $ANOVA$ results are shown in Table.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

Exercise 13.3.9

What is SS_{between}?

Exercise 13.3.10

What is the df for the numerator?

Answer

3

Exercise 13.3.11

What is MS_{between}?

Exercise 13.3.12

What is SS_{within}?

Answer

13.2

Exercise 13.3.13

What is the df for the denominator?

Exercise 13.3.14

What is MS_{within}?

Answer

0.825

Exercise 13.3.15

What is the F statistic?

Exercise 13.3.16

Judging by the F statistic, do you think it is likely or unlikely that you will reject the null hypothesis?

Answer

Because a one-way $ANOVA$ test is always right-tailed, a high F statistic corresponds to a low p-value, so it is likely that we will reject the null hypothesis.

DIRECTIONS

Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in [link].

Q 13.4.1

Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10%, test the hypothesis that the three formulas produce the same mean weight gain.

Weights of Student Lab Rats

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

a. $H_0 : \mu_L = \mu_T = \mu_J$
b. at least any two of the means are different
c. $df(\text{num}) = 2; df(\text{denom}) = 12$
d. F distribution
e. 0.67
f. 0.5305
g. Check student's solution.
h. Decision: Do not reject null hypothesis; Conclusion: There is insufficient evidence to conclude that the means are different.

Q 13.4.2

A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in Table. Using a 5% significance level, test the hypothesis that the three mean commuting mileages are the same.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Q 13.4.3

Examine the seven practice laps from [link]. Determine whether the mean lap time is statistically the same for the seven practice laps, or if there is at least one lap that has a different mean time from the others.

S 13.4.3

a. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_T$
b. At least two mean lap times are different.
c. $df(\text{num}) = 6; df(\text{denom}) = 98$
d. F distribution
e. 1.69
f. 0.1319
g. Check student's solution.
h. Decision: Do not reject null hypothesis; Conclusion: There is insufficient evidence to conclude that the mean lap times are different.

Use the following information to answer the next two exercises. Table lists the number of pages in four different types of magazines.

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Q 13.4.4

Using a significance level of 5%, test the hypothesis that the four magazine types have the same mean length.

Q 13.4.5

Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Use a new solution sheet. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

S 13.4.6

a. $H_a : \mu_d = \mu_n = \mu_h$
b. At least any two of the magazines have different mean lengths.
c. $df(\text{num}) = 2, df(\text{denom}) = 12$
d. F distribtuion
e. $F = 15.28$
f. $p\text{-value} = 0.001$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the Null Hypothesis.
   iii. Reason for decision: $p\text{-value} < \alpha$
   iv. Conclusion: There is sufficient evidence to conclude that the mean lengths of the magazines are different.

Q 13.4.7

A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that Table shows the results of a study.

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Q 13.4.8

Are the means for the final exams the same for all statistics class delivery types? Table shows the scores on final exams from several randomly selected classes that used the different delivery types.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

S 13.4.8

a. $H_0 : \mu_o = \mu_h = \mu_f$
b. At least two of the means are different.
c. $df(\text{n}) = 2, df(\text{d}) = 13$
d. $F_{2,13}$
e. 0.64
f. 0.5437
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: $p\text{-value} < \alpha$
iv. Conclusion: The mean scores of different class delivery are not different.

Q 13.4.9

Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics and Asians? Suppose that Table shows the results of a study.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

Q 13.4.10

Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that Table shows the results of a study.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

S 13.4.11

a. $H_0 : \mu_p = \mu_m = \mu_h$
b. At least any two of the means are different.
c. $df(\text{n}) = 2, df(\text{d}) = 12$
d. $F_{2,12}$
e. 3.13
f. 0.0807
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: $p\text{-value} < \alpha$
iv. Conclusion: There is not sufficient evidence to conclude that the mean numbers of daily visitors are different.

Q 13.4.12

Sanjay made identical paper airplanes out of three different weights of paper, light, medium and heavy. He made four airplanes from each of the weights, and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper Type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters

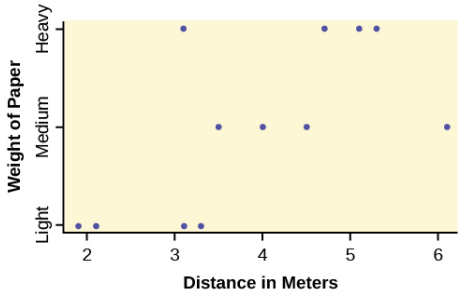


Figure 13.4.1.

a. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group? Yes or No.
b. Why is this a balanced design?
c. Calculate the sample mean and sample standard deviation for each group.
d. Does the weight of the paper have an effect on how far the plane will travel? Use a 1% level of significance. Complete the test using the method shown in the bean plant example in Example.

- variance of the group means _____
- $MS_{\text{between}} =$ _____
- mean of the three sample variances _____
- $MS_{\text{within}} =$ _____
- F statistic = _____
- $df(\text{num}) =$ _____, $df(\text{denom}) =$ _____
- number of groups _____
- number of observations _____
- p-value = _____ ($P(F >$ _____$) =$ _____)
- Graph the p-value.
- decision: _____
- conclusion: _____

Q 13.4.13

DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective, but persisted in the environment and over time became seen as harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). Finally there was a control line of non-selected or typical fruitflies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
16.4	26.7	28.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
14.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1% level of significance, are the mean rates of egg selection for the three strains of fruitfly different? If so, in what way? Specifically, the researchers were interested in whether or not the selectively bred strains were different from the nonselected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:

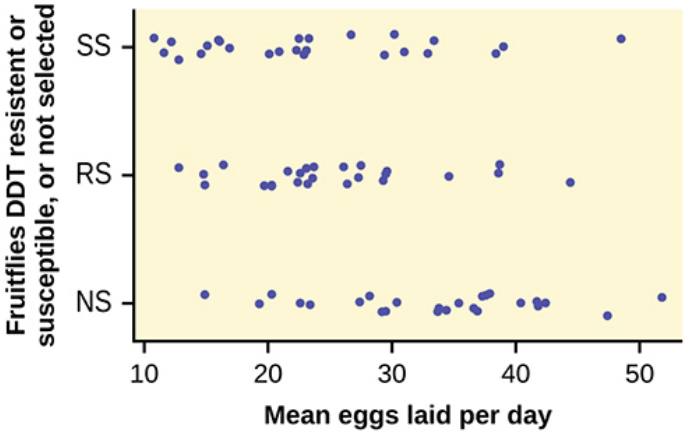


Figure 13.4.2.

S 13.4.13

The data appear normally distributed from the chart and of similar spread. There do not appear to be any serious outliers, so we may proceed with our ANOVA calculations, to see if we have good evidence of a difference between the three groups.

$H_0 : \mu_1 = \mu_2 = \mu_3$;

$H_a : \mu_i \neq \mu_j$ some $i \neq j$.

Define μ_1, μ_2, μ_3, as the population mean number of eggs laid by the three groups of fruit flies.

F statistic $= 8.6657$;

p-value $= 0.0004$

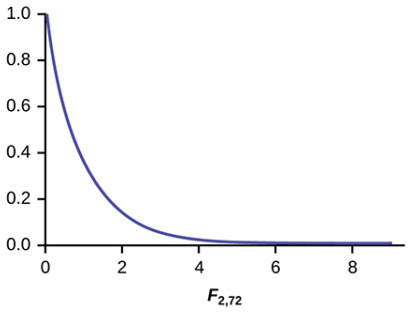


Figure 13.4.3.

Decision: Since the p-value is less than the level of significance of 0.01, we reject the null hypothesis.

Conclusion: We have good evidence that the average number of eggs laid during the first 14 days of life for these three strains of fruitflies are different.

Interestingly, if you perform a two sample t-test to compare the RS and NS groups they are significantly different ($p = 0.0013$). Similarly, SS and NS are significantly different ($p = 0.0006$). However, the two selected groups, RS and SS are not significantly different ($p = 0.5176$). Thus we appear to have good evidence that selection either for resistance or for susceptibility involves a reduced rate of egg production (for these specific strains) as compared to flies that were not selected for resistance or susceptibility to DDT. Here, genetic selection has apparently involved a loss of fecundity.

Q 13.4.14

The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms.

Traditionally we are taught that the normal human body temperature is 98.6 F. This is not quite correct for everyone. Are the mean temperatures among the four groups different?

Calculate 95% confidence intervals for the mean body temperature in each group and comment about the confidence intervals.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

13.5: Test of Two Variances

Use the following information to answer the next two exercises. There are two assumptions that must be true in order to perform an ***F*** test of two variances.

Exercise 13.5.2

Name one assumption that must be true.

Answer

The populations from which the two samples are drawn are normally distributed.

Exercise 13.5.3

What is the other assumption that must be true?

Use the following information to answer the next five exercises. Two coworkers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times and that his commute time is shorter. Test the claim at the 10% level.

Exercise 13.5.4

State the null and alternative hypotheses.

Answer

$H_0 : \sigma_1 = \sigma_2$

$H_a : \sigma_1 < \sigma_2$

or

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 < \sigma_2^2$

Exercise 13.5.5

What is s_1 in this problem?

Exercise 13.5.6

What is s_2 in this problem?

Answer

4.11

Exercise 13.5.7

What is n?

Exercise 13.5.8

What is the F statistic?

Answer

0.7159

Exercise 13.5.9

What is the p-value?

Exercise 13.5.10

Is the claim accurate?

Answer

No, at the 10% level of significance, we do not reject the null hypothesis and state that the data do not show that the variation in drive times for the first worker is less than the variation in drive times for the second worker.

Use the following information to answer the next four exercises. Two students are interested in whether or not there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are lower.

Exercise 13.5.11

State the null and alternative hypotheses.

Exercise 13.5.12

What is the F Statistic?

Answer

2.8674

Exercise 13.5.13

What is the p-value?

Exercise 13.5.14

At the 5% significance level, do we reject the null hypothesis?

Answer

Reject the null hypothesis. There is enough evidence to say that the variance of the grades for the first student is higher than the variance in the grades for the second student.

Use the following information to answer the next three exercises. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8 and the second cyclist has a variance of 32.1. The cyclists want to see if their variances are the same or different.

Exercise 13.5.15

State the null and alternative hypotheses.

Exercise 13.5.16

What is the F Statistic?

Answer

0.7414

Exercise 13.5.17

At the 5% significance level, what can we say about the cyclists' variances?

Q 13.5.1

Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Determine whether or not the variance in weight gain is statistically the same among Javier's and Linda's rats. Test at a significance level of 10%.

S 13.5.1

a. $H_0 : \sigma_1^2 = \sigma_2^2$
b. $H_a : \sigma_1^2 \neq \sigma_1^2$
c. $df(\text{num}) = 4; df(\text{denom}) = 4$
d. $F_{4,4}$
e. 3.00
f. $2(0.1563) = 0.3126$. Using the TI-83+/84+ function 2-SampFtest, you get the test statistic as 2.9986 and *p*-value directly as 0.3127. If you input the lists in a different order, you get a test statistic of 0.3335 but the p-value is the same because this is a two-tailed test.
g. Check student's solution.
h. Decision: Do not reject the null hypothesis; Conclusion: There is insufficient evidence to conclude that the variances are different.

Q 13.5.2

A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Determine whether or not the variance in mileage driven is statistically the same among the working class and professional (middle income) groups. Use a 5% significance level.

Q 13.5.3

Refer to the data from [link].

Examine practice laps 3 and 4. Determine whether or not the variance in lap time is statistically the same for those practice laps.

Use the following information to answer the next two exercises. The following table lists the number of pages in four different types of magazines.

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

S 13.5.3

a. $H_0 : \sigma_1^2 = \sigma_2^2$
b. $H_a : \sigma_1^2 \neq \sigma_1^2$
c. $df(\text{n}) = 19, df(\text{d}) = 19$
d. $F_{19,19}$
e. 1.13
f. 0.786
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: $p\text{-value} > \alpha$
iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

Q 13.5.4

Which two magazine types do you think have the same variance in length?

Q 13.5.5

Which two magazine types do you think have different variances in length?

S 13.5.5

The answers may vary. Sample answer: Home decorating magazines and news magazines have different variances.

Q 13.5.6

Is the variance for the amount of money, in dollars, that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that the Table shows the results of a study.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Q 13.5.7

Are the variances for incomes on the East Coast and the West Coast the same? Suppose that Table shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a level of significance of 0.05.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

S 13.5.7

a. $H_0 : \sigma_1^2 = \sigma_2^2$
b. $H_a : \sigma_1^2 \neq \sigma_1^2$
c. $df(\text{n}) = 7, df(\text{d}) = 6$
d. $F_{7,6}$
e. 0.8117
f. 0.7825
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: $p\text{-value} > \alpha$
iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

Q 13.5.8

Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of ten, with each receiving one of three doses of caffeine: 0 mg, 100 mg, 200 mg. This is approximately the amount in no, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?

Here are the data:

0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
242	248	246	245	246	248
244	245	250	248	247	252
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

Q 13.5.9

King Manuel I, Komnenus ruled the Byzantine Empire from Constantinople (Istanbul) during the years 1145 to 1180 A.D. The empire was very powerful during his reign, but declined significantly afterwards. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from a fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

Did the silver content of the coins change over the course of Manuel's reign?

Here are the means and variances of each coinage. The data are unbalanced.

	First	Second	Third	Fourth
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

S 13.5.9

Here is a strip chart of the silver content of the coins:

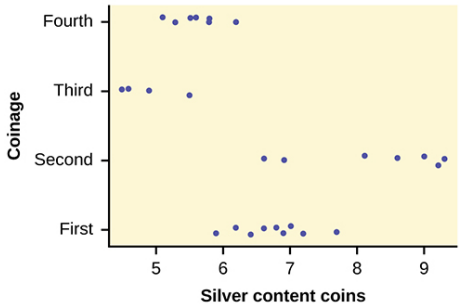


Figure 13.5.1.

While there are differences in spread, it is not unreasonable to use $ANOVA$ techniques. Here is the completed $ANOVA$ table:

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	37.748	$4-1=3$	12.5825	26.272
Error (Within)	11.015	$27-4=23$	0.4789	
Total	48.763	$27-1=26$		

$P(F > 26.272) = 0$;

Reject the null hypothesis for any alpha. There is sufficient evidence to conclude that the mean silver content among the four coinages are different. From the strip chart, it appears that the first and second coinages had higher silver contents than the third and fourth.

Q 13.5.10

The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to test whether the mean number of wins per team in the three American League divisions were the same or not. Note that the data are not balanced, as two divisions had five teams, while one had only four.

Division	Team	Wins
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72
Central	Cleveland	68
Central	Minnesota	66

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

S 13.5.10

Here is a stripchart of the number of wins for the 14 teams in the AL for the 2012 season.

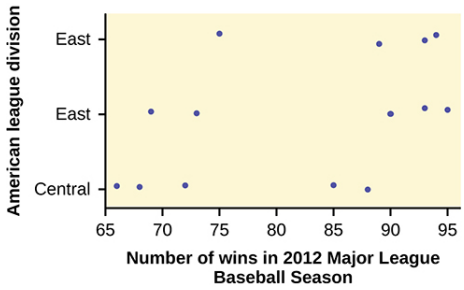


Figure 13.5.2.

While the spread seems similar, there may be some question about the normality of the data, given the wide gaps in the middle near the 0.500 mark of 82 games (teams play 162 games each season in MLB). However, one-way $ANOVA$ is robust.

Here is the $ANOVA$ table for the data:

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	344.16	3 – 1 = 2	172.08	26.272
Error (Within)	1,219.55	14 – 3 = 11	110.87	1.5521
Total	1,563.71	14 – 1 = 13		

$P(F > 1.5521) = 0.2548$

Since the p-value is so large, there is not good evidence against the null hypothesis of equal means. We decline to reject the null hypothesis. Thus, for 2012, there is not any have any good evidence of a significant difference in mean number of wins between the divisions of the American League.

