# INTRODUCTORY STATISTICS



*Barbara Illowsky & Susan Dean*
De Anza College

LibreTexts™

# De Anza College

# Introductory Statistics

Barbara Illowsky and Susan Dean

- **TitlePage** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

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: Sampling and Data

- 1.1: Introduction to Probability and Statistics
- 1.2: Key Terms and Definitions
- 1.3: Populations and Samples

## 2: Descriptive Statistics

- 2.1: Organizing and Graphing Qualitative Data
- 2.2: Organizing and Graphing Quantitative Data
- 2.3: Stem-and-Leaf Displays
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.5: Measures of Position- Percentiles and Quartiles
- 2.6: Box Plots
- 2.7: Measures of Spread- Variance and Standard Deviation
- 2.8: Skewness and the Mean, Median, and Mode

## 3: Introduction to Linear Regression and Correlation

- 3.1: Linear Equations
- 3.2: Scatter Plots
- 3.3: Simple Linear Regression
- 3.4: Prediction
- 3.5: Outliers

## 4: Probability Theory

- 4.1: Probability Experiments and Sample Spaces
- 4.2: Experiments Having Equally Likely Outcomes
- 4.3: Conditional Probability and Independence
- 4.4: Counting Basics- the Multiplication and Addition Rules
- 4.5: Intersection and Union of Events and Venn Diagrams
- 4.6: Joint and Marginal Probabilities and Contingency Tables
- 4.7: More Counting- Factorials, Combinations, and Permutations

## 5: Discrete Random Variables

- 5.1: Introduction to Random Variables
- 5.2: The Probability Distribution Function
- 5.3: Expectation, Variance and Standard Deviation
- 5.4: The Binomial Distribution
- 5.5: The Geometric Distribution
- 5.6: The Hypergeometric Distribution
- 5.7: The Poisson Distribution

## 6: Continuous Random Variables

- 6.1: Probability Density Functions
- 6.2: The Uniform and Other Simple Continuous Distributions
- 6.3: The Standard Normal Distribution
- 6.4: Applications of Finding Normal Probabilities

## 7: Sampling Distributions

- 7.1: The Sample Mean and Sources of Error
- 7.2: The Sum Distribution

## 8: Confidence Intervals

- 8.1: Estimating Population Means
- 8.2: The t-distribution
- 8.3: Estimating Proportions
- 8.4: Confidence Intervals

## 9: Hypothesis Testing for a Single Variable and Population

- 9.1: Hypothesis Tests- An Introduction
- 9.2: Type I and Type II Errors
- 9.3: Hypothesis Tests about μ- p-value Approach
- 9.4: Hypothesis Tests about μ- Critical Region Approach
- 9.5: Hypothesis Tests for a Proportion

## 10: Hypothesis Testing for Paired and Unpaired Data

- 10.1: Two Population Means
- 10.2: Two Independent Population Proportions
- 10.3: Matched or Paired Samples
- 10.4: Two Population Means with Known Standard Deviations
- 10.5: Difference of Two Means

## 11: Linear Regression and Hypothesis Testing

- 11.1: Testing the Hypothesis that β = 0

## 12: The Chi-Square Distribution

- 12.1: The Chi-Square Distribution
- 12.2: A Goodness-of-Fit Test
- 12.3: A Test of Independence or Homogeneity
- 12.4: Test of a Single Variance
- 12.5: Test for Homogeneity
- 12.6: Comparison of the Chi-Square Tests

## 13: F Distribution and One-Way ANOVA

- 13.1: Prelude to F Distribution and One-Way ANOVA
- 13.2: One-Way ANOVA
- 13.3: The F Distribution and the F-Ratio
- 13.4: Facts About the F Distribution

- 13.5: Test of Two Variances

Index

Glossary

Index

Glossary

Detailed Licensing

# Licensing

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

# CHAPTER OVERVIEW

## 1: Sampling and Data

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 to Probability and Statistics

1.2: Key Terms and Definitions

1.3: Populations and Samples


### 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: Sampling and 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.

# 1.1: Introduction to Probability and Statistics

**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."

## 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 1.1: Introduction to Probability and 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** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 1.2: Key Terms and Definitions

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:

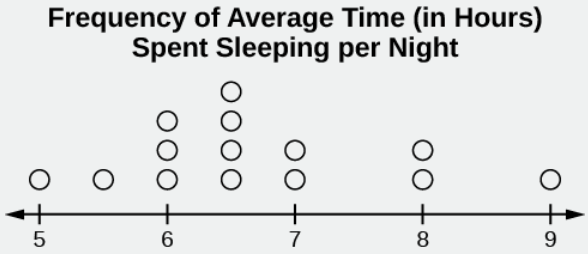



Figure **1.2.1**

- 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 $X$ = 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 $X$ = 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.

## WeBWorK Problems

## References

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

## 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

## 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 1.2: Key Terms and Definitions 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: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 1.3: Populations and Samples

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 $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, $\pi$, $\frac{3\pi}{4}$, 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**.

### 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.

### 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.

## 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.

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.

**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.

## WeBWorK Problems

## 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.

## 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: Populations and 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.

- **1.3: Data, Sampling, and Variation in Data and Sampling** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **1.2: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 2: Descriptive Statistics

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: Organizing and Graphing Qualitative Data
2.2: Organizing and Graphing Quantitative Data
2.3: Stem-and-Leaf Displays
2.4: Measures of Central Tendency- Mean, Median and Mode
2.5: Measures of Position- Percentiles and Quartiles
2.6: Box Plots
2.7: Measures of Spread- Variance and Standard Deviation
2.8: Skewness and the Mean, Median, and Mode


### 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: 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.

# 2.1: Organizing and Graphing Qualitative Data

**Learning Objectives**

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

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- 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.

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, 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.

## 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 2.1.1: 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 2.1.3 and 2.1.4 and determine which graph (pie or bar) you think displays the comparisons better.

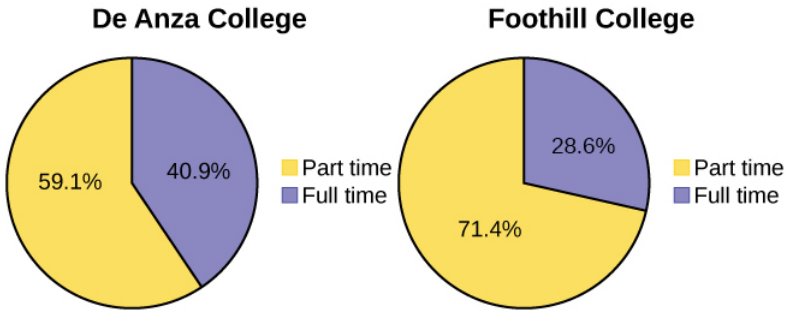



Figure 2.1.3: Pie Charts

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.

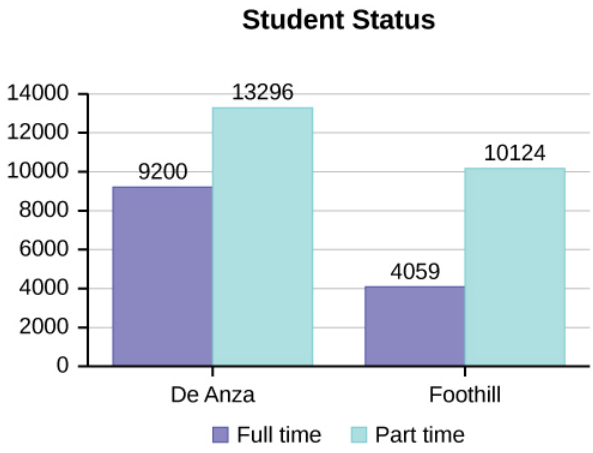



Figure 2.1.4: Bar chart

## 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 **2.1.2**: 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% |

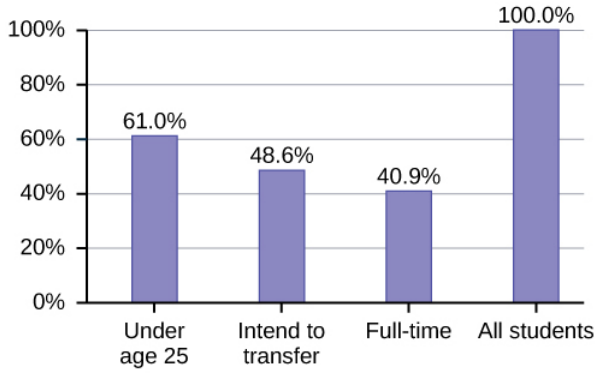



Figure **2.1.2**: Bar chart of data in Table **2.1.2**.

## 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 **2.1.2**: 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% |

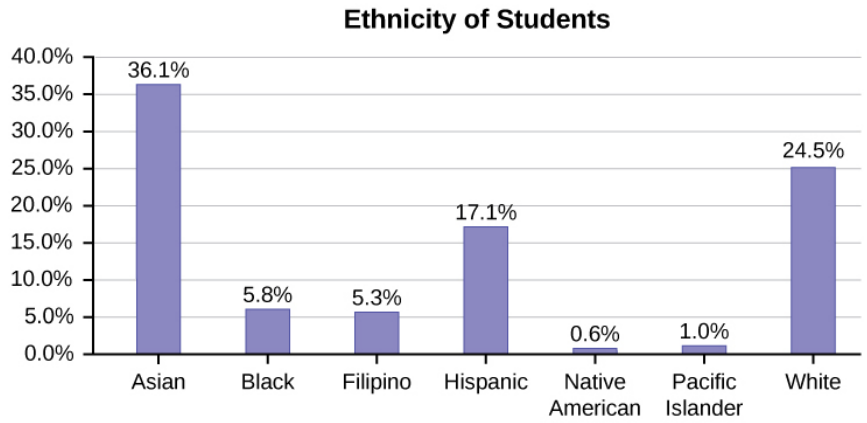



Figure **2.1.3**: Enrollment of De Anza College (Spring 2010)

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 **2.1.4** can be difficult to understand visually. The graph in Figure **2.1.5** is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

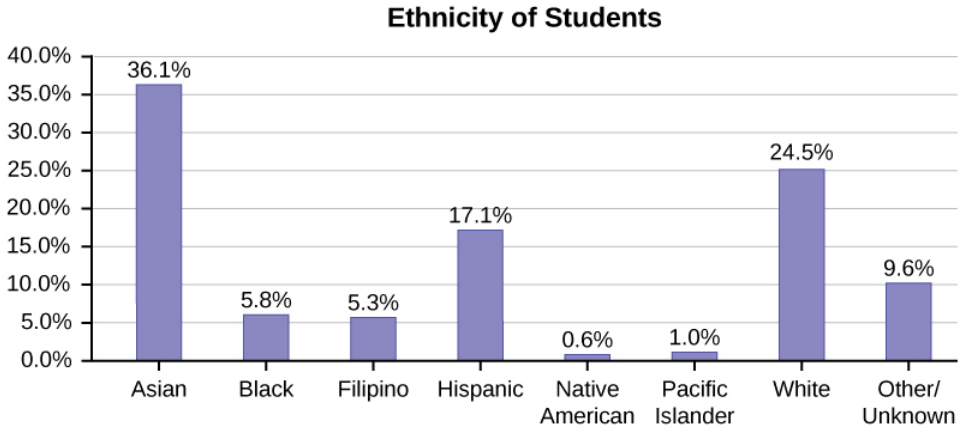



Figure **2.1.4**: Bar Graph with Other/Unknown Category

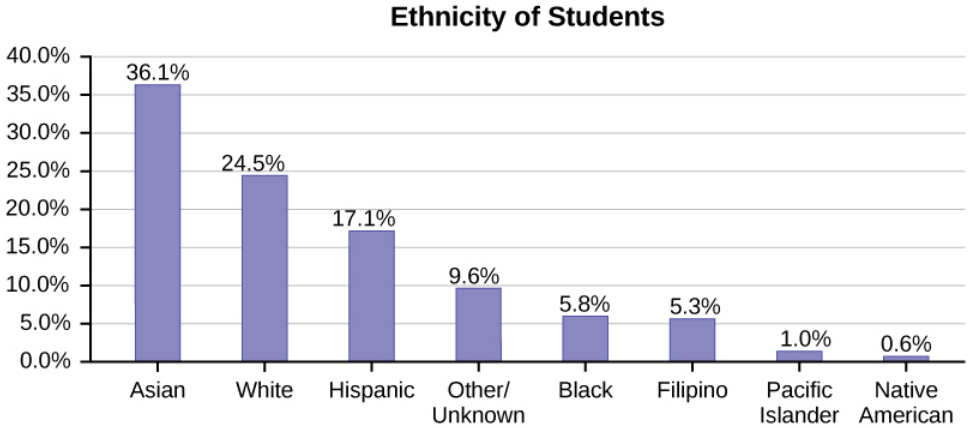



Figure **2.1.5**: Pareto Chart With Bars Sorted by Size

## 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 **2.1.6** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure **2.1.6**.