13.6: Lab: One-Way ANOVA

This page titled 11.E: F Distribution and One-Way ANOVA (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **13.E: F Distribution and One-Way ANOVA (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

11.E: The Chi-Square Distribution (Optional Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

11.1: Introduction

11.2: Facts about the Chi-Square Distribution

Decide whether the following statements are true or false.

Q 11.2.1

As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.

S 11.2.1

true

Q 11.2.2

The standard deviation of the chi-square distribution is twice the mean.

Q 11.2.3

The mean and the median of the chi-square distribution are the same if $df = 24$.

S 11.2.3

false

11:3: Goodness-of-Fit Test

For each problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Q 11.3.1

A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in Table are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

The marital status distribution of the U.S. male population, ages 15 and older, is as shown in Table.Q 11.3.2

Marital Status	Percent	Expected Frequency
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in Table, rounding to two decimal places.

Marital Status	Frequency
never married	140
married	238
widowed	2
divorced/separated	20

S 11.3.2

Marital Status	Percent	Expected Frequency
never married	31.3	125.2
married	56.1	224.4
widowed	2.5	10
divorced/separated	10.1	40.4

a. The data fits the distribution.
b. The data does not fit the distribution.
c. 3
d. chi-square distribution with $df = 3$
e. 19.27
f. 0.0002
g. Check student's solution.
h. i. $\alpha = 0.05$
ii. Decision: Reject null
iii. Reason for decision: p-value $< \alpha$
iv. Conclusion: Data does not fit the distribution.

Use the following information to answer the next two exercises: The columns in Table contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African-American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

Q 11.3.3

Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

Q 11.3.4

Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

S 11.3.4

a. H_0: The local results follow the distribution of the U.S. AP examinee population
b. H_0: The local results do not follow the distribution of the U.S. AP examinee population
c. $df = 5$
d. chi-square distribution with $df = 5$
e. chi-square test statistic = 13.4
f. $p\text{-value} = 0.0199$
g. Check student's solution.
h. i. $\alpha = 0.05$

ii. Decision: Reject null when $\alpha = 0.05$

iii. Reason for Decision: $p\text{-value} < \alpha$

iv. Conclusion: Local data do not fit the AP Examinee Distribution.

v. Decision: Do not reject null when $\alpha = 0.01$

vi. Conclusion: There is insufficient evidence to conclude that local data do not follow the distribution of the U.S. AP examinee distribution.

Q 11.3.5

The City of South Lake Tahoe, CA, has an Asian population of 1,419 people, out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the Manhattan, NY, area yielded the data in Table. Conduct a goodness-of-fit test to

determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Race	Lake Tahoe Frequency	Manhattan Frequency
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Use the following information to answer the next two exercises: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for Exercise and Exercise. The second column in each table does not add to 100% because of rounding.

Q 11.3.6

Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Women - Expected Major	Women - Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

S 11.3.6

a. H_0: The actual college majors of graduating females fit the distribution of their expected majors
b. H_a: The actual college majors of graduating females do not fit the distribution of their expected majors
c. $df = 10$
d. chi-square distribution with $df = 10$
e. $\text{test statistic} = 11.48$
f. $p\text{-value} = 0.3211$
g. Check student's solution.

h. i. $\alpha = 0.05$
ii. Decision: Do not reject null when $\alpha = 0.05$ and $\alpha = 0.01$
iii. Reason for decision: $p\text{-value} > \alpha$
iv. Conclusion: There is insufficient evidence to conclude that the distribution of actual college majors of graduating females fits the distribution of their expected majors.

Q 11.3.7

Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Men - Expected Major	Men - Actual Major
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

Read the statement and decide whether it is true or false.

Q 11.3.8

In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.

S 11.3.8

true

Q 11.3.9

In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

Q 11.3.10

Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week or not.

S 11.3.10

true

Q 11.3.11

The test to use to determine if a six-sided die is fair is a goodness-of-fit test.

Q 11.3.12

In a goodness-of fit test, if the *p*-value is 0.0113, in general, do not reject the null hypothesis.

S 11.3.12

false

Q 11.3.13

A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means any one type of recyclable material such as plastic or aluminum. Table shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5% significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business Type	Number in class	Observed Number that recycle one commodity	Expected number that recycle one commodity
Office	35	19	17.5
Retail/Wholesale	48	27	24
Food/Restaurants	53	35	26.5
Manufacturing/Medical	52	21	26
Hotel/Mixed	24	9	12

Q 11.3.14

Table contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (Years)	Obese (Percentage)	Expected USA average (Percentage)
20–30	75.0	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21.0	39.7

S 11.3.14

a. H_0: Surveyed obese fit the distribution of expected obese
b. H_a: Surveyed obese do not fit the distribution of expected obese
c. $df = 4$
d. chi-square distribution with $df = 4$
e. test statistic $= 54.01$
f. p-value $= 0$
g. Check student's solution.
h. i. $\alpha : 0.05$

ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p-value $< \alpha$
iv. Conclusion: At the 5% level of significance, from the data, there is sufficient evidence to conclude that the surveyed obese do not fit the distribution of expected obese.

11.4: Test of Independence

For each problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E for the chi-square solution sheet. Round expected frequency to two decimal places.

Q 11.4.1

A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Q 11.4.2

Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in Table. Conduct a test of independence.

Family Size	Sub & Compact	Mid-size	Full-size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

S 11.4.2

a. H_0: Car size is independent of family size.
b. H_a: Car size is dependent on family size.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic $= 15.8284$
f. p-value $= 0.0706$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value $> \alpha$
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that car size and family size are dependent.

Q 11.4.3

College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Table shows the data. Conduct a test of independence.

Major	< $50,000	$50,000 – $68,999	$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Q 11.4.4

Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in Table. Conduct a test of independence.

Location	20–29	30–39	40–49	50 and over
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

a. H_0: Honeymoon locations are independent of bride's age.
b. H_a: Honeymoon locations are dependent on bride's age.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic $= 15.7027$
f. p-value $= 0.0734$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value $> \alpha$
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that honeymoon location and bride age are dependent.

Q 11.4.5

A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	18 - 25	26 - 30	31 - 40	41 and over
racquetball	42	58	30	46
tennis	58	76	38	65
swimming	72	60	65	33

Q 11.4.6

A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in Table. Conduct a test of independence.

Type of Fries	Northeast	South	Central	West
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

S 11.4.6

a. H_0: The types of fries sold are independent of the location.
b. H_a: The types of fries sold are dependent on the location.
c. $df = 6$
d. chi-square distribution with $df = 6$
e. test statistic $= 18.8369$
f. p-value $= 0.0044$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p-value $> \alpha$
iv. Conclusion: At the 5% significance level, There is sufficient evidence that types of fries and location are dependent.

Q 11.4.7

According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $200,000	$200,000–$400,000	$401,001–$1,000,000	$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Q 11.4.8

Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< $30,000	15	25	10	5
$30,000–$40,000	20	40	70	30
$40,000–$50,000	10	20	40	55
$50,000–$60,000	5	10	20	60
$60,000+	0	5	10	150

S 11.4.8

a. H_0: Salary is independent of level of education.
b. H_a: Salary is dependent on level of education.
c. $df = 12$
d. chi-square distribution with $df = 12$
e. test statistic $= 255.7704$
f. p-value $= 0$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, There is sufficient evidence that types of fries and location are dependent.

Read the statement and decide whether it is true or false.

Q 11.4.9

The number of degrees of freedom for a test of independence is equal to the sample size minus one.

Q 11.4.10

The test for independence uses tables of observed and expected data values.

S 11.4.10

true

Q 11.4.11

The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

Q 11.4.12

In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

S 11.4.12

true

Q 11.4.13

An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on Table, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

U.S. region/Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Row total
East	8	31	27	8	15	7	96
Midwest	10	32	22	11	15	6	96
West	12	21	22	19	15	8	97
South	15	28	30	8	15	6	102
Column Total	45	112	101	46	60	27	391

Q 11.4.14

Table provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Age Group\ Net Worth Value (in millions of US dollars)	**1–5**	**6–24**	**≥25**	**Row Total**
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

S 11.4.14

a. H_0: Age is independent of the youngest online entrepreneurs' net worth.
b. H_5: Age is dependent on the net worth of the youngest online entrepreneurs.
c. $df = 2$
d. chi-square distribution with $df = 2$
e. test statistic $= 1.76$
f. p-value $= 0.4144$
g. Check student's solution.
h.
   i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value $> \alpha$
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that age and net worth for the youngest online entrepreneurs are dependent.

Q 11.4.15

A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in Table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/Ethnicity	**Asian-American**	**White/Non-Hispanic**	**African-American**	**Latino**	**Row Total**
Against tax	48	433	41	160	628
In Favor of tax	54	234	24	147	459
No opinion	16	43	16	19	84
Column Total	118	710	71	272	1171

11.5: Test for Homogeneity

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Q 11.5.1

A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in Table. Conduct a test of homogeneity. Test at a 5% level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

S 11.5.1

a. H_0: The distribution for personality types is the same for both majors
b. H_a: The distribution for personality types is not the same for both majors
c. $df = 4$
d. chi-square with $df = 4$
e. test statistic $= 3.01$
f. p-value $= 0.5568$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value $> \alpha$
   iv. Conclusion: There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors.

Q 11.5.2

Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place is shown in Table. Conduct a test for homogeneity at a 5% level of significance.

	French Toast	Pancakes	Waffles	Omelettes
Men	47	35	28	53
Women	65	59	55	60

Q 11.5.3

A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5% level of significance.

S 11.5.3

a. H_0: The distribution for fish caught is the same in Green Valley Lake and in Echo Lake.
b. H_a: The distribution for fish caught is not the same in Green Valley Lake and in Echo Lake.
c. $df = 3$
d. chi-square with $df = 3$
e. test statistic $= 11.75$
f. p-value $= 0.0083$
g. Check student's solution.
h. i. $\alpha : 0.05$
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value $> \alpha$

iv. Conclusion: There is evidence to conclude that the distribution of fish caught is different in Green Valley Lake and in Echo Lake

Q 11.5.4

In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In Table you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5% significance level. Did you expect the result you obtained?

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)	Row Total
Concern about the environment of other schools	1,321	309	1,630
Dissatisfaction with academic instruction at other schools	1,096	258	1,354
To provide religious or moral instruction	1,257	540	1,797
Child has special needs, other than physical or mental	315	55	370
Nontraditional approach to child's education	984	99	1,083
Other reasons (e.g., finances, travel, family time, etc.)	485	216	701
Column Total	5,458	1,477	6,935

Q 11.5.5

When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in Table shows the average energy use (in units of kg of oil equivalent per capita) in the USA and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5% significance level.

Year	European Union	United States	Row Total
2010	3,413	7,164	10,557
2009	3,302	7,057	10,359
2008	3,505	7,488	10,993
2007	3,537	7,758	11,295
2006	3,595	7,697	11,292
2005	3,613	7,847	11,460
Column Total	45,011	20,965	65,976

S 11.5.5

a. H_0: The distribution of average energy use in the USA is the same as in Europe between 2005 and 2010.
b. H_a: The distribution of average energy use in the USA is not the same as in Europe between 2005 and 2010.

c. $df = 4$
d. chi-square with $df = 4$
e. test statistic $= 2.7434$
f. p-value $= 0.7395$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value $> \alpha$
iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the average energy use values in the US and EU are not derived from different distributions for the period from 2005 to 2010.

Q 11.5.6

The Insurance Institute for Highway Safety collects safety information about all types of cars every year, and publishes a report of Top Safety Picks among all cars, makes, and models. Table presents the number of Top Safety Picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the Top Safety Picks safety award has remained the same between 2009 and 2013. Derive your results at the 5% significance level.

Year \ Car Type	Small	Mid-Size	Large	Small SUV	Mid-Size SUV	Large SUV	Row Total
2009	12	22	10	10	27	6	87
2013	31	30	19	11	29	4	124
Column Total	43	52	29	21	56	10	211

11.6: Comparison of the Chi-Square Tests

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Q 11.6.1

Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of some randomly selected community college students, 43 used a computer, 102 used a calculator with built in statistics functions, and 65 used a table from the textbook. Of some randomly selected university students, 28 used a computer, 33 used a calculator with built in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.

S 11.6.1

a. H_0: The distribution for technology use is the same for community college students and university students.
b. H_a: The distribution for technology use is not the same for community college students and university students.
c. $df = 2$
d. chi-square with $df = 2$
e. test statistic $= 7.05$
f. p-value $= 0.0294$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p-value $> \alpha$
iv. Conclusion: There is sufficient evidence to conclude that the distribution of technology use for statistics homework is not the same for statistics students at community colleges and at universities.

Read the statement and decide whether it is true or false.

Q 11.6.2

If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

11.7: Test of a Single Variance

Use the following information to answer the next twelve exercises: Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

Q 11.7.1

Is the traveler disputing the claim about the average or about the variance?

Q 11.7.2

A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.

S 11.7.2

225

Q 11.7.3

Is this a right-tailed, left-tailed, or two-tailed test?

Q 11.7.4

H_0: _____

S 11.7.4

$H_0 : \sigma^2 \leq 150$

Q 11.7.5

$df =$ _____

Q 11.7.6

chi-square test statistic = _____

S 11.7.6

36

Q 11.7.7

p-value = _____

Q 11.7.8

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.

S 11.7.8

Check student's solution.

Q 11.7.9

Let $\alpha = 0.05$

Decision: _____

Conclusion (write out in a complete sentence.): _____

Q 11.7.10

How did you know to test the variance instead of the mean?

S 11.7.10

The claim is that the variance is no more than 150 minutes.

Q 11.7.11

If an additional test were done on the claim of the average delay, which distribution would you use?

Q 11.7.12

If an additional test were done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

S 11.7.12

a Student's t- or normal distribution

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Q 11.7.13

A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

Q 11.7.14

Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of \$84 and a sample standard deviation of \$12, test the claim that the standard deviation is greater than \$15.

S 11.7.14

a. $H_0 : \sigma = 15$
b. $H_a : \sigma > 15$
c. $df = 42$
d. chi-square with $df = 42$
e. test statistic = 26.88
f. $p\text{-value} = 0.9663$
g. Check student's solution.
h.
   i. $alpha = 0.05$
   ii. Decision: Do not reject null hypothesis.
   iii. Reason for decision: $p\text{-value} > \alpha$
   iv. Conclusion: There is insufficient evidence to conclude that the standard deviation is greater than 15.

Q 11.7.15

Isabella, an accomplished **Bay to Breakers** runner, claims that the standard deviation for her time to run the 7.5 mile race is at most three minutes. To test her claim, Rupinder looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

Q 11.7.16

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights

surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

S 11.7.16

a. $H_0 : \sigma \leq 3$
b. $H_a : \sigma > 3$
c. $df = 17$
d. chi-square distribution with $df = 17$
e. test statistic = 28.73
f. $p\text{-value} = 0.0371$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: $p\text{-value} < \alpha$
iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is greater than three.

Q 11.7.17

The number of births per woman in China is 1.6 down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had had. The results are shown in Table. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of births	Frequency
0	5
1	30
2	10
3	5

Q 11.7.18

According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

S 11.7.18

a. $H_0 : \sigma = 2$
b. $H_a : \sigma \neq 2$
c. $df = 14$
d. chi-square distribution with $df = 14$
e. chi-square test statistic = 5.2094
f. $p\text{-value} = 0.0346$
g. Check student's solution.
h. i. $\alpha : 0.05$
ii. Decision: Reject the null hypothesis
iii. Reason for decision: $p\text{-value} < \alpha$
iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is different than 2.

Q 11.7.19

The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a ten-ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of two oz.

Q 11.7.20

You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $1,249 with a very narrow standard deviation of $25. You find a website that has a price comparison for the same computer at a series of stores as follows: $1,299; $1,229.99; $1,193.08; $1,279; $1,224.95; $1,229.99; $1,269.95; $1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5% significance level. As a potential buyer, what would be the practical conclusion from your analysis?

S 11.7.20

a. $H_0 : \sigma = 25^2$
b. $H_a : \sigma > 25^2$
c. $df = n - 1 = 7$
d. test statistic: $\chi^2 = \chi_7^2 = \frac{(n-1)s^2}{25^2} = \frac{(8-1)(34.29)^2}{25^2} = 13.169$
e. $p\text{-value} : P(\chi_7^2 > 13.169) = 1 - P(\chi_7^2 \leq 13.169) = 0.0681$
f. i. $\alpha : 0.05$
   ii. Decision: Do not reject the null hypothesis
   iii. Reason for decision: $p\text{-value} < \alpha$
   iv. Conclusion: At the 5% level, there is insufficient evidence to conclude that the variance is more than 625.

Q 11.7.21

A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 g, and there is a maximum allowed weight tolerance of 5% above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements. Conduct an appropriate hypothesis test.

a. at the 5% significance level
b. at the 1% significance level

Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; 172;

11.8: Lab 1: Chi-Square Goodness-of-Fit

11.9: Lab 2: Chi-Square Test of Independence

This page titled 11.E: The Chi-Square Distribution (Optional Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **11.E: The Chi-Square Distribution (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

12: Nonparametric Statistics

Because distribution-free tests do not assume normality, they can be less susceptible to non-normality and extreme values. Therefore, they can be more powerful than the standard tests of means that assume normality.

12.1: Benefits of Distribution Free Tests
12.2: Randomization Tests - Two Conditions
12.3: Randomization Tests - Two or More Conditions
12.4: Randomization Association
12.5: Fisher's Exact Test
12.6: Rank Randomization Two Conditions
12.7: Rank Randomization Two or More Conditions
12.8: Rank Randomization for Association
12.9: Statistical Literacy Standard
12.10: Wilcoxon Signed-Rank Test
12.11: Kruskal–Wallis Test
12.12: Spearman Rank Correlation
12.13: Choosing the Right Test
12.E: Distribution Free Tests (Exercises)

Contributors and Attributions

- Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

This page titled 12: Nonparametric Statistics is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

12.1: Benefits of Distribution Free Tests

Learning Objectives

- State how distribution-free tests can avoid an inflated **Type I** error rate
- State how how distribution-free tests can affect power

Most tests based on the normal distribution are said to be robust when the assumption of normality is violated. To the extent to which actual probability values differ from nominal probability values, the actual probability values tend to be higher than the nominal p values. For example, if the probability of a difference as extreme or more extreme were 0.04, the test might report that the probability value is 0.06. Although this sounds like a good thing because the Type I error rate is lower than the nominal rate, it has a serious downside: reduced power. When the null hypothesis is false, then the probability of rejecting the null hypothesis can be substantially lower than it would have been if the distributions were distributed normally.

Tests assuming normality can have particularly low power when there are extreme values or outliers. A contributing factor is the sensitivity of the mean to extreme values. Although transformations can ameliorate this problem in some situations, they are not a universal solution.

Tests assuming normality often have low power for leptokurtic distributions. Transformations are generally less effective for reducing kurtosis than for reducing skew.

Because distribution-free tests do not assume normality, they can be less susceptible to non-normality and extreme values. Therefore, they can be more powerful than the standard tests of means that assume normality.

This page titled 12.1: Benefits of Distribution Free Tests is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.1: Benefits of Distribution Free Tests** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.2: Randomization Tests - Two Conditions

Learning Objectives

- Compute a randomization test of the difference between independent groups

The data in Table 12.2.1 are from a fictitious experiment comparing an experimental group with a control group. The scores in the Experimental Group are generally higher than those in the Control Group with the Experimental Group mean of 14 being considerably higher than the Control Group mean of 4. Would a difference this large or larger be likely if the two treatments had identical effects? The approach taken by randomization tests is to consider all possible ways the values obtained in the experiment could be assigned to the two groups. Then, the location of the actual data within the list is used to assess how likely a difference that large or larger would occur by chance.

Table 12.2.1: Fictitious data

Experimental	Control
7	0
8	2
11	5
30	9

First, consider all possible ways the 8 values could be divided into two sets of 4. We can apply the formula from the section on Permutations and Combinations for the number of combinations of n items taken r at a time and find that there are 70 ways.

$$ {}_n\mathrm{C}_r = \frac{n!}{(n-r)!r!} = \frac{8!}{(8-4)!4!} = 70 \tag{12.2.1} $$

Of these 70 ways of dividing the data, how many result in a difference between means of 10 or larger? From Table 12.2.1you can see that there are two rearrangements that would lead to a bigger difference than 10:

a. the score of 7 could have been in the Control Group with the score of 9 in the Experimental Group and
b. the score of 8 could have been in the Control Group with the score of 9 in the Experimental Group

Therefore, including the actual data, there are 3 ways to produce a difference as large or larger than the one obtained. This means that if assignments to groups were made randomly, the probability of this large or a larger advantage of the Experimental Group is $3/70 = 0.0429$. Since only one direction of difference is considered (Experimental larger than Control), this is a one-tailed probability. The two-tailed probability is 0.0857 since there are 6/70 ways to arrange the data so that the absolute value of the difference between groups is as large or larger than the one obtained.

Clearly, this type of analysis would be very time consuming for even moderate sample sizes. Therefore, it is most useful for very small sample sizes.

An alternate approach made practical by computer software is to randomly divide the data into groups thousands of times and count the proportion of times the difference is as big or bigger than that found with the actual data. If the number of times the data are divided randomly is very large, then this proportion will be very close to the proportion you would get if you had listed all possible ways the data could be divided. The link below goes to a web page that can do these calculations.

Statkey

This page titled 12.2: Randomization Tests - Two Conditions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.2: Randomization Tests - Two Conditions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.3: Randomization Tests - Two or More Conditions

Learning Objectives

- Compute a randomization test for differences among more than two conditions

The method of randomization for testing differences among more than two means is essentially very similar to the method when there are exactly two means. Table 12.3.1 shows the data from a fictitious experiment with three groups.

Table 12.3.1: Fictitious data

T1	T2	Control
7	14	0
8	19	2
11	21	5
12	122	9

The first step in a randomization test is to decide on a test statistic. Then we compute the proportion of the possible arrangements of the data for which that test statistic is as large as or larger than the arrangement of the actual data. When comparing several means, it is convenient to use the F ratio. The F ratio is computed not to test for significance directly, but as a measure of how different the groups are. For these data, the F ratio for a one-way ANOVA is 2.06.

The next step is to determine how many arrangements of the data result in as large or larger F ratios. There are 6 arrangements that lead to the same F of 2.06: the six arrangements of the three columns. One such arrangement is shown in Table 12.3.2. The six are:

1. $T1$, $T2$, Control
2. $T1$, Control, $T2$
3. $T2$, $T1$, Control
4. $T2$, Control, $T1$
5. Control, $T1$, $T2$
6. Control, $T2$, $T1$

For each of the 6 arrangements there are two changes that lead to a higher F ratio: swapping the 7 for the 9 (which gives an F of 2.08) and swapping the 8 for the 9 (which gives an F of 2.07). The former of these two is shown in Table 12.3.3.

Table 12.3.2: Fictitious data with data for $T2$ and Control swapped

T1	Control	T2
7	14	0
8	19	2
11	21	5
12	122	9

***Table* 12.3.3:** *Data from Table* 12.3.1 *with the* 7 *and the* 9 *swapped*

T1	T2	Control
9	14	0
8	19	2
11	21	5
12	122	7

Thus, there are six arrangements, each with two swaps that lead to a larger F ratio. Therefore, the number of arrangements with an F as large or larger than the actual arrangement is 6 (for the arrangements with the same F) + 12 (for the arrangements with a larger F), which makes 18 in all.

The next step is to determine the total number of possible arrangements. This can be computed from the following formula:

$$\text{Arrangements} = (n!)^k = (4!)^3 = 13,824 \tag{12.3.1}$$

where n is the number of observations in each group (assumed to be the same for all groups), and k is the number of groups. Therefore, the proportion of arrangements with an F as large or larger than the F of 2.06 obtained with the data is

$$\frac{18}{13,824} = 0.0013. \tag{12.3.2}$$

Thus, if there were no treatment effect, it is very unlikely that an F as large or larger than the one obtained would be found.

This page titled 12.3: Randomization Tests - Two or More Conditions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.3: Randomization Tests - Two or More Conditions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.4: Randomization Association

skills to develop

- Compute a randomization test for Pearson's r

A significance test for Pearson's r is described in the section inferential statistics for b and r. The significance test described in that section assumes normality. This section describes a method for testing the significance of r that makes no distributional assumptions.

Table 12.4.1: Example data

X	1.0	2.4	3.8	4.0	11.0
Y	1.0	2.0	2.3	3.7	2.5

The approach is to consider the X variable fixed and compare the correlation obtained in the actual data to the correlations that could be obtained by rearranging the Y variable. For the data shown in Table 12.4.1, the correlation between X and Y is 0.385. There is only one arrangement of Y that would produce a higher correlation. This arrangement is shown in Table 12.4.2 and the r is 0.945. Therefore, there are two arrangements of Y that lead to correlations as high or higher than the actual data.

Table 12.4.1: The example data arranged to give the highest r

X	Y
1.0	1.0
2.4	2.0
3.8	2.3
4.0	2.5
11.0	3.7

The next step is to calculate the number of possible arrangements of Y. The number is simply $N!$, where N is the number of pairs of scores. Here, the number of arrangements is $5! = 120$. Therefore, the probability value is $2/120 = 0.017$. Note that this is a one-tailed probability since it is the proportion of arrangements that give an r as large or larger. For the two-tailed probability, you would also count arrangements for which the value of r were less than or equal to -0.385. In randomization tests, the two-tailed probability is not necessarily double the one-tailed probability.

This page titled 12.4: Randomization Association is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.4: Randomization Association** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.5: Fisher's Exact Test

Learning Objectives

- State the situation when Fisher's exact test can be used
- Calculate Fisher's exact test
- Describe how conservative the Fisher exact test is relative to a Chi Square test

The chapter on Chi Square showed one way to test the relationship between two nominal variables. A special case of this kind of relationship is the difference between proportions. This section shows how to compute a significance test for a difference in proportions using a randomization test. Suppose, in a fictitious experiment, 4 subjects in an Experimental Group and 4 subjects in a Control Group are asked to solve an anagram problem. Three of the 4 subjects in the Experimental Group and none of the subjects in the Control Group solved the problem. Table 12.5.1 shows the results in a contingency table.

Table 12.5.1: Anagram Problem Data

	Experimental	Control	Total
Solved	3	0	3
Did Not Solve	1	4	5
Total	4	4	8

The significance test we are going to perform is called the Fisher Exact Test. The basic idea is to take the row totals and column totals as "given" and add the probability of obtaining the pattern of frequencies obtained in the experiment and the probabilities of all other patterns that reflect a greater difference between conditions. The formula for obtaining any given pattern of frequencies is:

$$\frac{n!(N-n)!R!(N-R)!}{r!(n-r)!(R-r)!(N-n-R+r)!N!} \tag{12.5.1}$$

where N is the total sample size (8), n is the sample size for the first group (4), r is the number of successes for the first group (3), and R is the total number of successes (3). For this example, the probability is

$$\frac{4!(8-4)!3!(8-3)!}{3!(4-3)!(3-3)!(8-4-3+8)!8!} = 0.0714 \tag{12.5.2}$$

Since more extreme outcomes do not exist given the row and column totals, the p value is 0.0714. This is a one-tailed probability since it only considers outcomes as extreme or more extreme favoring the Experimental Group. An equally extreme outcome favoring the Control Group is shown in Table 12.5.2, which also has a probability of 0.0714. Therefore, the two-tailed probability is 0.1428. Note that in the Fisher Exact Test, the two-tailed probability is not necessarily double the one-tailed probability.

Table 12.5.2: Anagram Problem Favoring Control Group

	Experimental	Control	Total
Solved	0	3	3
Did not Solve	4	1	5
Total	4	4	8

The Fisher Exact Test is "exact" in the sense that it is not based on a statistic that is approximately distributed as, for example, Chi Square. However, because it assumes that both marginal totals are fixed, it can be considerably less powerful than the Chi Square test. Even though the Chi Square test is an approximate test, the approximation is quite good in most cases and tends to have too low a Type I error rate more often than too high a Type I error rate (see for yourself using this simulation).

This page titled 12.5: Fisher's Exact Test is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.5: Fisher's Exact Test** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.6: Rank Randomization Two Conditions

Learning Objectives

- State the difference between a randomization test and a rank randomization test
- Describe why rank randomization tests are more common
- Be able to compute a Mann-Whitney U test

The major problem with randomization tests is that they are very difficult to compute. Rank randomization tests are performed by first converting the scores to ranks and then computing a randomization test. The primary advantage of rank randomization tests is that there are tables that can be used to determine significance. The disadvantage is that some information is lost when the numbers are converted to ranks. Therefore, rank randomization tests are generally less powerful than randomization tests based on the original numbers.

There are several names for rank randomization tests for differences in central tendency. The two most common are the Mann-Whitney U test and the Wilcoxon Rank Sum test.

Consider the data shown in Table **12.6.1** that were used as an example in the section on randomization tests.

Table **12.6.1**: Fictitious data

Experimental	Control
7	0
8	2
11	5
30	9

A rank randomization test on these data begins by converting the numbers to ranks.

Table **12.6.2**: Fictitious data converted to ranks. Rank sum = **24**

Experimental	Control
4	1
5	2
7	3
8	6

The probability value is determined by computing the proportion of the possible arrangements of these ranks that result in a difference between ranks as large or larger than those in the actual data (Table **12.6.2**). Since the sum of the ranks (the numbers $1-8$) is a constant (**36** in this case), we can use the computational shortcut of finding the proportion of arrangements for which the sum of the ranks in the Experimental Group is as high or higher than the sum here ($4+5+7+8=24$).

First, consider how many ways the **8** values could be divided into two sets of **4**. We can apply the formula from the section on Permutations and Combinations for the number of combinations of n items taken r at a time ($\mathbf{n = \text{the total number of observations; } r = \text{the number of observations in the first group}}$) and find that there are **70** ways.

$$ {}_n\mathrm{C}_r = \frac{n!}{(n-r)!r!} = \frac{8!}{(8-4)!4!} = 70 \qquad (12.6.1) $$

Of these **70** ways of dividing the data, how many result in a sum of ranks of **24** or more? Tables $3-5$ show three rearrangements that would lead to a rank sum of **24** or larger.

Table **12.6.3**: Rearrangement of data converted to ranks. Rank sum = **26**

Experimental	Control
6	1
5	2
7	3
8	4

Table 12.6.4: Rearrangement of data converted to ranks. Rank sum = 25

Experimental	Control
4	1
6	2
7	3
8	5

Table 12.6.5: Rearrangement of data converted to ranks. Rank sum = 24

Experimental	Control
3	1
6	2
7	4
8	5

Therefore, the actual data represent 1 arrangement with a rank sum of 24 or more and the 3 arrangements represent three others. Therefore, there are 4 arrangements with a rank sum of 24 or more. This makes the probability equal to $4/70 = 0.057$. Since only one direction of difference is considered (Experimental larger than Control), this is a one-tailed probability. The two-tailed probability is $(2)(0.057) = 0.114$ since there are $8/70$ ways to arrange the data so that the sum of the ranks is either

a. as large or larger or
b. as small or smaller than the sum found for the actual data.