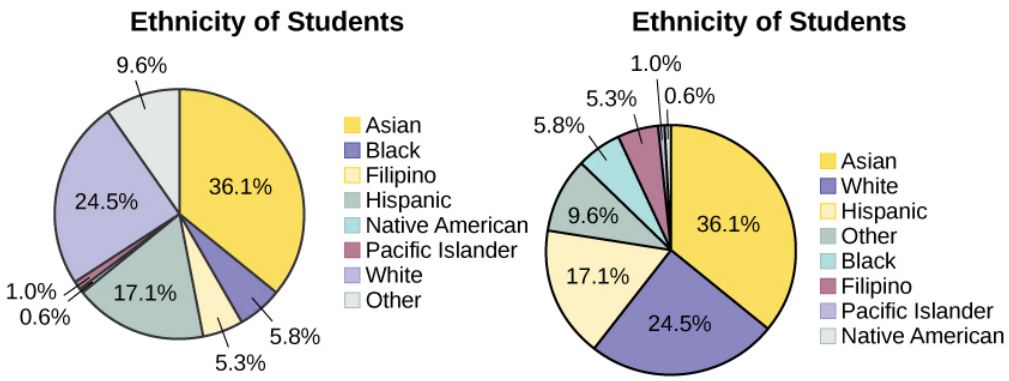



Figure 2.1.6.

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: Organizing and Graphing Qualitative 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.

- **2.1: Prelude to Descriptive Statistics** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **1.3: Data, Sampling, and Variation in Data and Sampling** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 2.2: Organizing and Graphing Quantitative Data

For most of the work you do in this course, you will be working with quantitative data, and you will use a frequency table and frequency histogram to organize and graph the data. An advantage of a frequency table and frequency histogram is that they can be used to organize and display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

## Frequency

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 **2.2.1**: Frequency Table of Student Work Hours

| DATA VALUE | FREQUENCY |
|---|---|
| 2 | 3 |
| 3 | 5 |
| 4 | 3 |
| 5 | 6 |
| 6 | 2 |
| 7 | 1 |

**Definition: Relative Frequency**

A frequency is the number of times a value of the data occurs. According to Table **2.2.1**, 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: 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 **2.2.2**: 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 **2.2.2** 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 **2.2.3**.

Table **2.2.3**: Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

| DATA VALUE | FREQUENCY | RELATIVE FREQUENCY | CUMULATIVE RELATIVE FREQUENCY |
|---|---|---|---|
| 2 | 3 | 320320 or 0.15 | 0.15 |
| 3 | 5 | 520520 or 0.25 | 0.15 + 0.25 = 0.40 |
| 4 | 3 | 320320 or 0.15 | 0.40 + 0.15 = 0.55 |
| 5 | 6 | 620620 or 0.30 | 0.55 + 0.30 = 0.85 |
| 6 | 2 | 220220 or 0.10 | 0.85 + 0.10 = 0.95 |
| 7 | 1 | 120120 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.

Table **2.2.4** represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table **2.2.4**: 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

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.

### Collaborative Exercise 2.2.7

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 2.2.7

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 **2.2.6** was produced:

Table **2.2.6**: 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. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

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 relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.(Remember, frequency is defined as the number of times an answer occurs.) If:

- $f$ is frequency
- $n$ is total number of data values (or the sum of the individual frequencies), and
- $RF$ is relative frequency,

then:

$$RF = \frac{f}{n} \tag{2.2.1}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and RF = fn = 340 = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many bars or intervals, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is $6.05 (6.1 – 0.05 = 6.05)$. We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is $1.495 (1.5 – 0.005 = 1.495)$. If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is $0.9995 (1.0 – 0.0005 = 0.9995)$. If all the data happen to be integers and the smallest value is two, then a convenient starting point is $1.5 (2 – 0.5 = 1.5)$. Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

**Example 2.2.1**

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76 \quad (2.2.2)$$

*We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.*

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.

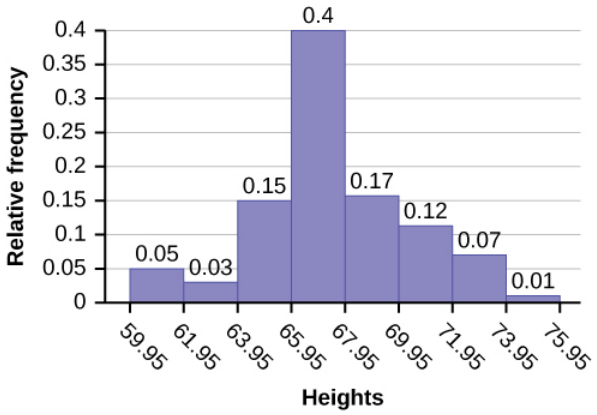



Figure **2.2.1**: Histogram of something

**Example 2.2.2**

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

**Answer**

Calculate the number of bars as follows:

$$\frac{6.5-0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

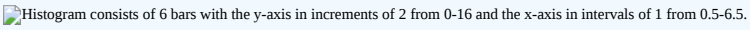



Figure **2.2.2**.

## Example 2.2.3

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends

| | | | | |
|---|---|---|---|---|
| 9.95 | 10 | 2.25 | 16.75 | 0 |
| 19.5 | 22.5 | 7.5 | 15 | 12.75 |
| 5.5 | 11 | 10 | 20.75 | 17.5 |
| 23 | 21.9 | 24 | 23.75 | 18 |
| 20 | 15 | 22.9 | 18.8 | 20.5 |

**Answer**

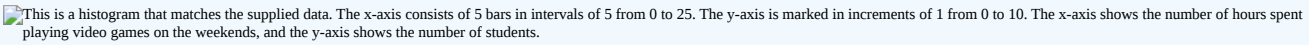



Figure **2.2.3**.

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

### 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.

**Example 2.2.4**

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 |

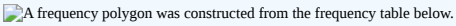



Figure 2.2.4.

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.

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 2.2.6**

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**

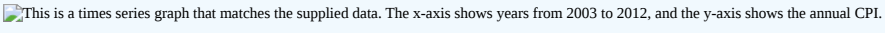



Figure **2.2.7**.

## 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 also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the

frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.Glossary

## WeBWorK Problems

### 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: Organizing and Graphing Quantitative 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.

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

# 2.3: Stem-and-Leaf Displays

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**

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.

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.

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 2.3.7**

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 |

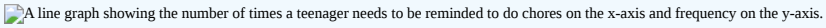

A line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis and frequency on the y-axis.

Figure **2.3.1**

**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 **2.3.9** has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

**Example 2.3.9**

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**

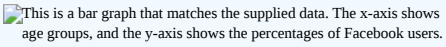

This is a bar graph that matches the supplied data. The x-axis shows age groups, and the y-axis shows the percentages of Facebook users.

Figure **2.3.3**.

**Exercise 2.3.10**

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**

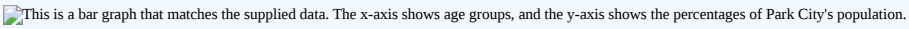



Figure 2.3.4.

## 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.

## WeBWorK Problems

## 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).

## 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.
- CUNY OER WeBWorK Fellows

---

This page titled 2.3: Stem-and-Leaf Displays 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: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 2.4: Measures of Central Tendency- Mean, Median and Mode

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 $\mu$ (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

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{2.4.1}$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \tag{2.4.2}$$

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

$$\frac{n+1}{2} \tag{2.4.3}$$

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

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \tag{2.4.4}$$

The median is the $49^{th}$ value in the ordered data. If the total number of data values is 100, then

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \tag{2.4.5}$$

The median occurs midway between the $50^{th}$ and $51^{st}$ 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 2.4.1**

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:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6 \quad (2.4.6)$$

To find the median, $M$, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \quad (2.4.7)$$

Starting at the smallest value, the median is located between the 20$^{th}$ and 21$^{st}$ 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

$$M = \frac{24+24}{2} = 24 \quad (2.4.8)$$

**Example 2.4.2**

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**

$$\bar{x} = \frac{5,000,000+49(30,000)}{50} = 129,400 \quad (2.4.9)$$

$M = 30,000$

(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.

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 2.4.3**

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.= 7.

**Example 2.4.4**

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.

### 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 $\mu$. 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 $\mu$.

## 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:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{2.4.10}$$

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

$$\frac{\text{lower boundary+upper boundary}}{2}. \tag{2.4.11}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{2.4.12}$$

where $f$ is the frequency of the interval and $m$ is the midpoint of the interval.

Example 2.4.5

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. $\sum fm 53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

WeBWorK Problems


https://stats.libretexts.org/@go/page/26027

## 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

$$\mu = \frac{\sum fm}{\sum f} \tag{2.4.13}$$

where $f$ = interval frequencies and $m$ = interval midpoints.

## 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 $\mu$) 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 2.4: Measures of Central Tendency- 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.

- **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.
- **1.2: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 2.5: Measures of Position- Percentiles and Quartiles

The common measures of location are **quartiles** and **percentiles.** Quartiles are special percentiles. The first quartile, $Q_1$, is the same as the 25$^{th}$ percentile, and the third quartile, $Q_3$, is the same as the 75$^{th}$ percentile. The median, $M$, is called both the second quartile and the 50$^{th}$ 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 90$^{th}$ 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 75$^{th}$ 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.

$$\frac{6.8 + 7.2}{2} = 7 \tag{2.5.1}$$

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**.

**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.

## Example 2.4.3

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 28$^{th}$ 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 28$^{th}$ percentile. They include the two 4s, the five 5s, and the seven 6s. The 28$^{th}$ percentile is between the last six and the first seven. **The 28$^{th}$ percentile is 6.5.**

**Find the median**. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50$^{th}$ 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 50$^{th}$ percentile is between the 25$^{th}$, or seven, and 26$^{th}$, or seven, values. **The median is seven.**

**Find the third quartile**. The third quartile is the same as the 75$^{th}$ 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 75$^{th}$ 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 38$^{th}$ 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.)

## Example 2.4.4

Using Table:

a. Find the 80$^{th}$ percentile.
b. Find the 90$^{th}$ 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.

**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 $i = \frac{k}{100}(n+1)$i=k100(n+1)

If $i$ is an integer, then the $k^{th}$ percentile is the data value in the $i^{th}$ 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
   - $i = theindex$
   - $n = 29$

$i = \frac{k}{100}(n+1) = (\frac{83}{100})(29+1) = 24.9$, 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.

**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.

## 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.)

On a 20 question math test, the $70^{th}$ percentile for number of correct answers was 16. Interpret the $70^{th}$ 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.

**Example 2.4.9**

At a community college, it was found that the $30^{th}$ percentile of credit units that students are enrolled for is seven units. Interpret the $30^{th}$ 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.

**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.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \quad (2.5.2)$$

.

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.

## WeBWorK Problems

## 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 50$^{th}$ 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 25$^{th}$ percentile,the second quartile ($Q_2$ or median) is 50$^{th}$ percentile, and the third quartile ($Q_3$) is the the 75$^{th}$ 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 k$^{th}$ 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 $50^{th}$ percentile. The first and third quartiles are the $25^{th}$ and the $75^{th}$ 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.

## 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 2.5: Measures of Position- Percentiles and Quartiles 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.

# 2.6: Box Plots

*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.

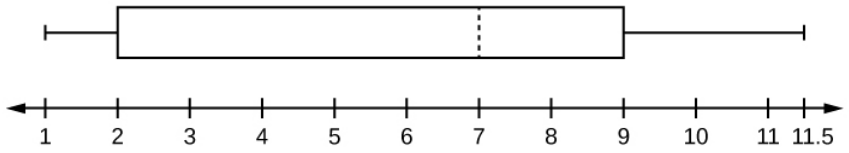



Figure **2.6.1**

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 2.6.1**

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; 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

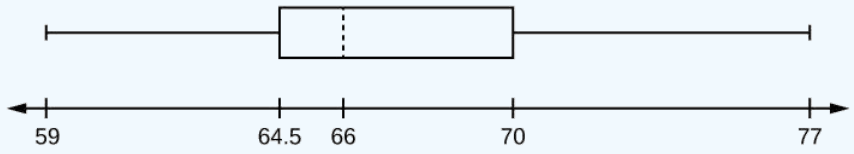



Figure **2.6.2**

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.

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 **2.6.4**

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 2.6.2

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.

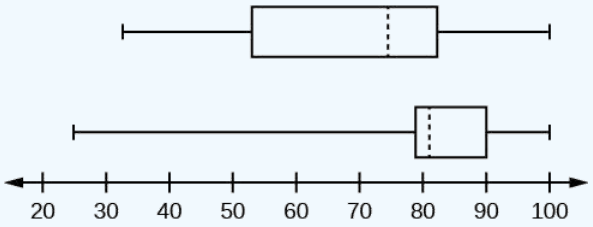



***Figure* 2.6.5**

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.

**Example 2.6.3**

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.