The beginning of this section stated that rank randomization tests were easier to compute than randomization tests because tables are available for rank randomization tests. Table 12.6.6 can be used to obtain the critical values for equal sample sizes of $4 - 10$.

Table for unequal sample sizes

For the present data, both n_1 and $n_2 = 4$ so, as can be determined from the table, the rank sum for the Experimental Group must be at least 25 for the difference to be significant at the 0.05 level (one-tailed). Since the sum of ranks equals 24, the probability value is somewhat above 0.05. In fact, by counting the arrangements with the sum of ranks greater than or equal to 24, we found that the probability value is 0.057. Naturally a table can only give the critical value rather than the p value itself. However, with a larger sample size such as 10 subjects per group, it becomes very time-consuming to count all arrangements equaling or exceeding the rank sum of the data. Therefore, for practical reasons, the critical value sometimes suffices.

Table 12.6.6: Critical values. One-Tailed Test. Rank Sum for Higher Group

n_1	n_2	0.20	0.10	0.05	0.025	0.01	0.005
4	4	22	23	25	26	.	.
5	5	33	35	36	38	39	40
6	6	45	48	50	52	54	55
7	7	60	64	66	69	71	73
8	8	77	81	85	87	91	93
9	9	96	101	105	109	112	115
10	10	117	123	128	132	136	139

For larger sample sizes than covered in the tables, you can use the following expression that is approximately normally distributed for moderate to large sample sizes.

$$Z = \frac{W_a - n_a(n_a + n_b + 1)/2}{\sqrt{n_a n_b (n_a + n_b + 1)/12}} \tag{12.6.2}$$

where:

- W_a is the sum of the ranks for the first group
- n_a is the sample size for the first group
- n_b is the sample size for the second group
- Z is the test statistic

The probability value can be determined from Z using the normal distribution calculator.

The data from the Stereograms Case Study can be analyzed using this test. For these data, the sum of the ranks for Group 1 (W_a) is **1911**, the sample size for Group 1 (n_a) is **43**, and the sample size for Group 2 (n_b) is **35**. Plugging these values into the formula results in a Z of **2.13**, which has a two-tailed p of **0.033**.

This page titled 12.6: Rank Randomization Two Conditions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.6: Rank Randomization Two Conditions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.7: Rank Randomization Two or More Conditions

Learning Objectives

- Compute the Kruskal-Wallis test

The Kruskal-Wallis test is a rank-randomization test that extends the Wilcoxon test to designs with more than two groups. It tests for differences in central tendency in designs with one between-subjects variable. The test is based on a statistic H that is approximately distributed as Chi Square. The formula for H is shown below:

$$H = -3(N+1) + \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} \tag{12.7.1}$$

where

- N is the total number of observations
- T_i is the sum of ranks for the ith group
- n_i is the sample size for the ith group
- k is the number of groups

The first step is to convert the data to ranks (ignoring group membership) and then find the sum of the ranks for each group. Then, compute H using the formula above. Finally, the significance test is done using a Chi Square distribution with $k-1$ degrees of freedom.

For the "Smiles and Leniency" case study, the sums of the ranks for the four conditions are:

- False: 2732.0
- Felt: 2385.5
- Miserable: 2424.5
- Neutral: 1776.0

Note that since there are "ties" in the data, the mean rank of the ties is used. For example, there were 10 scores of 2.5 which tied for ranks $4-13$. The average of the numbers $4, 5, 6, 7, 8, 9, 10, 11, 12, 13$ is 8.5. Therefore, all values of 2.5 were assigned ranks of 8.5.

The sample size for each group is 34.

$$H = -3(136+1) + \frac{12}{(136)(137)} \left(\frac{(2732)^2}{34} + \frac{(2385.5)^2}{34} + \frac{(2424.5)^2}{34} + \frac{(1776)^2}{34} \right) = 9.28 \tag{12.7.2}$$

Using the Chi Square Calculator for $\text{Chi Square} = 9.28$ with $4-1=3$ df results in a p value of 0.0258. Thus the null hypothesis of no leniency effect can be rejected.

This page titled 12.7: Rank Randomization Two or More Conditions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.7: Rank Randomization Two or More Conditions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.8: Rank Randomization for Association

Learning Objectives

- Compute Spearman's ρ
- Test Spearman's ρ for significance

The rank randomization test for association is equivalent to the randomization test for Pearson's r except that the numbers are converted to ranks before the analysis is done. Table 12.8.1 shows 5 values of X and Y. Table 12.8.2 shows these same data converted to ranks (separately for X and Y).

Table 12.8.1: Example data

X	Y
1.0	1.0
2.4	2.0
3.8	2.3
4.0	3.7
11.0	2.5

Table 12.8.2: Ranked data

X	Y
1	1
2	2
3	3
4	5
5	4

The approach is to consider the X variable fixed and compare the correlation obtained in the actual ranked data to the correlations that could be obtained by rearranging the Y variable ranks. For the ranked data shown in Table 12.8.2, the correlation between X and Y is 0.90. The correlation of ranks is called "Spearman's ρ."

Table 12.8.3: Ranked data with correlation of 1.0

X	Y
1	1
2	2
3	3
4	4
5	5

There is only one arrangement of Y that produces a higher correlation than 0.90: A correlation of 1.0 results if the fourth and fifth observations' Y values are switched (see Table 12.8.3). There are also three other arrangements that produce an r of 0.90 (see Tables 12.8.4, 12.8.5, and 12.8.6). Therefore, there are five arrangements of Y that lead to correlations as high or higher than the actual ranked data (Tables 12.8.2 through 12.8.6).

Table 12.8.4: Ranked data with correlation of 0.90

X	Y
1	1
2	2
3	4
4	3
5	5

Table 12.8.5: Ranked data with correlation of 0.90

X	Y
1	1
2	3
3	2
4	4
5	5

Table 12.8.6: Ranked data with correlation of 0.90

X	Y
1	2
2	1
3	3
4	4
5	5

The next step is to calculate the number of possible arrangements of Y. The number is simply $N!$, where N is the number of pairs of scores. Here, the number of arrangements is $5! = 120$. Therefore, the probability value is $5/120 = 0.042$. Note that this is a one-tailed probability since it is the proportion of arrangements that give a correlation as large or larger. The two-tailed probability is 0.084.

Since it is hard to count up all the possibilities when the sample size is even moderately large, it is convenient to have a table of critical values.

Table of critical values for Spearman's ρ

From the table linked to above, you can see that the critical value for a one-tailed test with 5 observations at the 0.05 level is 0.90. Since the correlation for the sample data is 0.90, the association is significant at the 0.05 level (one-tailed). As shown above, the probability value is 0.042. Since the critical value for a two-tailed test is 1.0, Spearman's ρ is not significant in a two-tailed test.

This page titled 12.8: Rank Randomization for Association is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.8: Rank Randomization for Association** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.9: Statistical Literacy Standard

Learning Objectives

- Troponin Concentration and Ventricular Strain

Cardiac troponins are markers of myocardial damage. The levels of troponin in subjects with and without signs of right ventricular strain in the electrocardiogram were compared in the experiment described here.

The Wilcoxon rank sum test was used to test for significance. The troponin concentration in patients with signs of right ventricular strain was higher ($median = 0.03$ ng/ml) than in patients without right ventricular strain ($median < 0.01$ ng/ml), $p < 0.001$.

Example 12.9.1: what do you think?

Why might the authors have used the Wilcoxon test rather than a t test? Do you think the conclusions would have been different?

Solution

Perhaps the distributions were very non-normal. Typically a transformation can be done to make a distribution more normal but that is not always the case. It is almost certain the same conclusion would have been reached, although it would have been described in terms of mean differences instead of median differences.

This page titled 12.9: Statistical Literacy Standard is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.9: Statistical Literacy Standard** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

12.10: Wilcoxon Signed-Rank Test

Learning Objectives

- To use the Wilcoxon signed-rank test when you'd like to use the paired ***t***–test, but the differences are severely non-normally distributed.

When to use it

Use the Wilcoxon signed-rank test when there are two nominal variables and one measurement variable. One of the nominal variables has only two values, such as "before" and "after," and the other nominal variable often represents individuals. This is the non-parametric analogue to the paired *t*–test, and you should use it if the distribution of differences between pairs is severely non-normally distributed.

For example, Laureysens et al. (2004) measured metal content in the wood of **13** popular clones growing in a polluted area, once in August and once in November. Concentrations of aluminum (in micrograms of Al per gram of wood) are shown below.

Clone	August	November	August−November
Columbia River	18.3	12.7	-5.6
Fritzi Pauley	13.3	11.1	-2.2
Hazendans	16.5	15.3	-1.2
Primo	12.6	12.7	0.1
Raspalje	9.5	10.5	1.0
Hoogvorst	13.6	15.6	2.0
Balsam Spire	8.1	11.2	3.1
Gibecq	8.9	14.2	5.3
Beaupre	10.0	16.3	6.3
Unal	8.3	15.5	7.2
Trichobel	7.9	19.9	12.0
Gaver	8.1	20.4	12.3
Wolterson	13.4	36.8	23.4

There are two nominal variables: time of year (August or November) and poplar clone (Columbia River, Fritzi Pauley, etc.), and one measurement variable (micrograms of aluminum per gram of wood). The differences are somewhat skewed; the Wolterson clone, in particular, has a much larger difference than any other clone. To be safe, the authors analyzed the data using a Wilcoxon signed-rank test, and I'll use it as the example.

Null hypothesis

The null hypothesis is that the median difference between pairs of observations is zero. Note that this is different from the null hypothesis of the paired ***t***–test, which is that the *mean* difference between pairs is zero, or the null hypothesis of the sign test, which is that the numbers of differences in each direction are equal.

How the test works

Rank the absolute value of the differences between observations from smallest to largest, with the smallest difference getting a rank of **1**, then next larger difference getting a rank of **2**, etc. Give average ranks to ties. Add the ranks of all differences in one direction,

then add the ranks of all differences in the other direction. The smaller of these two sums is the test statistic, W (sometimes symbolized T_s). Unlike most test statistics, *smaller* values of W are less likely under the null hypothesis. For the aluminum in wood example, the median change from August to November (**3.1** micrograms Al/g wood) is significantly different from zero ($W = 16,\ P = 0.040$).

Example

Buchwalder and Huber-Eicher (2004) wanted to know whether turkeys would be less aggressive towards unfamiliar individuals if they were housed in larger pens. They tested **10** groups of three turkeys that had been reared together, introducing an unfamiliar turkey and then counting the number of times it was pecked during the test period. Each group of turkeys was tested in a small pen and in a large pen. There are two nominal variables, size of pen (small or large) and the group of turkeys, and one measurement variable (number of pecks per test). The median difference between the number of pecks per test in the small pen vs. the large pen was significantly greater than zero ($W = 10,\ P = 0.04$).



Fig. 4.12.1 A turkey's head.

Ho et al. (2004) inserted a plastic implant into the soft palate of **12** chronic snorers to see if it would reduce the volume of snoring. Snoring loudness was judged by the sleeping partner of the snorer on a subjective **10**-point scale. There are two nominal variables, time (before the operations or after the operation) and individual snorer, and one measurement variable (loudness of snoring). One person left the study, and the implant fell out of the palate in two people; in the remaining nine people, the median change in snoring volume was significantly different from zero ($W = 0,\ P = 0.008$).

Graphing the results

You should graph the data for a Wilcoxon signed rank test the same way you would graph the data for a paired *t*–test, a bar graph with either the values side-by-side for each pair, or the differences at each pair.

Similar tests

You can analyze paired observations of a measurement variable using a paired *t*–test, if the null hypothesis is that the mean difference between pairs of observations is zero and the differences are normally distributed. If you have a large number of paired observations, you can plot a histogram of the differences to see if they look normally distributed. The paired *t*–test isn't very sensitive to non-normal data, so the deviation from normality has to be pretty dramatic to make the paired *t*–test inappropriate.

Use the sign test when the null hypothesis is that there are equal number of differences in each direction, and you don't care about the size of the differences.

How to do the test

Spreadsheet

I have prepared a spreadsheet to do the Wilcoxon signed-rank test signedrank.xls. It will handle up to **1000** pairs of observations.

Web pages

There is a web page that will perform the Wilcoxon signed-rank test. You may enter your paired numbers directly onto the web page; it will be easier if you enter them into a spreadsheet first, then copy them and paste them into the web page.

R

Salvatore Mangiafico's ***R*** *Companion* has a sample R program for the Wilcoxon signed rank test.

SAS

To do Wilcoxon signed-rank test in SAS, you first create a new variable that is the difference between the two observations. You then run PROC UNIVARIATE on the difference, which automatically does the Wilcoxon signed-rank test along with several others. Here's an example using the poplar data from above:

```
DATA POPLARS;
INPUT clone $ augal noval;
diff=augal - noval;
DATALINES;
Balsam_Spire 8.1 11.2
Beaupre 10.0 16.3
Hazendans 16.5 15.3
Hoogvorst 13.6 15.6
Raspalje 9.5 10.5
Unal 8.3 15.5
Columbia_River 18.3 12.7
Fritzi_Pauley 13.3 11.1
Trichobel 7.9 19.9
Gaver 8.1 20.4
Gibecq 8.9 14.2
Primo 12.6 12.7
Wolterson 13.4 36.8
;

PROC UNIVARIATE DATA=poplars;
VAR diff;
RUN;
```

PROC UNIVARIATE returns a bunch of descriptive statistics that you don't need; the result of the Wilcoxon signed-rank test is shown in the row labeled "Signed rank":

Tests for Location: Mu0=0

Test -Statistic- -----p Value------

Student's t t -2.3089 Pr > |t| 0.0396
Sign M -3.5 Pr >= |M| 0.0923
Signed Rank S -29.5 Pr >= |S| 0.0398

References

1. Picture of a turkey's head from Ohio State University 4-H Poultry.
2. Buchwalder, T., and B. Huber-Eicher. 2004. Effect of increased floor space on aggressive behaviour in male turkeys (*Melagris gallopavo*). Applied Animal Behavior Science 89: 207-214.
3. Ho, W.K., W.I. Wei, and K.F. Chung. 2004. Managing disturbing snoring with palatal implants: a pilot study. Archives of Otolaryngology Head and Neck Surgery 130: 753-758.
4. Laureysens, I., R. Blust, L. De Temmerman, C. Lemmens and R. Ceulemans. 2004. Clonal variation in heavy metal accumulation and biomass production in a poplar coppice culture. I. Seasonal variation in leaf, wood and bark concentrations. Environmental Pollution 131: 485-494.

This page titled 12.10: Wilcoxon Signed-Rank Test is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **4.12: Wilcoxon Signed-Rank Test** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

12.11: Kruskal–Wallis Test

Learning Objectives

- To learn to use the Kruskal–Wallis test when you have one nominal variable and one ranked variable. It tests whether the mean ranks are the same in all the groups.

When to use it

The most common use of the Kruskal–Wallis test is when you have one nominal variable and one measurement variable, an experiment that you would usually analyze using one-way anova, but the measurement variable does not meet the normality assumption of a one-way anova. Some people have the attitude that unless you have a large sample size and can clearly demonstrate that your data are normal, you should routinely use Kruskal–Wallis; they think it is dangerous to use one-way anova, which assumes normality, when you don't know for sure that your data are normal. However, one-way anova is not very sensitive to deviations from normality. I've done simulations with a variety of non-normal distributions, including flat, highly peaked, highly skewed, and bimodal, and the proportion of false positives is always around 5% or a little lower, just as it should be. For this reason, I don't recommend the Kruskal-Wallis test as an alternative to one-way anova. Because many people use it, you should be familiar with it even if I convince you that it's overused.

The Kruskal-Wallis test is a non-parametric test, which means that it does not assume that the data come from a distribution that can be completely described by two parameters, mean and standard deviation (the way a normal distribution can). Like most non-parametric tests, you perform it on ranked data, so you convert the measurement observations to their ranks in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on. You lose information when you substitute ranks for the original values, which can make this a somewhat less powerful test than a one-way anova; this is another reason to prefer one-way anova.

The other assumption of one-way anova is that the variation within the groups is equal (homoscedasticity). While Kruskal-Wallis does not assume that the data are normal, it does assume that the different groups have the same distribution, and groups with different standard deviations have different distributions. If your data are heteroscedastic, Kruskal–Wallis is no better than one-way anova, and may be worse. Instead, you should use Welch's anova for heteoscedastic data.

The only time I recommend using Kruskal-Wallis is when your original data set actually consists of one nominal variable and one ranked variable; in this case, you cannot do a one-way anova and must use the Kruskal–Wallis test. Dominance hierarchies (in behavioral biology) and developmental stages are the only ranked variables I can think of that are common in biology.

The Mann–Whitney U-test (also known as the Mann–Whitney–Wilcoxon test, the Wilcoxon rank-sum test, or the Wilcoxon two-sample test) is limited to nominal variables with only two values; it is the non-parametric analogue to two-sample *t*–test. It uses a different test statistic (U instead of the H of the Kruskal–Wallis test), but the P value is mathematically identical to that of a Kruskal–Wallis test. For simplicity, I will only refer to Kruskal–Wallis on the rest of this web page, but everything also applies to the Mann–Whitney U-test.

The Kruskal–Wallis test is sometimes called Kruskal–Wallis one-way anova or non-parametric one-way anova. I think calling the Kruskal–Wallis test an anova is confusing, and I recommend that you just call it the Kruskal–Wallis test.

Null hypothesis

The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same. The expected mean rank depends only on the total number of observations (for n observations, the expected mean rank in each group is $(\frac{n+1}{2})$, so it is not a very useful description of the data; it's not something you would plot on a graph.

You will sometimes see the null hypothesis of the Kruskal–Wallis test given as "The samples come from populations with the same distribution." This is correct, in that if the samples come from populations with the same distribution, the Kruskal–Wallis test will show no difference among them. I think it's a little misleading, however, because only some kinds of differences in distribution will be detected by the test. For example, if two populations have symmetrical distributions with the same center, but one is much wider than the other, their distributions are different but the Kruskal–Wallis test will not detect any difference between them.

The null hypothesis of the Kruskal–Wallis test is *not* that the means are the same. It is therefore incorrect to say something like "The mean concentration of fructose is higher in pears than in apples (Kruskal–Wallis test, $P = 0.02$)," although you will see data summarized with means and then compared with Kruskal–Wallis tests in many publications. The common misunderstanding of the null hypothesis of Kruskal-Wallis is yet another reason I don't like it.

The null hypothesis of the Kruskal–Wallis test is often said to be that the medians of the groups are equal, but this is only true if you assume that the shape of the distribution in each group is the same. If the distributions are different, the Kruskal–Wallis test can reject the null hypothesis even though the medians are the same. To illustrate this point, I made up these three sets of numbers. They have identical means (**43.5**), and identical medians (**27.5**), but the mean ranks are different (**34.6**, **27.5** and **20.4**, respectively), resulting in a significant ($P = 0.025$) Kruskal–Wallis test:

Group 1	Group 2	Group 3
1	10	19
2	11	20
3	12	21
4	13	22
5	14	23
6	15	24
7	16	25
8	17	26
9	18	27
46	37	28
47	58	65
48	59	66
49	60	67
50	61	68
51	62	69
52	63	70
53	64	71
342	193	72

How the test works

Here are some data on Wright's F_{ST} (a measure of the amount of geographic variation in a genetic polymorphism) in two populations of the American oyster, *Crassostrea virginica*. McDonald et al. (1996) collected data on F_{ST} for six anonymous DNA polymorphisms (variation in random bits of DNA of no known function) and compared the F_{ST} values of the six DNA polymorphisms to F_{ST} values on **13** proteins from Buroker (1983). The biological question was whether protein polymorphisms would have generally lower or higher F_{ST} values than anonymous DNA polymorphisms. McDonald et al. (1996) knew that the theoretical distribution of F_{ST} for two populations is highly skewed, so they analyzed the data with a Kruskal–Wallis test.

When working with a measurement variable, the Kruskal–Wallis test starts by substituting the rank in the overall data set for each measurement value. The smallest value gets a rank of **1**, the second-smallest gets a rank of **2**, etc. Tied observations get average ranks; in this data set, the two F_{ST} values of -0.005 are tied for second and third, so they get a rank of **2.5**.

gene	class	FST	Rank	Rank
CVJ5	DNA	-0.006	1	
CVB1	DNA	-0.005	2.5	
6Pgd	protein	-0.005		2.5
Pgi	protein	-0.002		4
CVL3	DNA	0.003	5	
Est-3	protein	0.004		6
Lap-2	protein	0.006		7
Pgm-1	protein	0.015		8
Aat-2	protein	0.016		9.5
Adk-1	protein	0.016		9.5
Sdh	protein	0.024		11
Acp-3	protein	0.041		12
Pgm-2	protein	0.044		13
Lap-1	protein	0.049		14
CVL1	DNA	0.053	15	
Mpi-2	protein	0.058		16
Ap-1	protein	0.066		17
CVJ6	DNA	0.095	18	
CVB2m	DNA	0.116	19	
Est-1	protein	0.163		20

You calculate the sum of the ranks for each group, then the test statistic, H. H is given by a rather formidable formula that basically represents the variance of the ranks among groups, with an adjustment for the number of ties. H is approximately chi-square distributed, meaning that the probability of getting a particular value of H by chance, if the null hypothesis is true, is the P value corresponding to a chi-square equal to H; the degrees of freedom is the number of groups minus 1. For the example data, the mean rank for DNA is 10.08 and the mean rank for protein is 10.68, $H = 0.043$, there is 1 degree of freedom, and the P value is 0.84. The null hypothesis that the F_{ST} of DNA and protein polymorphisms have the same mean ranks is not rejected.

For the reasons given above, I think it would actually be better to analyze the oyster data with one-way anova. It gives a P value of 0.75, which fortunately would not change the conclusions of McDonald et al. (1996).

If the sample sizes are too small, H does not follow a chi-squared distribution very well, and the results of the test should be used with caution. N less than 5 in each group seems to be the accepted definition of "too small."

Assumptions

The Kruskal–Wallis test does NOT assume that the data are normally distributed; that is its big advantage. If you're using it to test whether the medians are different, it does assume that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (one is skewed to the right and another is skewed to the left, for example, or they have different variances), the Kruskal–Wallis test may give inaccurate results (Fagerland and Sandvik 2009). If

you're interested in any difference among the groups that would make the mean ranks be different, then the Kruskal–Wallis test doesn't make any assumptions.

Heteroscedasticity is one way in which different groups can have different shaped distributions. If the distributions are heteroscedastic, the Kruskal–Wallis test won't help you; instead, you should use Welch's *t*–test for two groups, or Welch's anova for more than two groups.

Example



Fig. 4.8.1 Bluespotted salamander (Ambystoma laterale).

Bolek and Coggins (2003) collected multiple individuals of the toad *Bufo americanus*,, the frog *Rana pipiens*, and the salamander *Ambystoma laterale* from a small area of Wisconsin. They dissected the amphibians and counted the number of parasitic helminth worms in each individual. There is one measurement variable (worms per individual amphibian) and one nominal variable (species of amphibian), and the authors did not think the data fit the assumptions of an anova. The results of a Kruskal–Wallis test were significant ($H = 63.48$, $2 d.f.$, $P = 1.6 \times 10^{-14}$); the mean ranks of worms per individual are significantly different among the three species.

Dog	Sex	Rank
Merlino	Male	1
Gastone	Male	2
Pippo	Male	3
Leon	Male	4
Golia	Male	5
Lancillotto	Male	6
Mamy	Female	7
Nanà	Female	8
Isotta	Female	9
Diana	Female	10
Simba	Male	11
Pongo	Male	12
Semola	Male	13
Kimba	Male	14
Morgana	Female	15
Stella	Female	16
Hansel	Male	17
Cucciola	Male	18
Mammolo	Male	19
Dotto	Male	20
Gongolo	Male	21
Gretel	Female	22
Brontolo	Female	23
Eolo	Female	24
Mag	Female	25
Emy	Female	26
Pisola	Female	27

Cafazzo et al. (2010) observed a group of free-ranging domestic dogs in the outskirts of Rome. Based on the direction of 1815 observations of submissive behavior, they were able to place the dogs in a dominance hierarchy, from most dominant (Merlino) to most submissive (Pisola). Because this is a true ranked variable, it is necessary to use the Kruskal–Wallis test. The mean rank for males (11.1) is lower than the mean rank for females (17.7), and the difference is significant ($H = 4.61$, $1 d.f.$, $P = 0.032$).

Graphing the results

It is tricky to know how to visually display the results of a Kruskal–Wallis test. It would be misleading to plot the means or medians on a bar graph, as the Kruskal–Wallis test is not a test of the difference in means or medians. If there are relatively small number of observations, you could put the individual observations on a bar graph, with the value of the measurement variable on the Y axis and its rank on the X axis, and use a different pattern for each value of the nominal variable. Here's an example using the oyster F_{ST} data:

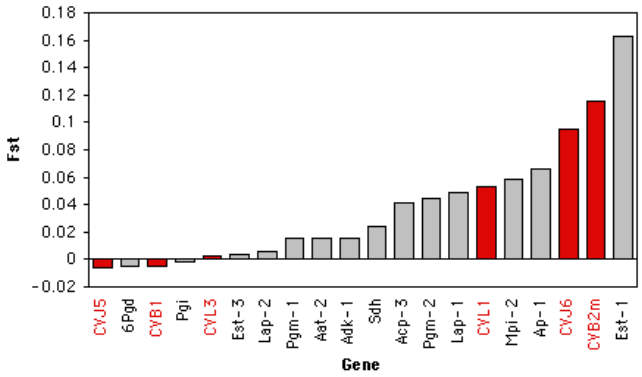


Fig. 4.8.2 F_{st} values for DNA and protein polymorphisms in the American oyster. DNA polymorphisms are shown in red.

If there are larger numbers of observations, you could plot a histogram for each category, all with the same scale, and align them vertically. I don't have suitable data for this handy, so here's an illustration with imaginary data:

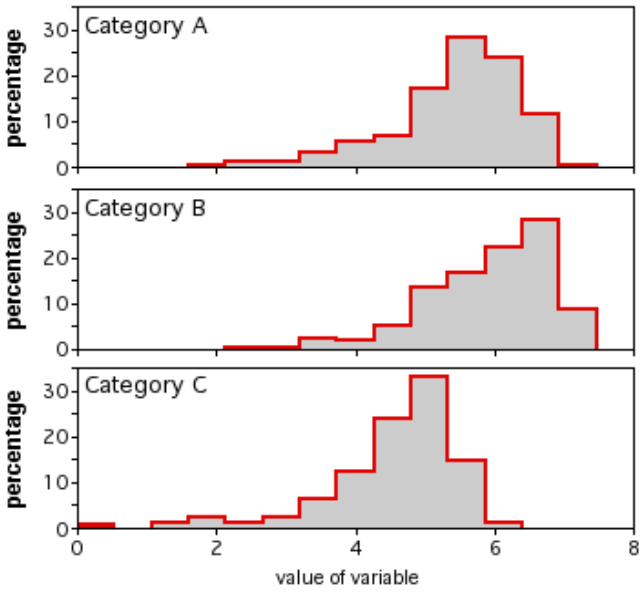


Fig. 4.8.3 Histograms of three sets of numbers.

Similar tests

One-way anova is more powerful and a lot easier to understand than the Kruskal–Wallis test, so unless you have a true ranked variable, you should use it.

How to do the test

Spreadsheet

I have put together a spreadsheet to do the Kruskal–Wallis test kruskalwallis.xls on up to 20 groups, with up to 1000 observations per group.

Web pages

Richard Lowry has web pages for performing the Kruskal–Wallis test for two groups, three groups, or four groups.

R

Salvatore Mangiafico's *R Companion* has a sample R program for the Kruskal–Wallis test.

SAS

To do a Kruskal–Wallis test in SAS, use the NPAR1WAY procedure (that's the numeral "one," not the letter "el," in NPAR1WAY). **WILCOXON** tells the procedure to only do the Kruskal–Wallis test; if you leave that out, you'll get several other statistical tests as well, tempting you to pick the one whose results you like the best. The nominal variable that gives the group names is given with the CLASS parameter, while the measurement or ranked variable is given with the VAR parameter. Here's an example, using the oyster data from above:

```
DATA oysters;
INPUT markername $ markertype $ fst;
DATALINES;
CVB1 DNA -0.005
CVB2m DNA 0.116
CVJ5 DNA -0.006
CVJ6 DNA 0.095
CVL1 DNA 0.053
CVL3 DNA 0.003
6Pgd protein -0.005
Aat-2 protein 0.016
Acp-3 protein 0.041
Adk-1 protein 0.016
Ap-1 protein 0.066
Est-1 protein 0.163
Est-3 protein 0.004
Lap-1 protein 0.049
Lap-2 protein 0.006
Mpi-2 protein 0.058
Pgi protein -0.002
Pgm-1 protein 0.015
Pgm-2 protein 0.044
Sdh protein 0.024
;
PROC NPAR1WAY DATA=oysters WILCOXON;
CLASS markertype;
VAR fst;
RUN;
```

The output contains a table of "Wilcoxon scores"; the "mean score" is the mean rank in each group, which is what you're testing the homogeneity of. "Chi-square" is the H-statistic of the Kruskal–Wallis test, which is approximately chi-square distributed. The "Pr > Chi-Square" is your P value. You would report these results as "$H = 0.04$, $1 d.f.$, $P = 0.84$."

Wilcoxon Scores (Rank Sums) for Variable fst classified by Variable markertype

```
Sum of Expected Std Dev Mean
markertype N Scores Under H0 Under H0 Score
-----------------------------------------------------------------
```

DNA 6 60.50 63.0 12.115236 10.083333
protein 14 149.50 147.0 12.115236 10.678571

Kruskal–Wallis Test

Chi-Square 0.0426
DF 1
Pr > Chi-Square 0.8365

Power analysis

I am not aware of a technique for estimating the sample size needed for a Kruskal–Wallis test.

References

1. Picture of a salamander from Cortland Herpetology Connection.
2. Bolek, M.G., and J.R. Coggins. 2003. Helminth community structure of sympatric eastern American toad, *Bufo americanus americanus*, northern leopard frog, *Rana pipiens*, and blue-spotted salamander, *Ambystoma laterale*, from southeastern Wisconsin. Journal of Parasitology 89: 673-680.
3. Buroker, N. E. 1983. Population genetics of the American oyster *Crassostrea virginica* along the Atlantic coast and the Gulf of Mexico. Marine Biology 75:99-112.
4. Cafazzo, S., P. Valsecchi, R. Bonanni, and E. Natoli. 2010. Dominance in relation to age, sex, and competitive contexts in a group of free-ranging domestic dogs. Behavioral Ecology 21: 443-455.
5. Fagerland, M.W., and L. Sandvik. 2009. The Wilcoxon-Mann-Whitney test under scrutiny. Statistics in Medicine 28: 1487-1497.
6. McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica*. Molecular Biology and Evolution 13: 1114-1118.

This page titled 12.11: Kruskal–Wallis Test is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **4.8: Kruskal–Wallis Test** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

12.12: Spearman Rank Correlation

Learning Objectives

- To use Spearman rank correlation to test the association between two ranked variables, or one ranked variable and one measurement variable. You can also use Spearman rank correlation instead of linear regression/correlation for two measurement variables if you're worried about non-normality, but this is not usually necessary.