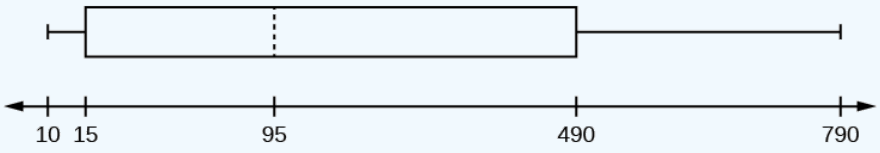



Figure **2.6.7**

## 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.

## WeBWorK Problems

## 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.

## 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 2.6: Box 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.

- **2.5: Box Plots** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **1.2: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 2.7: Measures of Spread- Variance and Standard Deviation

[NOTE from VS: The following is pulled from Shafer and Zhang]

Look at the two data sets in Table 2.7.1 and the graphical representation of each, called a *dot plot*, in Figure 2.7.1.

Table 2.7.1: *Two Data Sets*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Data Set I: | 40 | 38 | 42 | 40 | 39 | 39 | 43 | 40 | 39 | 40 |
| Data Set II: | 46 | 37 | 40 | 33 | 42 | 36 | 40 | 47 | 34 | 45 |

The two sets of ten measurements each center at the same value: they both have mean, median, and mode equal to 40. Nevertheless a glance at the figure shows that they are markedly different. In Data Set I the measurements vary only slightly from the center, while for Data Set II the measurements vary greatly. Just as we have attached numbers to a data set to locate its center, we now wish to associate to each data set numbers that measure quantitatively how the data either scatter away from the center or cluster close to it. These new quantities are called measures of variability, and we will discuss three of them.

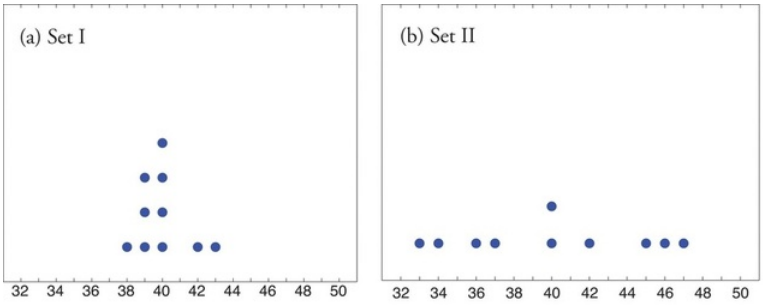



Figure 2.7.1: Dot Plots of Data Sets

## The Range

First we discuss the simplest measure of variability.

**Definition: range**

The *range* $R$ of a data set is difference between its largest and smallest values

$$R = x_{\max} - x_{\min} \tag{2.7.1}$$

where $x_{\max}$ is the largest measurement in the data set and $x_{\min}$ is the smallest.

**Example 2.7.1: Identifyig the Range of a dataset**

Find the range of each data set in Table 2.7.1.

**Solution**:

- For Data Set I the maximum is $43$ and the minimum is $38$, so the range is $R = 43 - 38 = 5$.
- For Data Set II the maximum is $47$ and the minimum is $33$, so the range is $R = 47 - 33 = 14$.

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite. The range is very limited in the information it gives us, as it is only based on the largest and smallest values. Anything can happen in between and the range tells us nothing about these. In order to get information about how all of the data points are spread out we can compare each one to the mean. We do this with the "Variance" and "Standard Deviation".

## The Variance and the Standard Deviation

The other two measures of variability that we will consider are the Variance and the Standard Deviation. They are intimately connected, as the standard deviation is just the square root of the variance. The word "deviation" gives us the clue of what we are trying to do. In order to measure how much variation there is in the data, we use the mean as the central value and then calculate all of the differences ("deviations") of each data value from the mean. The Variance is easier to calculate because it does not involve the square root. It has a drawback in that the quantities used are squared, so it will not represent the correct units for the data. The Standard Deviation takes the square root of the Variance and so the squared units are returned to regular units (such as inches, pounds and so forth based on the sampled data0.

### 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 $x - \bar{x}$.

To calculate the standard deviation, we need to calculate the variance first, and then take the square root. 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 $\sigma^2$ represents the population variance; the population standard deviation $\sigma$ 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.

In summary, 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 the same except that for the sample we divide by "sample size - 1: n-1" and for the population we divide by "Population size N". 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 $\sigma$ (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 $\sigma$.

**Formulas for the Sample Standard Deviation**

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{2.7.2}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

**Formulas for the Population Standard Deviation**

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{2.7.3}$$

For the population standard deviation, the denominator is $N$, the number of items in the population.

The Sample Variance is the calculation before taking the square root.

Definition: sample variance and sample Standard Deviation

The *sample variance* of a set of $n$ sample data is the number $s^2$ defined by the formula

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \tag{2.7.4}$$

An algebraically equivalent formula is sometimes used, because the calculations are easier to perform:

$$s^2 = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1} \tag{2.7.5}$$

The square root $\mathbf{s}$ of the sample variance is called the *sample standard deviation* of a set of $n$ *sample data* . It is given by the formulas

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}}. \tag{2.7.6}$$

Although the first formula in each case looks less complicated than the second, the latter is easier to use in hand computations, and is called a *shortcut formula*.

**Example 2.7.2: Identifying the Variance and Standard Deviation of a Dataset**

Find the sample variance and the sample standard deviation of Data Set II in Table 2.7.1.

**Solution**

To use the defining formula (the first formula) in the definition we first compute for each observation $x$ its deviation $x - \bar{x}$ from the sample mean. Since the mean of the data is $\bar{x} = 40$, we obtain the ten numbers displayed in the second line of the supplied table

| $x$ | 46 | 37 | 40 | 33 | 42 | 36 | 40 | 47 | 34 | 45 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x - \bar{x}$ | −6 | −3 | 0 | −7 | 2 | −4 | 0 | 7 | −6 | 5 |

Thus

$$\sum (x-\bar{x})^2 = 6^2 + (-3)^2 + 0^2 + (-7)^2 + 2^2 + (-4)^2 + 0^2 + 7^2 + (-6)^2 + 5^2 = 224$$

so the variance is

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{224}{9} = 24.\bar{8}$$

and the standard deviation is

$$s = \sqrt{24.\bar{8}} \approx 4.99$$

The student is encouraged to compute the ten deviations for Data Set I and verify that their squares add up to 20, so that the sample variance and standard deviation of Data Set I are the much smaller numbers

$$s^2 = 20/9 = 2.\bar{2} \tag{2.7.7}$$

and

$$s = 20/9 \approx 1.49 \tag{2.7.8}$$

**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.

## WeBWorK Problems

The number of standard deviations a data value is from the mean can be used as a measure of the closeness of a data value to the mean. Because the standard deviation measures the spread of the data this gives a uniform measure for any data set.

For example: 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.

**Rosa waits for seven minutes:**

- This is two minutes longer than the average wait time.
- Two minutes is the same as one standard deviation.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

**Binh waits for one minute.**

- This 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 **two standard deviations below the average** of five minutes.
- A "rule of thumb" is that more than two standard deviations away from the average is considered "far from the average". 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 **2.7.1**

- 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 $\sigma$ (sigma, lower case) represents the population standard deviation.

The symbol $\bar{x}$ is the sample mean and the Greek symbol $\mu$ is the population mean.

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 $\sigma_x$ or $s_x$ from the summary statistics. 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 calculator instructions appear at the end of this example.)

**Example 2.7.1**

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**

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.
- $(\bar{x} - 1.5s) = 10.53–(1.5)(0.72) = 9.45$
- $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

## 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 **2.7.1**, 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 $\sigma$, 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 $\sigma$ 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 2.7.2

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.

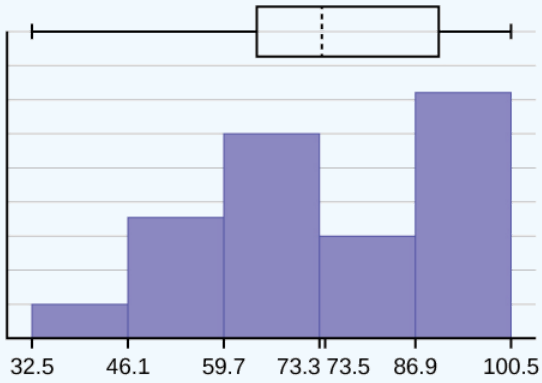



Figure 2.7.2.

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?) |

## 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:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{2.7.9}$$

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 2.7.3**

Find the standard deviation for the data in Table **2.7.3**.

Table **2.7.3**

| 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.

## 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 | $x = \bar{x} + zs$ | |
| Population | $x = \mu + z\sigma$ | |

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.

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, $\mu$, and the formula $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$. .

## Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{2.7.10}$$

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

## 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.

**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.

## 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 2.7: Measures of Spread- Variance 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.

- **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.
- **2.3: Measures of Variability** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

# 2.8: 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.

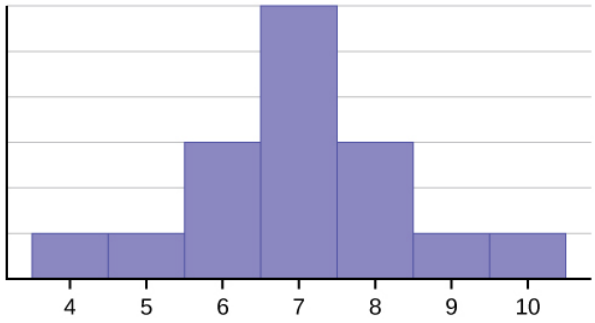



Figure **2.8.1**

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.

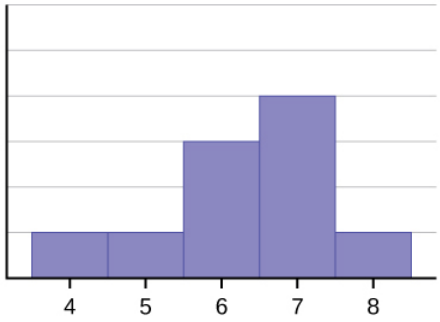



Figure **2.8.2**

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**.

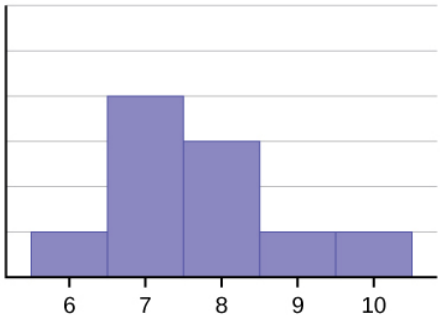



Figure **2.8.3**

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 2.8.1**

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.

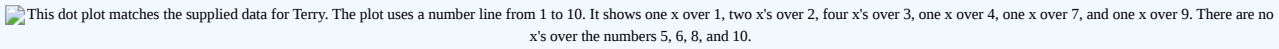



***Figure* 2.8.4**: Terry's distribution has a right (positive) skew.

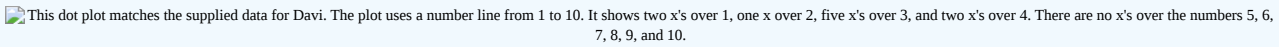



***Figure* 2.8.5**: Davis' distribution has a left (negative) skew

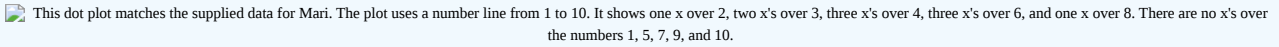



***Figure* 2.8.6**: Maris' distribution is symmetrically shaped.

b. Terry's mean is 3.7, Davis' mean is 2.7, Maris' mean is 4.6.
c. Terry's median is three, Davis' median is three. Maris' median is four.
d. 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.

## 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 **2.8.2**. A **right (or positive) skewed** distribution has a shape like Figure **2.8.3**. A **symmetrical** distribution looks like Figure **2.8.1**.

## 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 2.8: 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.

- **2.7: Skewness and the Mean, Median, and Mode** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 3: Introduction to Linear Regression and Correlation

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.


3.1: Linear Equations

3.2: Scatter Plots

3.3: Simple Linear Regression

3.4: Prediction

3.5: Outliers


### 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: Introduction 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.

# 3.1: 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 3.1.1**

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

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 3.1.2**

Graph the equation $y = -1 + 2x$.

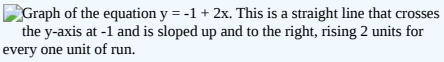



Figure 3.1.1.

**Example 3.1.3**

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**

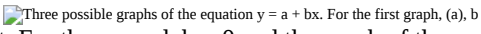
Let $x =$ the number of hours it takes to get the job done.

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: $y = 31.50 + 32x$

## 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.




0 and so the line slopes upward to the right. For the second, b = 0 and the graph of the equation is a horizontal line. In the third graph, (c), b < 0 and the line slopes downward to the right." src="

http://cnx.org/resources/917c2e46d01...ch12_03_01.jpg

" style="width: 725px; height: 170px;">

Figure 3.1.3:. 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 3.1.4**

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.