When to use it

Use Spearman rank correlation when you have two ranked variables, and you want to see whether the two variables covary; whether, as one variable increases, the other variable tends to increase or decrease. You also use Spearman rank correlation if you have one measurement variable and one ranked variable; in this case, you convert the measurement variable to ranks and use Spearman rank correlation on the two sets of ranks.

For example, Melfi and Poyser (2007) observed the behavior of **6** male colobus monkeys (*Colobus guereza*) in a zoo. By seeing which monkeys pushed other monkeys out of their way, they were able to rank the monkeys in a dominance hierarchy, from most dominant to least dominant. This is a ranked variable; while the researchers know that Erroll is dominant over Milo because Erroll pushes Milo out of his way, and Milo is dominant over Fraiser, they don't know whether the difference in dominance between Erroll and Milo is larger or smaller than the difference in dominance between Milo and Fraiser. After determining the dominance rankings, Melfi and Poyser (2007) counted eggs of *Trichuris* nematodes per gram of monkey feces, a measurement variable. They wanted to know whether social dominance was associated with the number of nematode eggs, so they converted eggs per gram of feces to ranks and used Spearman rank correlation.

Monkey name	Dominance rank	Eggs per gram	Eggs per gram (rank)
Erroll	1	5777	1
Milo	2	4225	2
Fraiser	3	2674	3
Fergus	4	1249	4
Kabul	5	749	6
Hope	6	870	5

Some people use Spearman rank correlation as a non-parametric alternative to linear regression and correlation when they have two measurement variables and one or both of them may not be normally distributed; this requires converting both measurements to ranks. Linear regression and correlation that the data are normally distributed, while Spearman rank correlation does not make this assumption, so people think that Spearman correlation is better. In fact, numerous simulation studies have shown that linear regression and correlation are not sensitive to non-normality; one or both measurement variables can be very non-normal, and the probability of a false positive ($P < 0.05$, when the null hypothesis is true) is still about **0.05** (Edgell and Noon 1984, and references therein). It's not incorrect to use Spearman rank correlation for two measurement variables, but linear regression and correlation are much more commonly used and are familiar to more people, so I recommend using linear regression and correlation any time you have two measurement variables, even if they look non-normal.

Null hypothesis

The null hypothesis is that the Spearman correlation coefficient, ρ ("rho"), is **0**. A ρ of **0** means that the ranks of one variable do not covary with the ranks of the other variable; in other words, as the ranks of one variable increase, the ranks of the other variable do not increase (or decrease).

Assumption

When you use Spearman rank correlation on one or two measurement variables converted to ranks, it does not assume that the measurements are normal or homoscedastic. It also doesn't assume the relationship is linear; you can use Spearman rank correlation even if the association between the variables is curved, as long as the underlying relationship is monotonic (as X gets larger, Y keeps getting larger, or keeps getting smaller). If you have a non-monotonic relationship (as X gets larger, Y gets larger and then gets smaller, or Y gets smaller and then gets larger, or something more complicated), you shouldn't use Spearman rank correlation.

Like linear regression and correlation, Spearman rank correlation assumes that the observations are independent.

How the test works

Spearman rank correlation calculates the P value the same way as linear regression and correlation, except that you do it on ranks, not measurements. To convert a measurement variable to ranks, make the largest value 1, second largest 2, etc. Use the average ranks for ties; for example, if two observations are tied for the second-highest rank, give them a rank of 2.5 (the average of 2 and 3).

When you use linear regression and correlation on the ranks, the Pearson correlation coefficient (r) is now the Spearman correlation coefficient, ρ, and you can use it as a measure of the strength of the association. For 11 or more observations, you calculate the test statistic using the same equation as for linear regression and correlation, substituting ρ for r: $t_s = \frac{\sqrt{d.f.} \times \rho^2}{\sqrt{(1-\rho^2)}}$. If the null hypothesis (that $\rho = 0$) is true, t_s is t-distributed with $n-2$ degrees of freedom.

If you have 10 or fewer observations, the P value calculated from the t-distribution is somewhat inaccurate. In that case, you should look up the P value in a table of Spearman t-statistics for your sample size. My Spearman spreadsheet does this for you.

You will almost never use a regression line for either description or prediction when you do Spearman rank correlation, so don't calculate the equivalent of a regression line.

For the Colobus monkey example, Spearman's ρ is 0.943, and the P value from the table is less than 0.025, so the association between social dominance and nematode eggs is significant.

Example



Fig. 5.2.1 Magnificent frigatebird, Fregata magnificens.

Volume (cm^3)	Frequency (Hz)
1760	529
2040	566
2440	473
2550	461
2730	465

2740	532
3010	484
3080	527
3370	488
3740	485
4910	478
5090	434
5090	468
5380	449
5850	425
6730	389
6990	421
7960	416

Males of the magnificent frigatebird (*Fregata magnificens*) have a large red throat pouch. They visually display this pouch and use it to make a drumming sound when seeking mates. Madsen et al. (2004) wanted to know whether females, who presumably choose mates based on their pouch size, could use the pitch of the drumming sound as an indicator of pouch size. The authors estimated the volume of the pouch and the fundamental frequency of the drumming sound in 18 males.

There are two measurement variables, pouch size and pitch. The authors analyzed the data using Spearman rank correlation, which converts the measurement variables to ranks, and the relationship between the variables is significant (Spearman's $\rho = -0.76,\ 16d.f.,\ P = 0.0002$). The authors do not explain why they used Spearman rank correlation; if they had used regular correlation, they would have obtained $r = -0.82,\ P = 0.00003$.

Graphing the results

You can graph Spearman rank correlation data the same way you would for a linear regression or correlation. Don't put a regression line on the graph, however; it would be misleading to put a linear regression line on a graph when you've analyzed it with rank correlation.

How to do the test

Spreadsheet

I've put together a spreadsheet that will perform a Spearman rank correlation spearman.xls on up to 1000 observations. With small numbers of observations (10 or fewer), the spreadsheet looks up the P value in a table of critical values.

Web page

This web page will do Spearman rank correlation.

R

Salvatore Mangiafico's *R Companion* has a sample R program for Spearman rank correlation.

SAS

Use PROC CORR with the SPEARMAN option to do Spearman rank correlation. Here is an example using the bird data from the correlation and regression web page:

PROC CORR DATA=birds SPEARMAN;
VAR species latitude;
RUN;
The results include the Spearman correlation coefficient ρ, analogous to the *r* value of a regular correlation, and the *P* value:

Spearman Correlation Coefficients, $N = 17$
Prob > |r| under H0: Rho=0

species latitude

species 1.00000 -0.36263 **Spearman correlation coefficient**
0.1526 ***P* value**

latitude -0.36263 1.00000
0.1526

References

1. Picture of magnificent frigatebird from CalPhoto, by Lloyd Glenn Ingles, © California Academy of Sciences.
2. Edgell, S.E., and S.M. Noon. 1984. Effect of violation of normality on the *t*–test of the correlation coefficient. Psychological Bulletin 95: 576-583.
3. Madsen, V., T.J.S. Balsby, T. Dabelsteen, and J.L. Osorno. 2004. Bimodal signaling of a sexually selected trait: gular pouch drumming in the magnificent frigatebird. Condor 106: 156-160.
4. Melfi, V., and F. Poyser. 2007. *Trichuris* burdens in zoo-housed *Colobus guereza*. International Journal of Primatology 28: 1449-1456.

This page titled 12.12: Spearman Rank Correlation is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **5.2: Spearman Rank Correlation** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

12.13: Choosing the Right Test

Learning Objectives

- This table is designed to help you decide which statistical test or descriptive statistic is appropriate for your experiment. In order to use it, you must be able to identify all the variables in the data set and tell what kind of variables they are.

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Exact test for goodness-of-fit	1	–	–	test fit of observed frequencies to expected frequencies	use for small sample sizes (less than 1000)	count the number of red, pink and white flowers in a genetic cross, test fit to expected $1:2:1$ ratio, total sample < 1000
Chi-square test of goodness-of-fit	1	–	–	test fit of observed frequencies to expected frequencies	use for large sample sizes (greater than 1000)	count the number of red, pink and white flowers in a genetic cross, test fit to expected $1:2:1$ ratio, total sample > 1000
G–test of goodness-of-fit	1	–	–	test fit of observed frequencies to expected frequencies	used for large sample sizes (greater than 1000)	count the number of red, pink and white flowers in a genetic cross, test fit to expected $1:2:1$ ratio, total sample > 1000
Repeated *G*–tests of goodness-of-fit	2	–	–	test fit of observed frequencies to expected frequencies in multiple experiments	-	count the number of red, pink and white flowers in a genetic cross, test fit to expected $1:2:1$ ratio, do multiple crosses

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Fisher's exact test	2	–	–	test hypothesis that proportions are the same in different groups	use for small sample sizes (less than 1000)	count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample < 1000
Chi-square test of independence	2	–	–	test hypothesis that proportions are the same in different groups	use for large sample sizes (greater than 1000)	count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample > 1000
G–test of independence	2	–	–	test hypothesis that proportions are the same in different groups	large sample sizes (greater than 1000)	count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is the same in the two treatments, total sample > 1000
Cochran-Mantel-Haenszel test	3	–	–	test hypothesis that proportions are the same in repeated pairings of two groups	alternate hypothesis is a consistent direction of difference	count the number of live and dead patients after treatment with drug or placebo, test the hypothesis that the proportion of live and dead is

						the same in the two treatments, repeat this experiment at different hospitals
test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Arithmetic mean	–	**1**	–	description of central tendency of data	-	-
Median	–	**1**	–	description of central tendency of data	more useful than mean for very skewed data	median height of trees in forest, if most trees are short seedlings and the mean would be skewed by a few very tall trees
Range	–	**1**	–	description of dispersion of data	used more in everyday life than in scientific statistics	-
Variance	–	**1**	–	description of dispersion of data	forms the basis of many statistical tests; in squared units, so not very understandable	-
Standard deviation	–	**1**	–	description of dispersion of data	in same units as original data, so more understandable than variance	-
Standard error of the mean	–	**1**	–	description of accuracy of an estimate of a mean	-	-
Confidence interval	–	**1**	–	description of accuracy of an estimate of a mean	-	-

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
One-sample *t*–test	–	1	–	test the hypothesis that the mean value of the measurement variable equals a theoretical expectation	-	blindfold people, ask them to hold arm at 45° angle, see if mean angle is equal to 45°
Two-sample *t*–test	1	1	–	test the hypothesis that the mean values of the measurement variable are the same in two groups	just another name for one-way anova when there are only two groups	compare mean heavy metal content in mussels from Nova Scotia and New Jersey
One-way anova	1	1	–	test the hypothesis that the mean values of the measurement variable are the same in different groups	-	compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey
Tukey-Kramer test	1	1	–	after a significant one-way anova, test for significant differences between all pairs of groups	-	compare mean heavy metal content in mussels from Nova Scotia vs. Maine, Nova Scotia vs. Massachusetts, Maine vs. Massachusetts, etc.
Bartlett's test	1	1	–	test the hypothesis that the standard deviation of a measurement variable is the same in different groups	usually used to see whether data fit one of the assumptions of an anova	compare standard deviation of heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut,

						New York and New Jersey
test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Nested anova	2+	1	–	test hypothesis that the mean values of the measurement variable are the same in different groups, when each group is divided into subgroups	subgroups must be arbitrary (model II)	compare mean heavy metal content in mussels from Nova Scotia, Maine, Massachusetts, Connecticut, New York and New Jersey; several mussels from each location, with several metal measurements from each mussel
Two-way anova	2	1	–	test the hypothesis that different groups, classified two ways, have the same means of the measurement variable	-	compare cholesterol levels in blood of male vegetarians, female vegetarians, male carnivores, and female carnivores
Paired *t*–test	2	1	–	test the hypothesis that the means of the continuous variable are the same in paired data	just another name for two-way anova when one nominal variable represents pairs of observations	compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet
Wilcoxon signed-rank test	2	1	–	test the hypothesis that the means of the measurement variable are the same in paired data	used when the differences of pairs are severely non-normal	compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet, when differences are non-normal

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Linear regression	–	**2**	–	see whether variation in an independent variable causes some of the variation in a dependent variable; estimate the value of one unmeasured variable corresponding to a measured variable	-	measure chirping speed in crickets at different temperatures, test whether variation in temperature causes variation in chirping speed; or use the estimated relationship to estimate temperature from chirping speed when no thermometer is available
Correlation	–	**2**	–	see whether two variables covary	-	measure salt intake and fat intake in different people's diets, to see if people who eat a lot of fat also eat a lot of salt
Polynomial regression	–	**2**	–	test the hypothesis that an equation with X^2, X^3, etc. fits the Y variable significantly better than a linear regression	-	-
Analysis of covariance (ancova)	**1**	**2**	–	test the hypothesis that different groups have the same regression lines	first test the homogeneity of slopes; if they are not significantly different, test the homogeneity of the Y-intercepts	measure chirping speed vs. temperature in four species of crickets, see if there is significant variation among the species in the slope or Y-intercept of the relationships

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Multiple regression	–	3+	–	fit an equation relating several X variables to a single Y variable	-	measure air temperature, humidity, body mass, leg length, see how they relate to chirping speed in crickets
Simple logistic regression	1	1	–	fit an equation relating an independent measurement variable to the probability of a value of a dependent nominal variable	-	give different doses of a drug (the measurement variable), record who lives or dies in the next year (the nominal variable)
Multiple logistic regression	1	2+	–	fit an equation relating more than one independent measurement variable to the probability of a value of a dependent nominal variable	-	record height, weight, blood pressure, age of multiple people, see who lives or dies in the next year

test	nominal variables	measurement variables	ranked variables	purpose	notes	example
Sign test	2	–	1	test randomness of direction of difference in paired data	-	compare the cholesterol level in blood of people before vs. after switching to a vegetarian diet, only record whether it is higher or lower after the switch
Kruskal–Wallis test	1	–	1	test the hypothesis that rankings are the same in different groups	often used as a non-parametric alternative to one-way anova	40 ears of corn (8 from each of 5 varieties) are ranked for tastiness, and the mean rank is compared among varieties
Spearman rank correlation	–	–	2	see whether the ranks of two variables covary	often used as a non-parametric alternative to regression or correlation	40 ears of corn are ranked for tastiness and prettiness, see whether prettier corn is also tastier

This page titled 12.13: Choosing the Right Test is shared under a not declared license and was authored, remixed, and/or curated by John H. McDonald via source content that was edited to the style and standards of the LibreTexts platform.

- **7.5: Choosing the Right Test** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

12.E: Distribution Free Tests (Exercises)

General Questions

Q1

For the following data, how many ways could the data be arranged (including the original arrangement) so that the advantage of the Experimental Group mean over the Control Group mean is as large or larger then the original arrangement?

Experimental	Control
5	1
10	2
15	3
16	4
17	9

Q2

For the data in Problem 1, how many ways can the data be rearranged?

Q3

What is the one-tailed probability for a test of the difference?

Q4

For the following data, how many ways can the data be rearranged?

T1	T2	Control
7	14	0
8	19	2
11	21	5

Q5

In general, are rank randomization tests or randomization tests more powerful?

Q6

What is the advantage of rank randomization tests over randomization tests?

Q7

Test whether the differences among conditions for the data in Problem 1 is significant (one tailed) at the 0.01 level using a rank randomization test.

Questions from Case Studies

The following question uses data from the SAT and GPA case study.

Q8

Compute Spearman's ρ for the relationship between UGPA and SAT.

The following question uses data from the Stereograms case study.

Q9

Test the difference in central tendency between the two conditions using a rank-randomization test (with the normal approximation) with a one-tailed test. Give the Z and the p.

The following question uses data from the Smiles and Leniency case study.

Q10

Test the difference in central tendency between the four conditions using a rank-randomization test (with the normal approximation). Give the Chi Square and the p.

This page titled 12.E: Distribution Free Tests (Exercises) is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane via source content that was edited to the style and standards of the LibreTexts platform.

- **18.10: Distribution Free Tests (Exercises)** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.

CHAPTER OVERVIEW

13: Appendices

13.1: A | Statistical Table- Standard Normal (Z)

13.2: A | Statistical Table- Student t Distribution

13.3: A | Statistical Table- Chi-Square Distribution

13.4: A | Statistical Table- F Distribution

13.5: B | Mathematical Phrases, Symbols, and Formulas

13: Appendices is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

13.1: A | Statistical Table- Standard Normal (Z)

Standard Normal Probability Distribution

Numerical entries represent the probability that a standard normal random variable is between **0** and z where $z = \frac{x-\mu}{\sigma}$.



Figure A2

https://stats.libretexts.org/@go/page/11914

Standard Normal Probability Distribution: Z Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Table A11 Standard Normal Distribution

This page titled 13.1: A | Statistical Table- Standard Normal (Z) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

13.2: A | Statistical Table- Student t Distribution

Student's t Distribution



Figure A3 **Upper critical values of Student's t Distribution with v Degrees of Freedom**

For selected probabilities, a, the table shows the values $t_{v,a}$ such that $P(t_v > t_{v,a}) = a$, where t_v is a Student's t random variable with v degrees of freedom. For example, the probability is .10 that a Student's t random variable with 10 degrees of freedom exceeds 1.372.

df	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706*	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
100	1.290	1.660	1.984	2.364	2.626	3.174

v	0.10	0.05	0.025	0.01	0.005	0.001
∞	1.282	1.645	1.960	2.326	2.576	3.090

This page titled 13.2: A | Statistical Table- Student t Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

13.3: A | Statistical Table- Chi-Square Distribution

χ^2 Probability Distribution

Table A12 Probability of Exceeding the Critical Value



Figure A4

https://stats.libretexts.org/@go/page/13025

χ^2 Probability Distribution

df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490

df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table A13 Area to the Right of the Critical Value of χ^2

This page titled 13.3: A | Statistical Table- Chi-Square Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

13.4: A | Statistical Table- F Distribution

F Distribution

Figure A1 Table entry for p is the critical value F^* with probability p lying to its right.

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	p	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33

Table A1 F critical values

	Degrees of freedom in the numerator										
Degrees of freedom in the denominator	10	12	15	20	25	30	40	50	60	120	1000
1	.100	60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
	.050	241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19
	.025	968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7
	.010	6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
	.001	605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
2	.100	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
	.050	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
	.025	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
	.010	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
	.001	999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
3	.100	5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
	.050	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
	.025	14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
	.010	27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
	.001	129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
4	.100	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
	.050	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
	.025	8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26
	.010	14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
	.001	48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
5	.100	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
	.050	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
	.025	6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
	.010	10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
	.001	26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
6	.100	2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
	.050	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
	.025	5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
	.010	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89

https://stats.libretexts.org/@go/page/13026

	Degrees of freedom in the numerator										
.001	18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
.100	2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
.050	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
7025	4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
.010	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
.001	14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72

Table A2 ***F*** critical values (continued)

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	***p***	1	2	3	4	5	6	7	8	9
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26

Table A3 ***F*** critical values (continued)

		Degrees of freedom in the numerator										
Degrees of freedom in the denominator	***p***	10	12	15	20	25	30	40	50	60	120	1000
8	.100	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30
	.050	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
	.025	4.30	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68
	.010	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87
	.001	11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36
9	.100	2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
	.050	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
	.025	3.96	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34
	.010	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
	.001	9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
10	.100	2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
	.050	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
	.025	3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09
	.010	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
	.001	8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
11	.100	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98
	.050	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41
	.025	3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89
	.010	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61
	.001	7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
12	.100	2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
	.050	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
	.025	3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
	.010	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
	.001	7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44
13	.100	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
	.050	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
	.025	3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
	.010	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
	.001	6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99
14	.100	2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
	.050	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
	.025	3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
	.010	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
	.001	6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
15	.100	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	.050	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
	.025	3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
	.010	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
	.001	6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33

Table A4 ***F*** critical values (continued)

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	***p***	1	2	3	4	5	6	7	8	9
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89

Table A5 ***F*** critical values (continued)

		Degrees of freedom in the numerator										
Degrees of freedom in the denominator	***p***	10	12	15	20	25	30	40	50	60	120	1000
16	.100	2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
	.050	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
	.025	2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
	.010	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
	.001	5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08
17	.100	2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
	.050	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
	.025	2.92	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26
	.010	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
	.001	5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87
18	.100	1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
	.050	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
	.025	2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
	.010	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
	.001	5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
19	.100	1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
	.050	2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
	.025	2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
	.010	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
	.001	5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
20	.100	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
	.050	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
	.025	2.77	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
	.010	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
	.001	5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
21	.100	1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
	.050	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
	.025	2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
	.010	3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
	.001	4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
22	.100	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
	.050	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
	.025	2.70	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
	.010	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
	.001	4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
23	.100	1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
	.050	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
	.025	2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
	.010	3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
	.001	4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08

Table A6 *F* critical values (continued)

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	***p***	1	2	3	4	5	6	7	8	9
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02

Table A7 ***F*** critical values (continued)

		Degrees of freedom in the numerator										
Degrees of freedom in the denominator	***p***	10	12	15	20	25	30	40	50	60	120	1000
24	.100	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
	.050	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
	.025	2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
	.010	3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
	.001	4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99
25	.100	1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
	.050	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
	.025	2.61	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
	.010	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
	.001	4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
26	.100	1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
	.050	2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
	.025	2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
	.010	3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
	.001	4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
27	.100	1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
	.050	2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
	.025	2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
	.010	3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
	.001	4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78
28	.100	1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
	.050	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
	.025	2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
	.010	3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
	.001	4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
29	.100	1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
	.050	2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
	.025	2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
	.010	3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
	.001	4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
30	.100	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
	.050	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
	.025	2.51	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
	.010	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
	.001	4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
40	.100	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
	.050	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
	.025	2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
	.010	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
	.001	3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25

Table A8 ***F*** critical values (continued)

		Degrees of freedom in the numerator								
Degrees of freedom in the denominator	***p***	1	2	3	4	5	6	7	8	9
50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

Table A9 ***F*** critical values (continued)

		Degrees of freedom in the numerator										
Degrees of freedom in the denominator	***p***	10	12	15	20	25	30	40	50	60	120	1000
50	.100	1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
	.050	2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
	.025	2.32	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
	.010	2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
	.001	3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
60	.100	1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
	.050	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
	.025	2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
	.010	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
	.001	3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
100	.100	1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
	.050	1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
	.025	2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
	.010	2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
	.001	3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
200	.100	1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
	.050	1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
	.025	2.11	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
	.010	2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
	.001	3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1000	.100	1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
	.050	1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
	.025	2.06	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
	.010	2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
	.001	2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

Table A10 *F* critical values (continued)

This page titled 13.4: A | Statistical Table- F Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

SECTION OVERVIEW

13.5: B | Mathematical Phrases, Symbols, and Formulas

Curated and edited by Kristin Kuter | Saint Mary's College, Notre Dame, IN

This page titled 13.5: B | Mathematical Phrases, Symbols, and Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax.

Index

A

Adding probabilities
- 4.3: The Addition and Multiplication Rules of Probability

anova
- 11.4.1: One-Way ANOVA

B

bar graph
- 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

bar graphs
- 2.3: Other Types of Graphs

Bernoulli trial
- 5.4: Binomial Distribution

binomial probability distribution
- 5.4: Binomial Distribution
- 5.4.1: Binomial Distribution Formula
- 7.4: Confidence Intervals and Sample Size for Proportions

blinding
- 1.6: Experimental Design and Ethics

box plots
- 3.5: Exploratory Data Analysis

C

central limit theorem
- 6.5: Normal Approximation to the Binomial Distribution

Chebyshev's Theorem
- 3.3.2: The Empirical Rule and Chebyshev's Theorem

Circular Permutations
- 4.4.2: Permutations with Similar Elements

cluster sample
- 1.6.2: Observational Studies and Sampling Strategies

cluster sampling
- 1.3: Variables and Types of Data

coefficient of determination
- 10.3: The Regression Equation

Combinations
- 4.4.3: Combinations

Comparing two population means
- 9.2: Inferences for Two Population Means- Large, Independent Samples
- 9.3: Inferences for Two Population Means - Unknown Standard Deviations

Comparing Two Population Proportions
- 9.5: Inferences for Two Population Proportions

complement
- 4.1.2: Terminology
- 4.2: Independent and Mutually Exclusive Events

conditional probability
- 4.1.2: Terminology

confidence interval
- 8.1: Steps in Hypothesis Testing

CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS
- 9.5: Inferences for Two Population Proportions

confounding variable
- 1.6.2: Observational Studies and Sampling Strategies

contingency table
- 4.3.1: Contingency Tables
- 11.3.1: Test of Independence

continuous data
- 1.3: Variables and Types of Data

control group
- 1.6: Experimental Design and Ethics

cumulative probability distributions
- 6.1: Introduction

cumulative relative frequency
- 1.3.1: Levels of Measurement
- 2.4: Organizing Data - Frequency Distributions

D

Decision
- 8.1.4: Rare Events, the Sample, Decision and Conclusion

direction of a relationship between the variables
- 10.1.2: Scatter Plots

discrete data
- 1.3: Variables and Types of Data

dot plot
- 2.3.2: Dot Plots

E

empirical rule
- 3.3.2: The Empirical Rule and Chebyshev's Theorem

Equal variance
- 10.2: Testing the Significance of the Correlation Coefficient

ethics
- 1.6: Experimental Design and Ethics

event
- 4.1.2: Terminology

expected value
- 5.3: Mean or Expected Value and Standard Deviation

experimental unit
- 1.6: Experimental Design and Ethics

explanatory variable
- 1.6: Experimental Design and Ethics

extrapolation
- 10.3.1: Prediction

F

F distribution
- 11.4: Prelude to F Distribution and One-Way ANOVA

factorial
- 4.4.1: Permutations
- 5.4.1: Binomial Distribution Formula

Fisher's Exact Test
- 12.5: Fisher's Exact Test

frequency
- 1.3.1: Levels of Measurement
- 2.4: Organizing Data - Frequency Distributions

Frequency Polygons
- 2.2.1: Frequency Polygons and Time Series Graphs

frequency table
- 1.3.1: Levels of Measurement
- 2.4: Organizing Data - Frequency Distributions

G

goodness of fit
- 11.2: Goodness-of-Fit Test

H

histograms
- 2.2.1: Frequency Polygons and Time Series Graphs

homogeneity
- 11.3.2: Test for Homogeneity

Hypothesis Testing
- 8.1: Steps in Hypothesis Testing
- 8.1.1: Null and Alternative Hypotheses
- 8.1.3: Distribution Needed for Hypothesis Testing
- 8.1.5: Additional Information on Hypothesis Tests
- 8.2: Hypothesis Test Examples for Means
- 8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation
- 8.4: Hypothesis Test Examples for Proportions

I

independent events
- 4.2: Independent and Mutually Exclusive Events
- 4.3: The Addition and Multiplication Rules of Probability
- 11.3.1: Test of Independence

inferential statistics
- 7.1: Confidence Intervals

Institutional Review Board
- 1.6: Experimental Design and Ethics

interpolation
- 10.3.1: Prediction

K

Kruskal–Wallis Test
- 12.11: Kruskal–Wallis Test

L

law of large numbers
- 6.5: Normal Approximation to the Binomial Distribution

level of measurement
- 1.3.1: Levels of Measurement
- 2.4: Organizing Data - Frequency Distributions

line graph
- 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

linear correlation coefficient
- 10.2: Testing the Significance of the Correlation Coefficient
- 10.3: The Regression Equation

linear equations
- 10.1.1: Review- Linear Equations

LINEAR REGRESSION MODEL
- 10.3: The Regression Equation

lurking variable
- 1.6: Experimental Design and Ethics

M

margin of error
- 7.2: Confidence Intervals for the Mean with Known Standard Deviation

mean
- 3.2.1: Skewness and the Mean, Median, and Mode
- 5.3: Mean or Expected Value and Standard Deviation

median
- 3.2: Measures of the Center of the Data
- 3.2.1: Skewness and the Mean, Median, and Mode
- 3.4: Measures of Position

mode
- 3.2: Measures of the Center of the Data
- 3.2.1: Skewness and the Mean, Median, and Mode

multiplication rule
- 4.5: Probability And Counting Rules

Multiplying probabilities
- 4.3: The Addition and Multiplication Rules of Probability

mutually exclusive
- 4.2: Independent and Mutually Exclusive Events
- 4.3: The Addition and Multiplication Rules of Probability

N

Normal Approximation to the Binomial Distribution
- 5.4.1: Binomial Distribution Formula
- 6.5: Normal Approximation to the Binomial Distribution

normal distribution
- 6.3: Applications of the Normal Distribution
- 6.4: The Central Limit Theorem

O

outcome
- 4.1.2: Terminology

outliers
- 3.4: Measures of Position
- 10.4: Outliers

P

paired difference samples
- 9.4: Inferences for Two Population Means - Paired Samples

paired samples
- 9.4: Inferences for Two Population Means - Paired Samples

parameter
- 1.2: Descriptive and Inferential Statistics
- 4.1: Sample Spaces and Probability

Pareto chart
- 1.3: Variables and Types of Data

Pareto charts
- 2.3: Other Types of Graphs

permutation
- 4.4.1: Permutations

pie charts
- 2.3: Other Types of Graphs

placebo
- 1.4: Data Collection and Sampling Techniques
- 1.6: Experimental Design and Ethics

pooled variance
- 9.3: Inferences for Two Population Means - Unknown Standard Deviations
- 11.4.2: The F Distribution and the F-Ratio

population
- 1.2: Descriptive and Inferential Statistics
- 4.1: Sample Spaces and Probability

population mean
- 3.2: Measures of the Center of the Data

Population Standard Deviation
- 3.3: Measures of Variation

power of the test
- 8.1.2: Outcomes and the Type I and Type II Errors
- 8.1.5: Additional Information on Hypothesis Tests
- 8.2: Hypothesis Test Examples for Means
- 8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation
- 8.4: Hypothesis Test Examples for Proportions

prediction
- 10.3.1: Prediction

probability
- 1.2: Descriptive and Inferential Statistics
- 4.1: Sample Spaces and Probability

probability distribution function
- 5.2: Probability Distribution Function (PDF) for a Discrete Random Variable
- 6.3: Applications of the Normal Distribution

prospective study
- 1.6.2: Observational Studies and Sampling Strategies

Q

qualitative data
- 1.3: Variables and Types of Data

quantitative data
- 1.3: Variables and Types of Data

quartiles
- 3.4: Measures of Position

R

random assignment
- 1.6: Experimental Design and Ethics

Randomization Association
- 12.4: Randomization Association

Ranked variables
- 12.12: Spearman Rank Correlation

rare events
- 8.1.4: Rare Events, the Sample, Decision and Conclusion

response variable
- 1.6: Experimental Design and Ethics

Retrospective studies
- 1.6.2: Observational Studies and Sampling Strategies

rounding
- 1.3.1: Levels of Measurement
- 2.4: Organizing Data - Frequency Distributions