## 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.

## 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: 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.

- **12.2: Linear Equations** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 3.2: 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 3.2.1**

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 **3.2.1**: 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 |

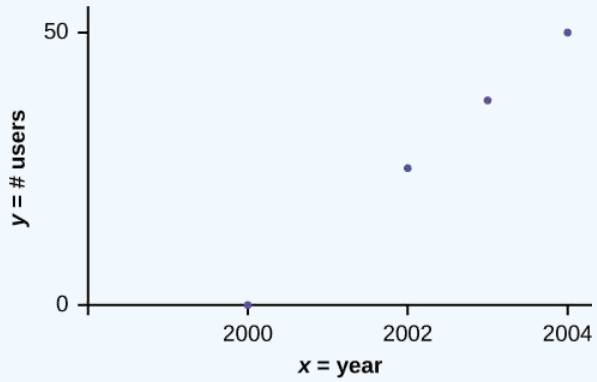



Figure **3.2.1**: Scatter plot showing the number of m-commerce users (in millions) by year.

**Exercise 3.2.1**

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**

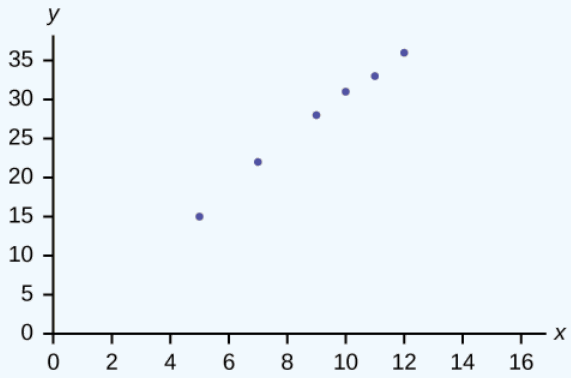



***Figure* 3.2.2**

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.

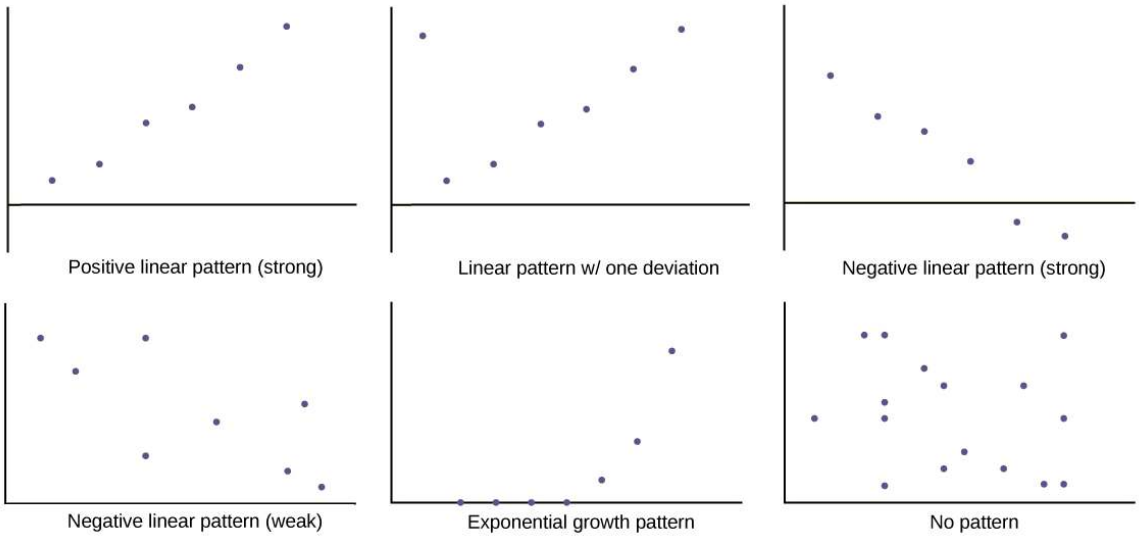



Figure **3.2.3**:

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.

## 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: 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.

- **12.3: Scatter Plots** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 3.3: Simple Linear Regression

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.

**Example 3.3.1**

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 |

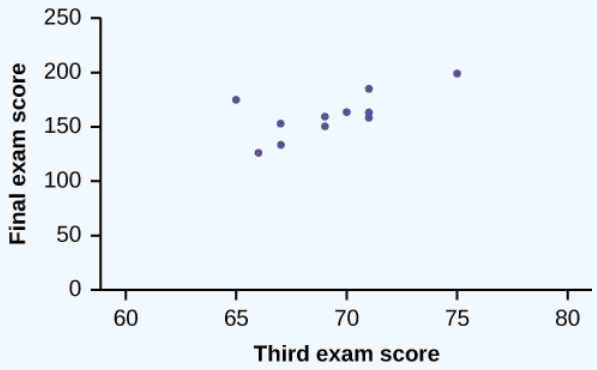



Figure 3.3.1: Scatter plot showing the scores on the final exam based on scores from the third exam.

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.

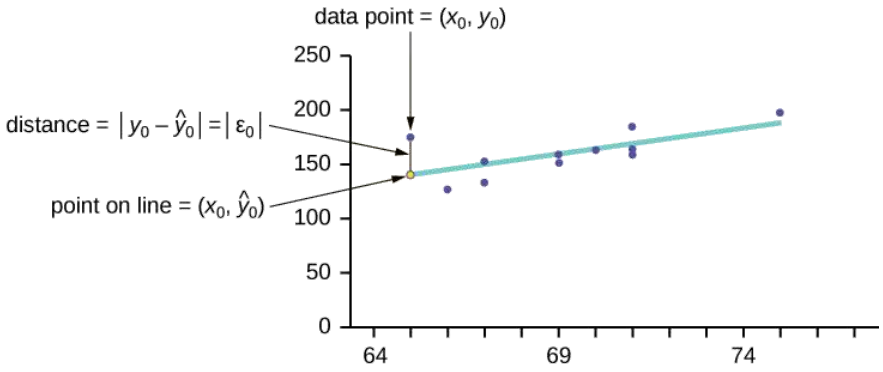



Figure **3.3.2**

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 $\varepsilon$ values. If you square each $\varepsilon$ and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + ... + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \tag{3.3.1}$$

Equation3.3.1 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:

$$\hat{y} = a + bx \tag{3.3.2}$$

where

- $a = \bar{y} - b\bar{x}$ and
- $b = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$.

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 $b = r\left(\dfrac{s_y}{s_x}\right)$ 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** .

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs.

**THIRD EXAM vs FINAL EXAM EXAMPLE:**

The graph of the line of best fit for the third-exam/final-exam example is as follows:

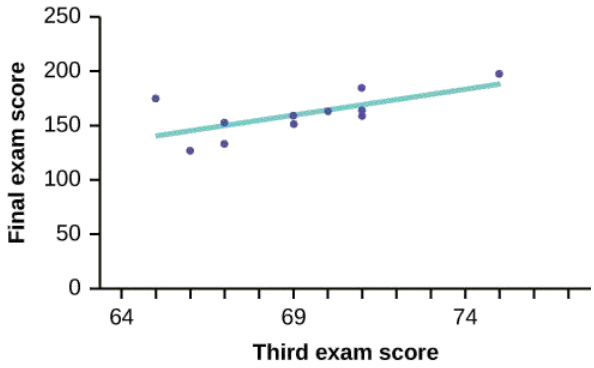



Figure **3.3.3**

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{3.3.3}$$

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.

## 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

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{3.3.4}$$

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.

> Strong correlation does not suggest that $x$ causes $y$ or $y$ causes $x$. We say **"correlation does not imply causation."**

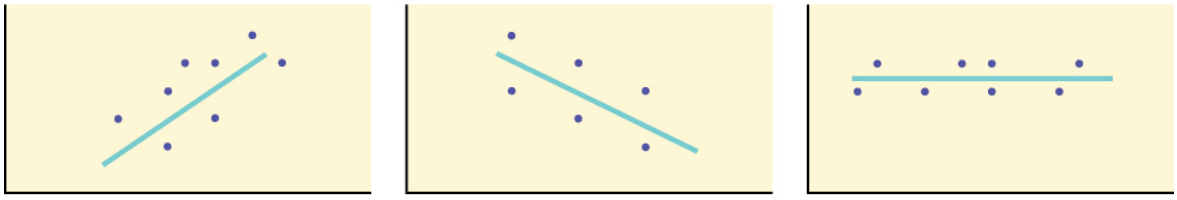



Figure **3.3.5**: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

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.

## WeBWorK Problems

## 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:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{3.3.5}$$

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 $y$.

## 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 3.3: Simple Linear 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.

- **12.4: The Regression Equation** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **1.2: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 3.4: 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 3.4.1**

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 3.4.1**

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

## WeBWorK Problems

## 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.

## 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 3.4: 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.

- **12.6: Prediction** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 3.5: 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.

> *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.*

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 3.5.1**

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.

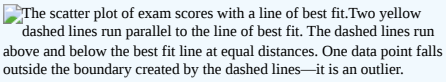

The scatter plot of exam scores with a line of best fit.Two yellow dashed lines run parallel to the line of best fit. The dashed lines run above and below the best fit line at equal distances. One data point falls outside the boundary created by the dashed lines—it is an outlier.

Figure 12.7.1.

## Numerical Identification of Outliers

In Table, 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:

$$s = \sqrt{\frac{SSE}{n-2}}$$

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.)

> 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 3.5.3: 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.)

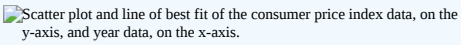



Figure 12.7.3.

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?)

## 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 |

## 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

## 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 3.5: 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.

- **12.7: Outliers** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 4: Probability Theory

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: Probability Experiments and Sample Spaces
4.2: Experiments Having Equally Likely Outcomes
4.3: Conditional Probability and Independence
4.4: Counting Basics- the Multiplication and Addition Rules
4.5: Intersection and Union of Events and Venn Diagrams
4.6: Joint and Marginal Probabilities and Contingency Tables
4.7: More Counting- Factorials, Combinations, and Permutations


### 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 Theory 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: Probability Experiments and Sample Spaces

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 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).

- $P(A) = 0$ 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).

**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 $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. $A$ 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

$$P(A) + P(A') = 1.$$

For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = 1, 2, 3, 4$. Then, $A' = 5, 6$ and $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and

$$P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1.$$

The conditional probability of $A$ given $B$ is written $P(A|B)$. $P(A|B)$ 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 $P(A|B)$ is

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let $A =$ face is 2 or 3 and $B =$ face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2, 4, 6\}$. 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.

$$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{P(\text{B})}$$
$$= \frac{\frac{\text{the number of outcomes that are 2 or 3 and even in S}}{6}}{\frac{\text{the number of outcomes that are even in S}}{6}}$$
$$= \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

## 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 4.1.1**

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. $P(\text{A}) =$ _____, $P(\text{B}) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(\text{A AND B}) =$ _____, $P(\text{A OR B}) =$ _____
f. $\text{A}' =$ _____, $P(\text{A}') =$ _____
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

**Example 4.1.2A**

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 T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event 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 $\mathbf{N}$ = the outcome is a prime number.
k. Event $\mathbf{I}$ = the outcome is seven.

**Solution**

a. $\mathbf{T} = \{2\}$, $P(\mathrm{T}) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(\mathrm{A}) = \frac{1}{2}$
c. $\mathbf{B} = \{1, 2, 3\}$, $P(\mathrm{B}) = \frac{1}{2}$
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. $\mathrm{A\ AND\ B} = 2, P(\mathrm{A\ AND\ B}) = \frac{1}{6}$
h. $\mathrm{A\ OR\ B} = \{1, 2, 3, 4, 6\}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
i. $\mathrm{A\ OR\ B}' = \{2, 4, 5, 6\}$, $P(\mathrm{A\ OR\ B}') = \frac{2}{3}$
j. $\mathbf{N} = \{2, 3, 5\}$, $P(\mathrm{N}) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

## Example 4.1.2B

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. $P(\mathrm{M\ OR\ R}) = 0.96$

i. $P(\text{F OR L}) = 0.57$
j. $P(\text{M'}) = 0.48$
k. $P(\text{R}|\text{M}) = 0.8269$ (rounded to four decimal places)
l. $P(\text{F}|\text{L}) = 0.3077$ (rounded to four decimal places)
m. $P(\text{L}|\text{F}) = 0.0833$

## WeBWorK Problems

## 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

$\text{A}$ and $\text{B}$ are events

$P(\text{S}) = 1$ where $\text{S}$ is the sample space

$0 \leq P(\text{A}) \leq 1$

$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{\text{P(B)}}$

## 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 $\text{A}$ and $\text{B}$ 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.

---

This page titled 4.1: Probability Experiments and Sample Spaces 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: Terminology** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 4.2: Experiments Having Equally Likely Outcomes

**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 $P(\mathbf{A}) = \frac{2}{4} = 0.5$.

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}. $P(\mathbf{E}) = \frac{2}{6}$. 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.

## WeBWorK Problems

4.2: Experiments Having Equally Likely Outcomes is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.2: Terminology** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 4.3: Conditional Probability and Independence

## Conditional Probability

What is the probability of an event A "given that" we have some partial information about the outcome of the experiment? This is called *conditional probability*.