S

sample mean
- 3.2: Measures of the Center of the Data

sample space
- 4.1.2: Terminology

sample standard deviation
- 3.3: Measures of Variation

sampling
- 1: The Nature of Statistics

sampling bias
- 1.3: Variables and Types of Data

sampling distribution of the mean
- 6.4: The Central Limit Theorem

Sampling Error
- 1.3: Variables and Types of Data

sampling with replacement
- 1.3: Variables and Types of Data
- 4.2: Independent and Mutually Exclusive Events
- 4.3.2: Tree and Venn Diagrams

sampling without replacement
- 1.3: Variables and Types of Data
- 4.2: Independent and Mutually Exclusive Events
- 4.3.2: Tree and Venn Diagrams

scatter plot
- 10.1.2: Scatter Plots

significance level
- 8.1.4: Rare Events, the Sample, Decision and Conclusion

simple random sampling
- 1.6.2: Observational Studies and Sampling Strategies

Skewed
- 3.2.1: Skewness and the Mean, Median, and Mode
- 3.5: Exploratory Data Analysis

slope
- 10.1.1: Review- Linear Equations

Spearman Rank Correlation
- 12.12: Spearman Rank Correlation

standard deviation
- 3.3: Measures of Variation
- 5.3: Mean or Expected Value and Standard Deviation

standard error of the mean
- 6.4: The Central Limit Theorem

standard normal distribution
- 6.2: The Normal Distribution
- 6.2.1: The Standard Normal Distribution

statistic
- 1.2: Descriptive and Inferential Statistics
- 4.1: Sample Spaces and Probability

stemplot
- 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

stratified sampling
- 1.6.2: Observational Studies and Sampling Strategies

strength of a relationship between the variables
- 10.1.2: Scatter Plots

T

test for homogeneity
- 11.3.2: Test for Homogeneity

The alternative hypothesis
- 8.1.1: Null and Alternative Hypotheses

The AND Event
- 4.1.2: Terminology

The null hypothesis
- 8.1.1: Null and Alternative Hypotheses

The Or Event
- 4.1.2: Terminology

The OR of Two Events
4.2: Independent and Mutually Exclusive Events
Time Series Graphs
2.2.1: Frequency Polygons and Time Series Graphs
treatments
1.6: Experimental Design and Ethics
tree diagram
4.3.2: Tree and Venn Diagrams
tree diagrams
4.5: Probability And Counting Rules
Type I error
8.1.2: Outcomes and the Type I and Type II Errors
Type II error
8.1.2: Outcomes and the Type I and Type II Errors

V

variable
1.2: Descriptive and Inferential Statistics
4.1: Sample Spaces and Probability
variation due to error or unexplained variation
11.4.2: The F Distribution and the F-Ratio
variation due to treatment or explained variation
11.4.2: The F Distribution and the F-Ratio
Venn diagram
4.3.2: Tree and Venn Diagrams

W

Wilcoxon Rank Sum test
12.6: Rank Randomization Two Conditions

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: STAT C1000 (LPC): Statistics and Probability

Webpages: 167

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 70.7% (118 pages)
- Undeclared: 14.4% (24 pages)
- Public Domain: 7.2% (12 pages)
- CC BY-NC-SA 4.0: 3.6% (6 pages)
- CC BY-SA 3.0: 2.4% (4 pages)
- CC BY-SA 4.0: 1.8% (3 pages)

By Page

- STAT C1000 (LPC): Statistics and Probability - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: The Nature of Statistics - *CC BY 4.0*
    - 1.1: Introduction - *CC BY 4.0*
    - 1.2: Descriptive and Inferential Statistics - *CC BY 4.0*
    - 1.3: Variables and Types of Data - *CC BY 4.0*
      - 1.3.1: Levels of Measurement - *CC BY 4.0*
    - 1.4: Data Collection and Sampling Techniques - *CC BY-SA 3.0*
    - 1.5: Computers and Calculators - *Undeclared*
      - 1.5.1: Using Spreadsheets for Statistics - *Undeclared*
    - 1.6: Experimental Design and Ethics - *CC BY 4.0*
      - 1.6.1: More on Experiments - *CC BY-SA 3.0*
      - 1.6.2: Observational Studies and Sampling Strategies - *CC BY-SA 3.0*
    - 1.E: Sampling and Data (Exercises) - *CC BY 4.0*
  - 2: Frequency Distributions and Graphs - *CC BY 4.0*
    - 2.1: Prelude to Graphs - *CC BY 4.0*
    - 2.2: Histograms, Ogives, and Frequency Polygons - *CC BY-SA 4.0*
      - 2.2.1: Frequency Polygons and Time Series Graphs - *CC BY 4.0*
    - 2.3: Other Types of Graphs - *CC BY-SA 4.0*
      - 2.3.1: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs - *CC BY 4.0*
      - 2.3.2: Dot Plots - *Public Domain*
      - 2.3.3: Guide to Fairly Good Graphs - *Undeclared*
      - 2.3.4: Presenting Data in Tables - *Undeclared*
    - 2.4: Organizing Data - Frequency Distributions - *CC BY 4.0*
    - 2.E: Graphs (Optional Exercises) - *CC BY 4.0*
  - 3: Data Description - *CC BY 4.0*
    - 3.1: Prelude to Descriptive Statistics - *CC BY 4.0*
    - 3.2: Measures of the Center of the Data - *CC BY 4.0*
      - 3.2.1: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
    - 3.3: Measures of Variation - *CC BY 4.0*
      - 3.3.1: Coefficient of Variation - *Undeclared*
      - 3.3.2: The Empirical Rule and Chebyshev's Theorem - *CC BY-NC-SA 4.0*
    - 3.4: Measures of Position - *CC BY 4.0*
      - 3.4.1: Measures of Location- Deciles - *Undeclared*
      - 3.4.2: Z-scores - *CC BY-NC-SA 4.0*
    - 3.5: Exploratory Data Analysis - *CC BY 4.0*
    - 3.E: Descriptive Statistics (Optional Exercises) - *CC BY 4.0*
      - 1.E: Measures of Position (Optional Exercises) - *CC BY 4.0*
  - 4: Probability and Counting - *CC BY 4.0*
    - 4.1: Sample Spaces and Probability - *CC BY 4.0*
      - 4.1.1: Introduction to Probability - *CC BY 4.0*
      - 4.1.2: Terminology - *CC BY 4.0*

- 4.2: Independent and Mutually Exclusive Events - *CC BY 4.0*
- 4.3: The Addition and Multiplication Rules of Probability - *CC BY 4.0*
  - 4.3.1: Contingency Tables - *CC BY 4.0*
  - 4.3.2: Tree and Venn Diagrams - *CC BY 4.0*
- 4.4: Counting Rules - *CC BY 4.0*
  - 4.4.1: Permutations - *CC BY 4.0*
  - 4.4.2: Permutations with Similar Elements - *CC BY 4.0*
  - 4.4.3: Combinations - *CC BY 4.0*
- 4.5: Probability And Counting Rules - *CC BY 4.0*
- 4.E: Probability Topics (Optional Exercises) - *CC BY 4.0*
  - 1.E: Combinations (Optional Exercises) - *CC BY 4.0*
  - 1.E: Permutations (Optional Exercises) - *CC BY 4.0*
  - 1.E: Permutations with Similar Elements (Optional Exercises) - *CC BY 4.0*
  - 1.E: Probability Using Tree Diagrams and Combinations (Optional Exercises) - *CC BY 4.0*
  - 1.E: Tree Diagrams and the Multiplication Axiom (Optional Exercises) - *CC BY 4.0*

5: Discrete Probability Distributions - *CC BY 4.0*

- 5.1: Prelude to Discrete Random Variables - *CC BY 4.0*
- 5.2: Probability Distribution Function (PDF) for a Discrete Random Variable - *CC BY 4.0*
- 5.3: Mean or Expected Value and Standard Deviation - *CC BY 4.0*
- 5.4: Binomial Distribution - *CC BY 4.0*
  - 5.4.1: Binomial Distribution Formula - *CC BY-SA 3.0*
- 5.E: Discrete Random Variables (Optional Exercises) - *CC BY 4.0*

6: Continuous Random Variables and the Normal Distribution - *CC BY 4.0*

- 6.1: Introduction - *CC BY 4.0*
  - 6.1.1: Continuous Probability Functions - *CC BY 4.0*
  - 6.1.2: The Uniform Distribution - *CC BY 4.0*
- 6.2: The Normal Distribution - *CC BY 4.0*
  - 6.2.1: The Standard Normal Distribution - *CC BY 4.0*
- 6.3: Applications of the Normal Distribution - *CC BY 4.0*
- 6.4: The Central Limit Theorem - *CC BY 4.0*
- 6.5: Normal Approximation to the Binomial Distribution - *CC BY 4.0*
- 6.E: The Normal Distribution (Optional Exercises) - *CC BY 4.0*
  - 1.E: The Central Limit Theorem for Sample Means (Optional Exercises) - *CC BY 4.0*
  - 1.E: The Standard Normal Distribution (Optional Exercises) - *CC BY 4.0*

7: Confidence Intervals and Sample Size - *CC BY 4.0*

- 7.1: Confidence Intervals - *CC BY 4.0*
- 7.2: Confidence Intervals for the Mean with Known Standard Deviation - *CC BY 4.0*
- 7.3: Confidence Intervals for the Mean with Unknown Standard Deviation - *CC BY 4.0*
- 7.4: Confidence Intervals and Sample Size for Proportions - *CC BY 4.0*
- 7.5: Confidence Intervals (Summary) - *CC BY 4.0*
- 7.E: Confidence Intervals (Optional Exercises) - *CC BY 4.0*
  - 1.E: Confidence Intervals for the Mean with Known Standard Deviation (Optional Exercises) - *CC BY 4.0*

8: Hypothesis Testing with One Sample - *CC BY 4.0*

- 8.1: Steps in Hypothesis Testing - *CC BY 4.0*
  - 8.1.1: Null and Alternative Hypotheses - *CC BY 4.0*
  - 8.1.2: Outcomes and the Type I and Type II Errors - *CC BY 4.0*
  - 8.1.3: Distribution Needed for Hypothesis Testing - *CC BY 4.0*
  - 8.1.4: Rare Events, the Sample, Decision and Conclusion - *CC BY 4.0*
  - 8.1.5: Additional Information on Hypothesis Tests - *CC BY 4.0*
- 8.2: Hypothesis Test Examples for Means - *CC BY 4.0*
- 8.3: Hypothesis Test Examples for Means with Unknown Standard Deviation - *CC BY 4.0*
- 8.4: Hypothesis Test Examples for Proportions - *CC BY 4.0*
- 8.E: Hypothesis Testing (Optional Exercises) - *Undeclared*
  - 1.E: Distribution Needed for Hypothesis Testing (Optional Exercises) - *CC BY 4.0*
  - 1.E: Hypothesis Testing with One Sample (Optional Exercises) - *CC BY 4.0*

- 1.E: Null and Alternative Hypotheses (Optional Exercises) - *CC BY 4.0*
- 1.E: Outcomes and the Type I and Type II Errors (Optional Exercises) - *CC BY 4.0*
- 1.E: Rare Events, the Sample, Decision and Conclusion (Optional Exercises) - *CC BY 4.0*

- 9: Inferences with Two Samples - *CC BY 4.0*
  - 9.1: Prelude to Hypothesis Testing with Two Samples - *CC BY 4.0*
  - 9.2: Inferences for Two Population Means- Large, Independent Samples - *CC BY-NC-SA 4.0*
  - 9.3: Inferences for Two Population Means - Unknown Standard Deviations - *CC BY-NC-SA 4.0*
  - 9.4: Inferences for Two Population Means - Paired Samples - *CC BY-NC-SA 4.0*
  - 9.5: Inferences for Two Population Proportions - *CC BY-NC-SA 4.0*
  - 9.6: Which Analysis Should You Conduct? - *CC BY-SA 4.0*
  - 9.E: Hypothesis Testing with Two Samples (Optional Exercises) - *CC BY 4.0*
- 10: Correlation and Regression - *CC BY 4.0*
  - 10.1: Prelude to Linear Regression and Correlation - *CC BY 4.0*
    - 10.1.1: Review- Linear Equations - *CC BY 4.0*
    - 10.1.2: Scatter Plots - *CC BY 4.0*
  - 10.2: Testing the Significance of the Correlation Coefficient - *CC BY 4.0*
  - 10.3: The Regression Equation - *CC BY 4.0*
    - 10.3.1: Prediction - *CC BY 4.0*
  - 10.4: Outliers - *CC BY 4.0*
  - 10.E: Linear Regression and Correlation (Optional Exercises) - *CC BY 4.0*
    - 1.E: Linear Equations (Optional Exercises) - *CC BY 4.0*
    - 1.E: Outliers (Optional Exercises) - *CC BY 4.0*
    - 1.E: Prediction (Optional Exercises) - *CC BY 4.0*
    - 1.E: Scatter Plots (Optional Exercises) - *CC BY 4.0*
    - 1.E: Testing the Significance of the Correlation Coefficient (Optional Exercises) - *CC BY 4.0*
    - 1.E: The Regression Equation (Optional Exercise) - *CC BY 4.0*
- 11: Chi-Square and Analysis of Variance (ANOVA) - *CC BY 4.0*
  - 11.1: Prelude to The Chi-Square Distribution - *CC BY 4.0*
    - 11.1.1: Facts About the Chi-Square Distribution - *CC BY 4.0*
  - 11.2: Goodness-of-Fit Test - *CC BY 4.0*
  - 11.3: Tests Using Contingency tables - *Undeclared*
    - 11.3.1: Test of Independence - *CC BY 4.0*
    - 11.3.2: Test for Homogeneity - *CC BY 4.0*
    - 11.3.3: Comparison of the Chi-Square Tests - *CC BY 4.0*
  - 11.4: Prelude to F Distribution and One-Way ANOVA - *CC BY 4.0*
    - 11.4.1: One-Way ANOVA - *CC BY 4.0*
    - 11.4.2: The F Distribution and the F-Ratio - *CC BY 4.0*
    - 11.4.3: Facts About the F Distribution - *CC BY 4.0*
    - 11.4.4: How to Use Microsoft Excel® for Regression Analysis - *Undeclared*
  - 11.E: F Distribution and One-Way ANOVA (Optional Exercises) - *CC BY 4.0*
  - 11.E: The Chi-Square Distribution (Optional Exercises) - *CC BY 4.0*
- 12: Nonparametric Statistics - *Public Domain*
  - 12.1: Benefits of Distribution Free Tests - *Public Domain*
  - 12.2: Randomization Tests - Two Conditions - *Public Domain*
  - 12.3: Randomization Tests - Two or More Conditions - *Public Domain*
  - 12.4: Randomization Association - *Public Domain*
  - 12.5: Fisher's Exact Test - *Public Domain*
  - 12.6: Rank Randomization Two Conditions - *Public Domain*
  - 12.7: Rank Randomization Two or More Conditions - *Public Domain*
  - 12.8: Rank Randomization for Association - *Public Domain*
  - 12.9: Statistical Literacy Standard - *Public Domain*
  - 12.10: Wilcoxon Signed-Rank Test - *Undeclared*
  - 12.11: Kruskal–Wallis Test - *Undeclared*
  - 12.12: Spearman Rank Correlation - *Undeclared*
  - 12.13: Choosing the Right Test - *Undeclared*
  - 12.E: Distribution Free Tests (Exercises) - *Public Domain*
- 13: Appendices - *Undeclared*
  - 13.1: A | Statistical Table- Standard Normal (Z) - *CC BY 4.0*
  - 13.2: A | Statistical Table- Student t Distribution - *CC BY 4.0*

- 13.3: A | Statistical Table- Chi-Square Distribution - *CC BY 4.0*
- 13.4: A | Statistical Table- F Distribution - *CC BY 4.0*
- 13.5: B | Mathematical Phrases, Symbols, and Formulas - *CC BY 4.0*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*