**Definition: Conditional Probability**

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(AandB)}{P(B)} \tag{4.3.1}$$

## 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 $\mathrm{A}$ and $\mathrm{B}$ 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 **4.3.1**):

- **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*.

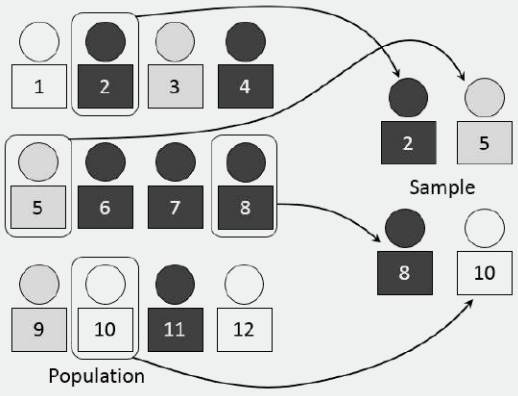



Figure **4.3.1**: A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". (CC BY-SA 4.0; Dan Kernler).

If it is not known whether $\mathrm{A}$ and $\mathrm{B}$ are independent or dependent, assume they are dependent until you can show otherwise.

**Example 4.3.1: 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.

**Example 4.3.2**

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

## 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 4.3.3**

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 $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for A and for B are $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{4}$.
- Let C = the event of getting all heads. $C = \{HH\}$. Since $B = \{TT\}$, $P(B \text{ AND } C) = 0$. 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. $D = \{TT\}$. $P(D) = \frac{1}{4}$
- 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.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$
- 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. $F = \{HT, TH, TT\}$. $P(F) = \frac{3}{4}$

**Example 4.3.4**

Flip two fair coins. Find the probabilities of the events.

a. Let F = the event of getting at most one tail (zero or one tail).
b. Let G = the event of getting two faces that are the same.
c. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let J = the event of getting all tails. Are J and H mutually exclusive?

**Solution**

Look at the sample space in Example 4.3.3.

1. Zero (0) or one (1) tails occur when the outcomes $HH, TH, HT$ show up. $P(F) = \frac{3}{4}$
2. Two faces are the same if $HH$ or $TT$ show up. $P(G) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when $HH$ or $HT$ show up. $P(H) = \frac{2}{4}$
4. F and G share $HH$ so $P(F \text{ AND } G)$ is not equal to zero (0). F and G are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins ($TT$). H's outcomes are $HH$ and $HT$.

J and H have nothing in common so $P(J \text{ AND } H) = 0$. J and H are mutually exclusive.

**Example 4.3.5**

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then $A = \{1, 3, 5\}$. Let event B = a face is even. Then $B = \{2, 4, 6\}$.

- Find the complement of A, A′. The complement of A, A′, is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then $C = \{3, 5\}$. Let event D = all even faces smaller than five. Then $D = \{2, 4\}$. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. $E = \{1, 2, 3, 4\}$.

Are C and 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, $P(C|A) = \frac{2}{3}$.

## Example 4.3.6

Let event G = taking a math class. Let event H = taking a science class. Then, G 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.

## Example 4.3.7

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(D|C)$?

**Solution**

a. Yes, because $P(C|D) = P(C)$.
b. No, because $P(C \text{ AND } D)$ is not equal to zero.
c. $P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{0.75} = 0.3$

**Example 4.3.8**

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

- $\mathrm{R}$ = red card is drawn,
- $\mathrm{B}$ = blue card is drawn,
- $\mathrm{E}$ = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$.

$S$ has eight outcomes.

- $P(\mathrm{R}) = \frac{3}{8}$. $P(\mathrm{B}) = \frac{5}{8}$. $P(\mathrm{R \text{ AND } B}) = 0$. (You cannot draw one card that is both red and blue.)
- $P(\mathrm{E}) = \frac{3}{8}$. (There are three even-numbered cards, $R2, B2$, and $B4$.)
- $P(\mathrm{E|B}) = \frac{2}{5}$. (There are five blue cards: $B1, B2, B3, B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(\mathrm{B|E}) = \frac{2}{3}$. (There are three even-numbered cards: $R2, B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events $\mathrm{R}$ and $\mathrm{B}$ are mutually exclusive because $P(\mathrm{R \text{ AND } B}) = 0$.
- Let $\mathrm{G}$ = card with a number greater than 3. $\mathrm{G} = \{B4, B5\}$. $P(\mathrm{G}) = \frac{2}{8}$. Let $\mathrm{H}$ = blue card numbered between one and four, inclusive. $\mathrm{H} = \{B1, B2, B3, B4\}$. $P(\mathrm{G|H}) = \frac{1}{4}$. (The only card in $\mathrm{H}$ that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(\mathrm{G}) = P(\mathrm{G|H})$, which means that $\mathrm{G}$ and $\mathrm{H}$ are independent.

**Example 4.3.9**

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 $\mathrm{F}$ be the event that a student is female. Let $\mathrm{L}$ 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:
- $P(\mathrm{F}) = 0.60$; $P(\mathrm{L}) = 0.50$
- $P(\mathrm{F \text{ AND } L}) = 0.45$
- $P(\mathrm{L|F}) = 0.75$

*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 $P(\mathrm{F|L})$ yet, so you cannot use the second condition.*

**Solution 1**

Check whether $P(\mathrm{F \text{ AND } L}) = P(\mathrm{F})P(\mathrm{L})$. We are given that $P(\mathrm{F \text{ AND } L}) = 0.45$, but $P(\mathrm{F})P(\mathrm{L}) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(\mathrm{F \text{ AND } L})$ does not equal $P(\mathrm{F})P(\mathrm{L})$.

**Solution 2**

Check whether $P(\mathrm{L|F})$ equals $P(\mathrm{L})$. We are given that $P(\mathrm{L|F}) = 0.75$, but $P(\mathrm{L}) = 0.50$; 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.

**Example 4.3.10**

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 A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.
A = {_____}. Find $P(A)$.
f. Event B = heads on the coin followed by a three on the die. B = {_____}. Find $P(B)$.
g. Are A and B mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.)
h. Are A and B independent? (Hint: Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then A and 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. $A = \{H2, H4, H6\}$; $P(A) = \frac{3}{12}$
f. $B = \{H3\}$; $P(B) = \frac{1}{12}$
g. Yes, because $P(A \text{ AND } B) = 0$
h. $P(A \text{ AND } B) = 0$. $P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so A and B are dependent.

## WeBWorK Problems

## 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, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.
- If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B) \text{and} P(A \text{ AND } B) = 0$.

## 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 4.3.11**

E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E|F)$.

**Exercise 4.3.12**

J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

**Answer**

$P(J) = 0.3$

**Exercise 4.3.13**

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

a. $P(U \text{ AND } V) =$
b. $P(U|V) =$
c. $P(U \text{ OR } V) =$

**Exercise 4.3.14**

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

**Answer**

$P(Q \text{ AND } R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

## Bringing It Together

**Exercise 4.3.16**

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 4.3.17**

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. $P(C) =$ _____
b. $P(P|C) =$ _____

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 4.3.18**

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 4.3.19**

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. $P(\text{NOT (J OR K)}) = 1 - P(\text{J OR K}) = 1 - 0.45 = 0.55$

## WeBWorK Problems

## 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 4.3: Conditional Probability and 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.

- **3.3: Independent and Mutually Exclusive Events** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **2.2: Conditional Probability I** by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: https://www.openintro.org/book/os.

# 4.4: Counting Basics- the Multiplication and Addition Rules

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:

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{4.4.1}$$

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 4.4.1 becomes

$$P(A \text{ AND } B) = P(A)P(B).$$

## The Addition Rule

If A and B are defined on a sample space, then:

If A and B are **mutually exclusive**, then

$$P(A \text{ AND } B) = 0.$$

and Equation ??? becomes

$$P(A \text{ OR } B) = P(A) + P(B).$$

**Example 4.4.1**

Klaus is trying to choose where to go on vacation. His two choices are: $\mathbf{A} = \mathbf{New\ Zealand}$ and $\mathbf{B} = \mathbf{Alaska}$.

- 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 $P(\mathbf{B}) = 0.35$.
- $P(\mathbf{A \text{ AND } B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\mathbf{A \text{ OR } B}) = P(\mathbf{A}) + P(\mathbf{B}) = 0.6 + 0.35 = 0.95$. 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 $P(\text{A OR B})$.

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\text{B AND A}) = 0.585$.

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \tag{4.4.2}$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \tag{4.4.3}$$

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.

## Example 4.4.2

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.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{4.4.4}$$

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{4.4.5}$$

$$0.0667 \neq 0.0996 \tag{4.4.6}$$

## Example 4.4.3

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, M|S = math given speech

a. What is the probability that Felicity enrolls in math and speech?
Find $P(\text{M AND S}) = P(\text{M}|\text{S})P(\text{S})$.
b. What is the probability that Felicity enrolls in math or speech classes?
Find $P(\text{M OR S}) = P(\text{M}) + P(\text{S}) - P(\text{M AND S})$.
c. Are M and S independent? Is $P(\text{M}|\text{S}) = P(\text{M})$?
d. Are M and S mutually exclusive? Is $P(\text{M AND S}) = 0$?

**Answer**

a. 0.1625, b. 0.6875, c. No, d. No

**Example 4.4.4**

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

**Example 4.4.5**

Refer to the information in Example 4.4.4. 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

## WeBWorK Problems

## 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 $\mathbf{A}$ and $\mathbf{B}$, as well as the probability of $\mathbf{A}$ or $\mathbf{B}$ for two given events $\mathbf{A}$, $\mathbf{B}$ 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 $\mathbf{A}$ and $\mathbf{B}$ 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:** $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

## 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. $P(\text{A}|\text{B}) = P(\text{A})$
2. $P(\text{B}|\text{A}) = P(\text{B})$
3. $P(\text{A AND B}) = P(\text{A})P(\text{B})$

**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 $P(\text{A AND B}) = 0$.

---

This page titled 4.4: Counting Basics- the Multiplication and Addition Rules 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.

# 4.5: Intersection and Union of Events 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 4.5.1: 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.

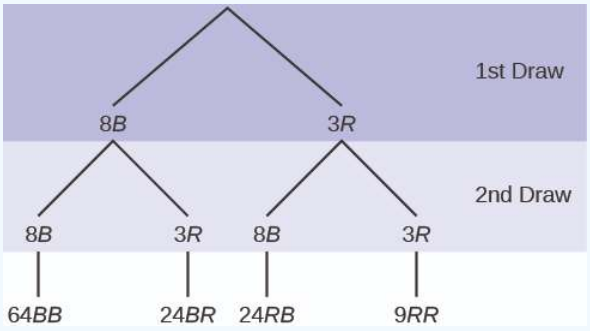



Figure 4.5.1: Total = 64 + 24 + 24 + 9 = 121

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.

**Example 4.5.2: 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, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

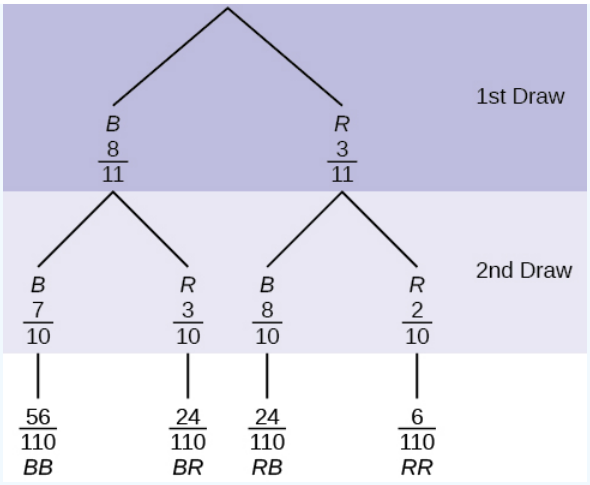



Figure 4.5.3: Total $= \frac{56+24+24+6}{110} = \frac{110}{110} = 1$

*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.*

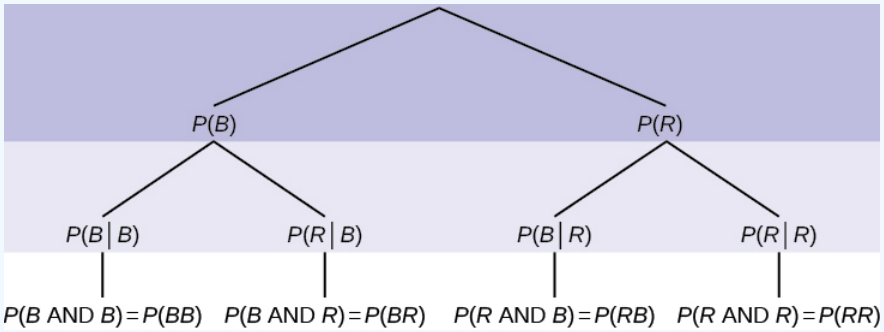
Calculate the following probabilities using the tree diagram.

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, $P(\text{B on 2nd}|\text{R on 1st}) = P(\text{R}|\text{B}) = \frac{8}{10}$.

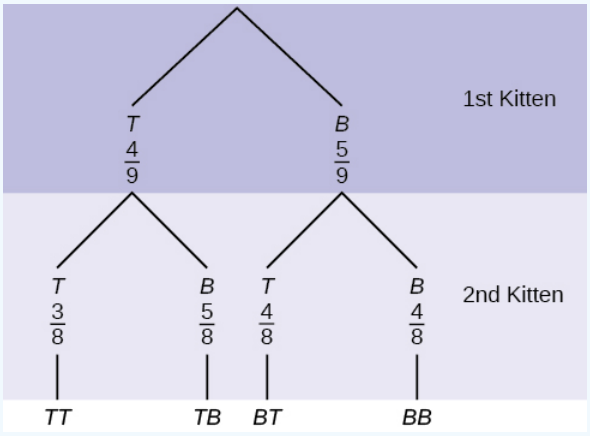
If we are using probabilities, we can label the tree in the following general way.




- $P(\text{R}|\text{R})$ 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 $P(\text{B on 2nd}|\text{B on 1st})$

**Example 4.5.3**

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. $\frac{32}{72}$

## 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 4.5.4**

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:

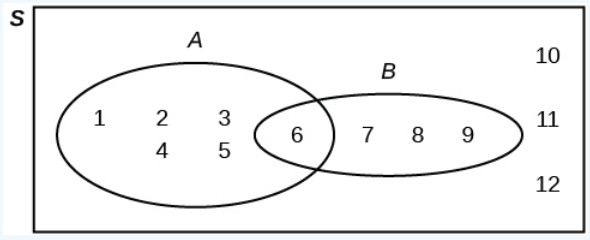



Figure 4.5.5:

**Example 4.5.5**

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:

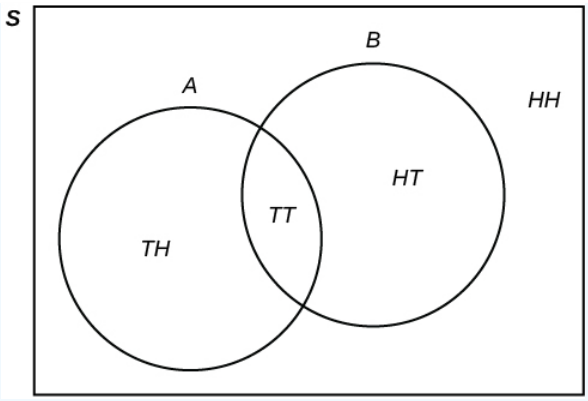



Figure 4.5.7:

**Example 4.5.6: 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 $\mathbf{C}$ = student belongs to a club and $\mathbf{PT}$ = student works part time.

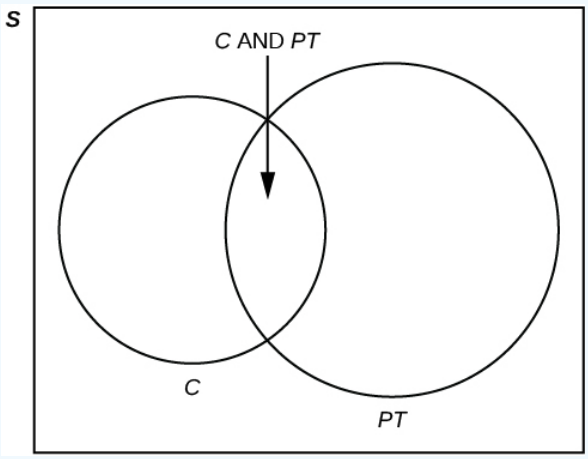



Figure 4.5.8:

If a student is selected at random, find

- the probability that the student belongs to a club. $P(\mathbf{C}) = 0.40$
- the probability that the student works part time. $P(\mathbf{PT}) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(\mathbf{C\ AND\ PT}) = 0.05$
- the probability that the student belongs to a club **given** that the student works part time. $P(\mathbf{C}|\mathbf{PT}) = \frac{P(\mathbf{C\ AND\ PT})}{P(\mathbf{PT})} = \frac{0.05}{0.50} = 0.1$
- the probability that the student belongs to a club **OR** works part time. $P(\mathbf{C\ OR\ PT}) = P(\mathbf{C}) + P(\mathbf{PT}) - P(\mathbf{C\ AND\ PT})$ $= 0.40 + 0.50 - 0.05 = 0.85$

**Example 4.5.7**

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.

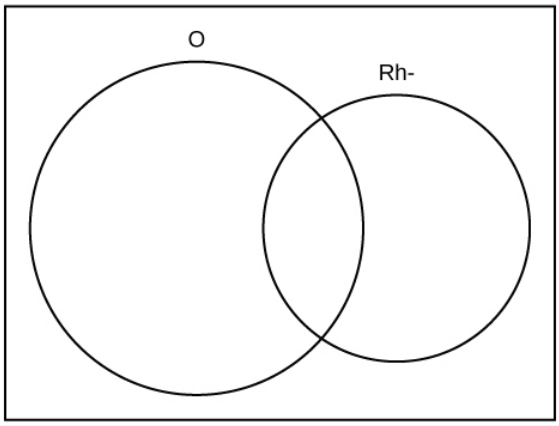



Figure **4.5.10**:

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.

## WeBWorK Problems

## 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 4.5: Intersection and Union of Events 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.

- **3.6: Tree and Venn Diagrams** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 4.6: Joint and Marginal Probabilities and 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 4.6.1**

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.)

**Example 4.6.2**

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.

$$P(\mathbf{F} \text{ AND } \mathbf{C}) = \frac{18}{100} = 0.18$$

$$P(\mathbf{F})P(\mathbf{C}) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(\mathbf{F} \text{ AND } \mathbf{C}) \neq P(\mathbf{F})P(\mathbf{C})$, so the events $\mathbf{F}$ and $\mathbf{C}$ are not independent.

c.

1. The word 'given' tells you that this is a conditional.
2. $P(\mathbf{M}|\mathbf{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

a. Find $P(\mathbf{F})$.
b. Find $P(\mathbf{P})$.
c. Find $P(\mathbf{F} \text{ AND } \mathbf{P})$.
d. Find $P(\mathbf{F} \text{ OR } \mathbf{P})$.

d.

1. $P(\mathbf{F}) = \frac{45}{100}$
2. $P(\mathbf{P}) = \frac{25}{100}$

3. $P(\text{F AND P}) = \frac{11}{100}$
4. $P(\text{F OR P}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

## Example 4.6.3

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 4.6.4

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

## 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 Loīc 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).


## WeBWork Problems

## 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 4.6: Joint and Marginal Probabilities and 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.

- **3.5: Contingency Tables** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- **1.2: Definitions of Statistics, Probability, and Key Terms** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 4.7: More Counting- Factorials, Combinations, and Permutations

## WeBWorK Problems

4.7: More Counting- Factorials, Combinations, and Permutations is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.2: Terminology** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 5: Discrete Random Variables


5.1: Introduction to Random Variables
5.2: The Probability Distribution Function
5.3: Expectation, Variance and Standard Deviation
5.4: The Binomial Distribution
5.5: The Geometric Distribution
5.6: The Hypergeometric Distribution
5.7: The Poisson Distribution


### 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 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.

# 5.1: Introduction to 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.

## 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 5.1: Introduction to 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.

- **4.1: Prelude to Discrete Random Variables** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 5.2: The Probability Distribution Function

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,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \quad (5.2.1)$$

**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 |

## WeBWorK Problems

## 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.

## 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.

*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.

## 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: The Probability Distribution Function 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: Probability Distribution Function (PDF) for a Discrete Random Variable** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 5.3: Expectation, Variance 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 $\mu$. 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, $\mu$, 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, $\mu$, 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 $\mu$ 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, $\mu$, 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.

## 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.

## 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, $\mu$? 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.

Like data, probability distributions have standard deviations. To calculate the standard deviation ($\sigma$) 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 $(x)$–$\mu^2 P(x)$ 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 $\mu$ and $\sigma$ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating $\mu$ and $\sigma$.

**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 $\mu$ and standard deviation $\sigma$ 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 $\mu$ and $\sigma$.

| $x$ | $P(x)$ | $xP(x)$ | $(x–\mu)^2 \cdot P(x)$ |
|---|---|---|---|
| 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:

$$\sigma = \sqrt{\frac{18}{36}} \approx 0.7071. \quad (5.3.1)$$

**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 $= \sqrt{648.0964 + 176.6636} \approx 28.7186$

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: $\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$

## WeBWorK Problems

## Glossary

**Expected Value**

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: $\mu$. 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 $\mu$) 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).


## 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 5.3: Expectation, Variance 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: The Binomial Distribution

Everyone is familiar with a multiple-choice test. Each question has a fixed number of possible answers but only one of them is correct. If we don't know anything about the question then we can still succeed if we guess the correct answer. What is the chance that we can pass the test just by guessing?

We can answer this by setting up a mathematical model that describes this situation. This is an example of a particular scenario called the Binomial Distribution. We can identify 4 specific characteristics of this problem:

1) There is an event with only 2 possible outcomes: success and failure. [This is the guess for a particular question.]

2) The event is repeated a fixed number of times ("trials") with exactly the same chance of success. [This is the number of questions. The chance of success = 1/number of choices]

3) Each separate repetition is independent of all the others. [Questions are independent of each other]

To make it specific, consider that there are 4 possible answers for each question and that there are 10 questions on the test.

Set $p$ = probability of success (guessing the correct answer on one question)

$n$ = the number of questions

$p = 0.25$.

$n = 10$

The "score", which is the number of correct answers, we denote by a random variable X.

We can set up a probability distribution table for X by listing all of the possible scores k = 0,1,2,…,9,10 together with their probabilities:

| Values for k (Possible scores) | P(X=k) |
|---|---|
| 0 | |
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

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 $N$.

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, $\mu$, and variance, $\sigma^2$, for the binomial probability distribution are

$$\mu = np \tag{5.4.1}$$

and

$$\sigma^2 = npq. \tag{5.4.2}$$

The standard deviation, $\sigma$, 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.

**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 $P(x = 15)$.

**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**

$P(x > 10)$

**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 ($\geq$). The probability question is $P(x \geq 40)$.

**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 $(x =$ value$)$ : binompdf$(n, p,$ number$)$** if "number" is left out, the result is the binomial probability table.

**To calculate $P(x \leq$ value$)$ : binomcdf$(n, p,$ 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 $1 -$ 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:

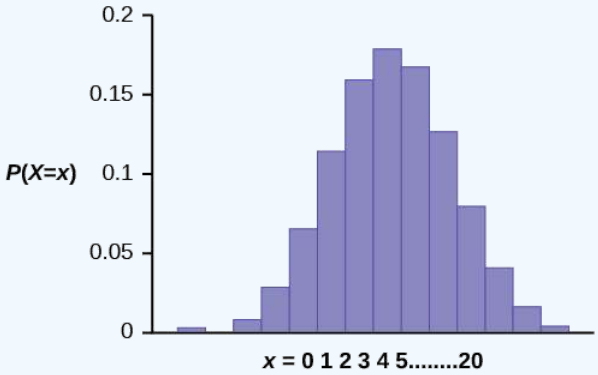



Figure 5.4.1 : The graph of $X \sim B(20, 0.41)$.

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 $\sigma = \sqrt{npq}$.

$$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20. \tag{5.4.5}$$

.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 $= \sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

## 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.

**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.

## WeBWorK Problems

## 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 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 $\sigma = \sqrt{npq}$.

## 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.

*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.

## 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: 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.

- **4.4: Binomial Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 5.5: The Geometric Distribution

There are three main characteristics of a geometric experiment.

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bullseye until you hit the bullseye. The first time you hit the bullseye is a "success" so you stop throwing the dart. It might take six tries until you hit the bullseye. You can think of the trials as failure, failure, failure, failure, failure, success, STOP.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, $p$, of a success and the probability, $q$, of a failure is the same for each trial. $p + q = 1$ and $q = 1 - p$. For example, the probability of rolling a three when you throw one fair die is $\frac{1}{6}$. This is true no matter how many times you roll the die. Suppose you want to know the probability of getting the first three on the fifth roll. On rolls one through four, you do not get a face with a three. The probability for each of the rolls is $q = \frac{5}{6}$, the probability of a failure. The probability of getting a three on the fifth roll is $\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = 0.0804$.

$X =$ the number of independent trials until the first success.

You play a game of chance that you can either win or lose (there are no other possibilities) **until** you lose. Your probability of losing is $p = 0.57$. What is the probability that it takes five games until you lose? Let $X =$ the number of games you play until you lose (includes the losing game). Then $X$ takes on the values 1, 2, 3, ... (could go on indefinitely). The probability question is $P(x = 5)$.

**Example 5.5.1**

You play a game of chance that you can either win or lose (there are no other possibilities) **until** you lose. Your probability of losing is $p = 0.57$. What is the probability that it takes five games until you lose? Let $X =$ the number of games you play until you lose (includes the losing game). Then $X$ takes on the values 1, 2, 3, ... (could go on indefinitely). The probability question is $P(x = 5)$.

**Example 5.5.2**

A safety engineer feels that 35% of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) **until** she finds one that shows an accident caused by failure of employees to follow instructions. On average, how many reports would the safety engineer **expect** to look at until she finds a report showing an accident caused by employee failure to follow instructions? What is the probability that the safety engineer will have to examine at least three reports until she finds a report showing an accident caused by employee failure to follow instructions?

Let $X$ = the number of accidents the safety engineer must examine **until** she finds a report showing an accident caused by employee failure to follow instructions. $X$ takes on the values 1, 2, 3, .... The first question asks you to find the **expected value** or the mean. The second question asks you to find $P(x \geq 3)$. ("At least" translates to a "greater than or equal to" symbol).

**Example 5.5.3**

Suppose that you are looking for a student at your college who lives within five miles of you. You know that 55% of the 25,000 students do live within five miles of you. You randomly contact students from the college **until** one says he or she lives within five miles of you. What is the probability that you need to contact four people?

*This is a geometric problem because you may have a number of failures before you have the one success you desire. Also, the probability of a success stays the same each time you ask a student if he or she lives within five miles of you. There is no definite number of trials (number of times you ask a student).*

a. Let $X =$ the number of _____ you must ask _____ one says yes.
b. What values does $X$ take on?
c. What are $p$ and $q$?

d. The probability question is $P$(_____).

**Solution**

a. Let $X =$ the number of **students** you must ask **until** one says yes.
b. 1, 2, 3, …, (total number of students)
c. $p = 0.55; q = 0.45$
d. $P(x = 4)$

## Notation for the Geometric: $G =$ Geometric Probability Distribution Function

$X \sim G(p)$

Read this as "$X$ is a random variable with a geometric distribution." The parameter is $p$; $p =$ the probability of a success for each trial.

## Example 5.5.4

Assume that the probability of a defective computer component is 0.02. Components are randomly selected. Find the probability that the first defect is caused by the seventh component tested. How many components do you expect to test until one is found to be defective?

Let $X$ = the number of computer components tested until the first defect is found.

$X$ takes on the values 1, 2, 3, ... where $p = 0.02$. $X \sim G(0.02)$

Find $P(x = 7)$. $P(x = 7) = 0.0177$.

To find the probability that $x = 7$,

- Enter 2nd, DISTR
- Scroll down and select geometpdf(
- Press ENTER
- Enter 0.02, 7); press ENTER to see the result: $P(x = 7) = 0.0177$

To find the probability that $x \leq 7$, follow the same instructions EXCEPT select E: geometcdf as the distribution function.

The probability that the seventh component is the first defect is 0.0177.

The graph of $X \sim G(0.02)$ is:




Figure 5.5.1

The $y$-axis contains the probability of $x$, where $X =$ the number of computer components tested.

The number of components that you would expect to test until you find the first defective one is the mean, $\mu = 50$.

The formula for the mean is

$$\mu = \frac{1}{p} = \frac{1}{0.02} = 50 \tag{5.5.1}$$

The formula for the variance is

$$\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right) = 2,450 \tag{5.5.2}$$

The standard deviation is

$$\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right)} = 49.5 \tag{5.5.3}$$

**Example 5.5.5**

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Let $X =$ the number of people you ask until one says he or she has pancreatic cancer. Then $X$ is a discrete random variable with a geometric distribution: $X \sim G\left(\frac{1}{78}\right)$ or $X \sim G(0.0128)$.

a. What is the probability of that you ask ten people before one says he or she has pancreatic cancer?
b. What is the probability that you must ask 20 people?
c. Find the (i) mean and (ii) standard deviation of $X$.

**Answer**

a. $P(x = 10) = \text{geometpdf}(0.0128, 10) = 0.0114$
b. $P(x = 20) = \text{geometpdf}(0.0128, 20) = 0.01$
c. i. Mean $= \mu = \frac{1}{p} = \frac{1}{0.0128} = 78$
   ii. Standard Deviation $= \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.0128}{0.0128^2}} \approx 77.6234$

## References


1. “Millennials: A Portrait of Generation Next,” PewResearchCenter. Available online at www.pewsocialtrends.org/files...-to-change.pdf (accessed May 15, 2013).
2. “Millennials: Confident. Connected. Open to Change.” Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at http://www.pewsocialtrends.org/2010/...pen-to-change/ (accessed May 15, 2013).
3. “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. 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).
5. “Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan,” The European Union and ICON-Institute. Available online at ec.europa.eu/europeaid/where/...summary_en.pdf (accessed May 15, 2013).
6. “The World FactBook,” Central Intelligence Agency. Available online at www.cia.gov/library/publicat...k/geos/af.html (accessed May 15, 2013).
7. “UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills,” UNICEF Television. Video available online at http://www.unicefusa.org/assets/vide...y-centers.html (accessed May 15, 2013).


## Review

There are three characteristics of a geometric experiment:

- There are one or more Bernoulli trials with all failures except the last one, which is a success.

- In theory, the number of trials could go on forever. There must be at least one trial.
- The probability, $p$, of a success and the probability, $q$, of a failure are the same for each trial.

In a geometric experiment, define the discrete random variable $X$ as the number of independent trials until the first success. We say that $X$ has a geometric distribution and write $X \sim G(p)$ where $p$ is the probability of success in a single trial. The mean of the geometric distribution $X \sim G(p)$ is $\mu = \frac{1-p}{p^2} = \sqrt{\frac{1}{p}\left(\frac{1}{p}-1\right)}$.

## 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.

## Formula Review

$X \sim G(p)$ means that the discrete random variable $X$ has a geometric probability distribution with probability of success in a single trial $p$.

$X =$ the number of independent trials until the first success

$X$ takes on the values $x = 1, 2, 3, ...$

$p =$ 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 is $\mu = \frac{1}{p}$.

The standard deviation is $\sigma = \frac{1-p}{p^2} = \sqrt{\frac{1}{p}\left(\frac{1}{p}-1\right)}$.

*Use the following information to answer the next six 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 select freshman from the study until you find one who replies “yes.” You are interested in the number of freshmen you must ask.

## Footnotes

[1]”Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...pi_data_value- last&sort=desc (accessed May 15, 2013).

## Glossary

**Geometric Distribution**

a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable $X$ is defined as the number of trials until the first success. Notation: $X \sim G(p)$. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma =$

$$\sqrt{\frac{1}{p}\left(\frac{1}{p}-1\right)} \tag{5.5.4}$$

. The probability of exactly $x$ failures before the first success is given by the formula: $P(X = x) = p(1–p)^{x-1}$.

**Geometric Experiment**

a statistical experiment with the following properties:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, $p$, of a success and the probability, $q$, of a failure do not change from trial to trial.

---

This page titled 5.5: The Geometric 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.

- **4.5: Geometric Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 5.6: The Hypergeometric Distribution

The hypergeometric distribution arises when one samples from a finite population, thus making the trials dependent on each other. There are five characteristics of a hypergeometric experiment.

### Characteristics of a hypergeometric experiment

1. You take samples from **two** groups.
2. You are concerned with a group of interest, called the first group.
3. You sample **without replacement** from the combined groups. For example, you want to choose a softball team from a combined group of 11 men and 13 women. The team consists of ten players.
4. Each pick is **not** independent, since sampling is without replacement. In the softball example, the probability of picking a woman first is $\frac{13}{24}$. The probability of picking a man second is $\frac{11}{23}$ if a woman was picked first. It is $\frac{10}{23}$ if a man was picked first. The probability of the second pick depends on what happened in the first pick.
5. You are **not** dealing with Bernoulli Trials.

The outcomes of a hypergeometric experiment fit a *hypergeometric probability* distribution. The random variable $X$ = the number of items from the group of interest.

### Example 5.6.1

A candy dish contains 100 jelly beans and 80 gumdrops. Fifty candies are picked at random. What is the probability that 35 of the 50 are gumdrops? The two groups are jelly beans and gumdrops. Since the probability question asks for the probability of picking gumdrops, the group of interest (first group) is gumdrops. The size of the group of interest (first group) is 80. The size of the second group is 100. The size of the sample is 50 (jelly beans or gumdrops). Let $X =$ the number of gumdrops in the sample of 50. $X$ takes on the values $x = 0, 1, 2, ..., 50$. What is the probability statement written mathematically?

**Answer**

$P(x = 35)$

### Example 5.6.2

Suppose a shipment of 100 DVD players is known to have ten defective players. An inspector randomly chooses 12 for inspection. He is interested in determining the probability that, among the 12 players, at most two are defective. The two groups are the 90 non-defective DVD players and the 10 defective DVD players. The group of interest (first group) is the defective group because the probability question asks for the probability of at most two defective DVD players. The size of the sample is 12 DVD players. (They may be non-defective or defective.) Let $X =$ the number of defective DVD players in the sample of 12. $X$ takes on the values $0, 1, 2, \ldots, 10$. $X$ may not take on the values 11 or 12. The sample size is 12, but there are only 10 defective DVD players. Write the probability statement mathematically.

**Answer**

$P(x \leq 2)$

### Example 5.6.3

You are president of an on-campus special events organization. You need a committee of seven students to plan a special birthday party for the president of the college. Your organization consists of 18 women and 15 men. You are interested in the number of men on your committee. If the members of the committee are randomly selected, what is the probability that your committee has more than four men?

This is a hypergeometric problem because you are choosing your committee from two groups (men and women).

a. Are you choosing with or without replacement?
b. What is the group of interest?

c. How many are in the group of interest?
d. How many are in the other group?
e. Let $X =$ _____ on the committee. What values does $X$ take on?
f. The probability question is $P$(_____).

**Solution**

a. without
b. the men
c. 15 men
d. 18 women
e. Let $X =$ the number of men on the committee. $x = 0, 1, 2, \ldots, 7$.
f. $P(x > 4)$

## Notation for the Hypergeometric: $H$ = Hypergeometric Probability Distribution Function

$$X \sim H(r, b, n) \tag{5.6.1}$$

Read this as "$X$ is a random variable with a hypergeometric distribution." The parameters are $r, b$, and $n$; $r =$ the size of the group of interest (first group), $b =$ the size of the second group, $n =$ the size of the chosen sample.

## Example 5.6.4

A school site committee is to be chosen randomly from six men and five women. If the committee consists of four members chosen randomly, what is the probability that two of them are men? How many men do you expect to be on the committee?

Let $X$ = the number of men on the committee of four. The men are the group of interest (first group).

$X$ takes on the values $0, 1, 2, 3, 4$, where $r = 6, b = 5$, and $n = 4$. $X \sim H(6, 5, 4)$

Find $P(x = 2)$. $P(x = 2) = 0.4545$ (calculator or computer)

> Currently, the TI-83+ and TI-84 do not have hypergeometric probability functions. There are a number of computer packages, including Microsoft Excel, that do.

The probability that there are two men on the committee is about 0.45.

The graph of $X \sim H(6, 5, 4)$ is:

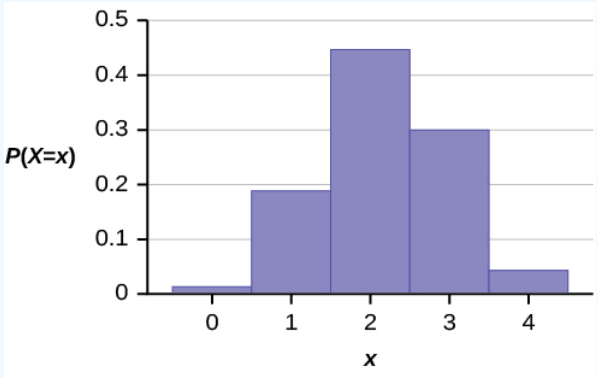



Figure 5.6.1.

The $y$-axis contains the probability of $X$, where $X =$ the number of men on the committee.

You would expect $m = 2.18$ (about two) men on the committee.

The formula for the mean is

$$\mu = \frac{nr}{r+b}\frac{(4)(6)}{6+5} = 2.18 \quad (5.6.2)$$

Summary

A hypergeometric experiment is a statistical experiment with the following properties:

- You take samples from two groups.
- You are concerned with a group of interest, called the first group.
- You sample without replacement from the combined groups.
- Each pick is not independent, since sampling is without replacement.
- You are not dealing with Bernoulli Trials.

The outcomes of a hypergeometric experiment fit a hypergeometric probability distribution. The random variable $X$ = the number of items from the group of interest. The distribution of $X$ is denoted $X \sim H(r, b, n)$, where $r =$ the size of the group of interest (first group), $b =$ the size of the second group, and $n =$ the size of the chosen sample. It follows that $n \leq r + b$. The mean of $X$ is $\mu = \frac{nr}{r+b}$ and the standard deviation is $\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$.

## Formula Review

$X \sim H(r, b, n)$ means that the discrete random variable $X$ has a hypergeometric probability distribution with $r =$ the size of the group of interest (first group), $b =$ the size of the second group, and $n =$ the size of the chosen sample.

$X$ = the number of items from the group of interest that are in the chosen sample, and $X$ may take on the values $x = 0, 1, \ldots$, up to the size of the group of interest. (The minimum value for $X$ may be larger than zero in some instances.)

$n \leq r + b$

The mean of $X$ is given by the formula $\mu = \frac{nr}{r+b}$ and the standard deviation is $= \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$.

*Use the following information to answer the next five exercises:* Suppose that a group of statistics students is divided into two groups: business majors and non-business majors. There are 16 business majors in the group and seven non-business majors in the group. A random sample of nine students is taken. We are interested in the number of business majors in the sample.

## Glossary

**Hypergeometric Experiment**

a statistical experiment with the following properties:

1. You take samples from two groups.
2. You are concerned with a group of interest, called the first group.
3. You sample without replacement from the combined groups.
4. Each pick is not independent, since sampling is without replacement.
5. You are not dealing with Bernoulli Trials.

**Hypergeometric Probability**

a discrete random variable (RV) that is characterized by:

1. A fixed number of trials.
2. The probability of success is not the same from trial to trial.

We sample from two groups of items when we are interested in only one group. $X$ is defined as the number of successes out of the total number of items chosen. Notation: $X \sim H(r, b, n)$, where $r =$ the number of items in the group of interest, $b =$ the number of items in the group not of interest, and $n =$ the number of items chosen.

## 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 5.6: The Hypergeometric 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.

- **4.6: Hypergeometric Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 5.7: The Poisson Distribution

The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

**two main characteristics of a Poisson experiment**

1. The Poisson probability distribution gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1,000). You will verify the relationship in the homework exercises. $n$ is the number of trials, and $p$ is the probability of a "success."

The random variable $X =$ the number of occurrences in the interval of interest.

**Example 5.7.1**

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in five minutes is three?

**Solution**

Let $X =$ the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, then the average number of loaves put on the shelf in five minutes is $\left(\frac{5}{30}\right)(12) = 2$ loaves of bread.

The probability question asks you to find $P(x = 3)$.

**Example 5.7.2**

A bank expects to receive six bad checks per day, on average. What is the probability of the bank getting fewer than five bad checks on any given day? Of interest is the number of checks the bank receives in one day, so the time interval of interest is one day. Let $X =$ the number of bad checks the bank receives in one day. If the bank expects to receive six bad checks per day then the average is six checks per day. Write a mathematical statement for the probability question.

**Answer**

$P(x < 5)$

**Example 5.7.3**

You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast. This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

a. What is the interval of interest?
b. What is the average number of times the news reporter says "uh" during one broadcast?
c. Let $X =$ _____. What values does $X$ take on?
d. The probability question is $P$(_____).

**Solutions**

a. one broadcast
b. 2

c. Let $X =$ the number of times the news reporter says "uh" during one broadcast.

$$x = 0, 1, 2, 3, ... \tag{5.7.1}$$

d. $P(x > 2)$

## Notation for the Poisson: $P =$ Poisson Probability Distribution Function

$$X \sim P(\mu) \tag{5.7.2}$$

Read this as "$X$ is a random variable with a Poisson distribution." The parameter is $\mu$ (or $\lambda$); $\mu$ (or $\lambda$) = the mean for the interval of interest.

### Example 5.7.4

Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call in the next 15 minutes?

**Solution**

Let $X$ = the number of calls Leah receives in 15 minutes. (The *interval of interest* is 15 minutes or $\frac{1}{4}$ hour.)

$$x = 0, 1, 2, 3, ... \tag{5.7.3}$$

If Leah receives, on the average, six telephone calls in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)(6) = 0.75$ calls in 15 minutes, on average. So, $\mu = 0.75$ for this problem.

$X \sim P(0.75)$

Find $P(x > 1)$. $P(x > 1) = 0.1734$ (calculator or computer)

- Press 1 – and then press 2nd DISTR.
- Arrow down to poissoncdf. Press ENTER.
- Enter (.75,1).
- The result is $P(x > 1) = 0.1734$.

> The TI calculators use $\lambda$ (lambda) for the mean.

The probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734:

$P(x > 1) = 1 - \text{poissoncdf}(0.75, 1)$.

The graph of $X \sim P(0.75)$ is:

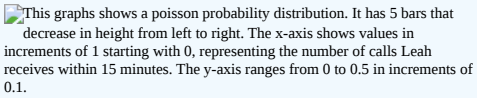



Figure 5.7.1

The *y*-axis contains the probability of $x$ where $X =$ the number of calls in 15 minutes.

### Example 5.7.5

According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let $X =$ the number of emails an email user receives per day. The discrete random variable $X$ takes on the values $x = 0, 1, 2 ....$ The random variable $X$ has a Poisson distribution: $X \sim P(147)$. The mean is 147 emails.

a. What is the probability that an email user receives exactly 160 emails per day?
b. What is the probability that an email user receives at most 160 emails per day?
c. What is the standard deviation?

**Solutions**

a. $P(x = 160) = \text{poissonpdf}(147, 160) \approx 0.0180$
b. $P(x \leq 160) = \text{poissoncdf}(147, 160) \approx 0.8666$
c. Standard Deviation $= \sigma = \sqrt{\mu} = \sqrt{147} \approx 12.1244$

**Example 5.7.6**

Text message users receive or send an average of 41.5 text messages per day.

a. How many text messages does a text message user receive or send per hour?
b. What is the probability that a text message user receives or sends two messages per hour?
c. What is the probability that a text message user receives or sends more than two messages per hour?

**Solutions**

a. Let $X =$ the number of texts that a user sends or receives in one hour. The average number of texts received per hour is $\frac{41.5}{24} \approx 1.7292$.
b. $X \sim P(1.7292)$, so $P(x = 2) = \text{poissonpdf}(1.7292, 2) \approx 0.2653$
c. $P(x > 2) = 1 – P(x \leq 2) = 1 – \text{poissoncdf}(1.7292, 2) \approx 1 – 0.7495 = 0.2505$

The Poisson distribution can be used to approximate probabilities for a binomial distribution. This next example demonstrates the relationship between the Poisson and the binomial distributions. Let $n$ represent the number of binomial trials and let $p$ represent the probability of a success for each trial. If $n$ is large enough and $p$ is small enough then the Poisson approximates the binomial very well. In general, $n$ is considered "large enough" if it is greater than or equal to 20. The probability $p$ from the binomial distribution should be less than or equal to 0.05. When the Poisson is used to approximate the binomial, we use the binomial mean $\mu = np$. The variance of $X$ is $\sigma^2 = \sqrt{\mu}$ and the standard deviation is $\sigma = \sqrt{\mu}$. The Poisson approximation to a binomial distribution was commonly used in the days before technology made both values very easy to calculate.

**Example 5.7.7**

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

**Answer**

Let $X$ = the number of days with low seismic activity.

Using the binomial distribution:

- $P(x = 10) = \text{binompdf}(200, .0102, 10) \approx 0.000039$

Using the Poisson distribution:

- Calculate $\mu = np = 200(0.0102) \approx 2.04$
- $P(x = 10) = \text{poissonpdf}(2.04, 10) \approx 0.000045$

We expect the approximation to be good because $n$ is large (greater than 20) and $p$ is small (less than 0.05). The results are close—both probabilities reported are almost 0.

## WeBWorK Problems

## References


1. "ATL Fact Sheet," Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at www.atl.com/about-atl/atl-factsheet/ (accessed May 15, 2013).
2. Center for Disease Control and Prevention. "Teen Drivers: Fact Sheet," Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafet...factsheet.html (accessed May 15, 2013).
3. "Children and Childrearing," Ministry of Health, Labour, and Welfare. Available online at http://www.mhlw.go.jp/english/policy...ing/index.html (accessed May 15, 2013).

4. “Eating Disorder Statistics,” South Carolina Department of Mental Health, 2006. Available online at http://www.state.sc.us/dmh/anorexia/statistics.htm (accessed May 15, 2013).
5. “Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day,” theguardian, 2013. Available online at www.theguardian.com/world/gal...471900&index=2 (accessed May 15, 2013).
6. “How Americans Use Text Messaging,” Pew Internet, 2013. Available online at pewinternet.org/Reports/2011/...in-Report.aspx (accessed May 15, 2013).
7. Lenhart, Amanda. “Teens, Smartphones & Testing: Texting volum is up while the frequency of voice calling is down. About one in four teens say they own smartphones,” Pew Internet, 2012. Available online at www.pewinternet.org/~/media/F...nd_Texting.pdf (accessed May 15, 2013).
8. “One born every minute: the maternity unit where mothers are THREE to a bed,” MailOnline. Available online at http://www.dailymail.co.uk/news/arti...thers-bed.html (accessed May 15, 2013).
9. Vanderkam, Laura. “Stop Checking Your Email, Now.” CNNMoney, 2013. Available online at management.fortune.cnn.com/20...our-email-now/ (accessed May 15, 2013).
10. “World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. www.world-earthquakes.com/ind...thq_prediction (accessed May 15, 2013).

## Review

A Poisson probability distribution of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is "small" (less than or equal to 0.05) and the number of trials is "large" (greater than or equal to 20).

## Formula Review

$X \sim P(\mu)$ means that $X$ has a Poisson probability distribution where $X =$ the number of occurrences in the interval of interest.

$X$ takes on the values $x = 0, 1, 2, 3, \ldots$

The mean $\mu$ is typically given.

The variance is $\sigma = \mu$, and the standard deviation is

$$\sigma = \sqrt{\mu} \tag{5.7.4}$$

.

When $P(\mu)$ is used to approximate a binomial distribution, $\mu = np$ where $n$ represents the number of independent trials and $p$ represents the probability of success in a single trial.

## 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.

*Use the following information to answer the next six exercises:* On average, a clothing store gets 120 customers per day.

## Glossary

**Poisson Probability Distribution**

a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

- The probability that the event occurs in a given interval is the same for all intervals.
- The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean $\mu$ of the event in the interval. Notation: $X \sim P(\mu)$. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly $x$ successes in $r$ trials is $P(X = x) =$

$$(e^{-\mu}) \frac{\mu^x}{x!} \tag{5.7.5}$$

. The Poisson distribution is often used to approximate the binomial distribution, when $n$ is “large” and $p$ is “small” (a general rule is that $n$ should be greater than or equal to 20 and $p$ should be less than or equal to 0.05).

---

This page titled 5.7: The Poisson 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.

- **4.7: Poisson Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 6: Continuous Random Variables

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: Probability Density Functions

6.2: The Uniform and Other Simple Continuous Distributions

6.3: The Standard Normal Distribution

6.4: Applications of Finding Normal Probabilities


### 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 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: Probability Density 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**

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

$$f(x) = \frac{1}{20} \text{ for } 0 \leq x \leq 20. \tag{6.1.1}$$

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 = $\frac{1}{20}$.

$$AREA = 20\left(\frac{1}{20}\right) = 1 \tag{6.1.2}$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the $x$-axis where $0 < x < 2$.




Figure 6.1.2

$$AREA = (2-0)\left(\frac{1}{20}\right) = 0.1 \tag{6.1.3}$$

$$(2-0) = 2 = \text{base of a rectangle}$$

*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 $4 < x < 15$.**

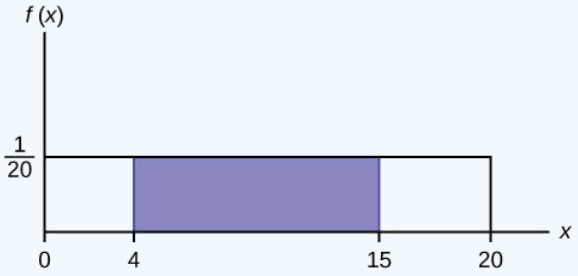



Figure 6.1.3

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$\text{AREA} = (15–4)(\frac{1}{20}) = 0.55$

$(15–4) = 11 = \text{the base of a rectangle} (15–4) = 11 = \text{the base of a rectangle}$

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, $P(x = 15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$

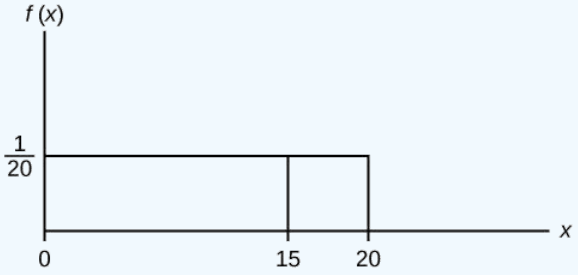



Figure 6.1.4

$P(X \leq x)$ (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: $P(X > x) = 1–P(X < x)$.

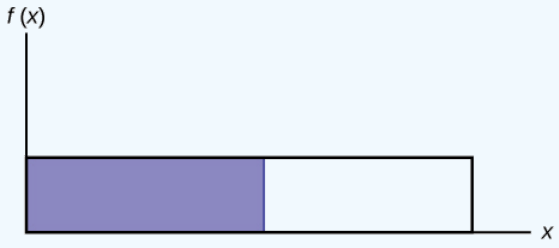



Figure 6.1.5

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.

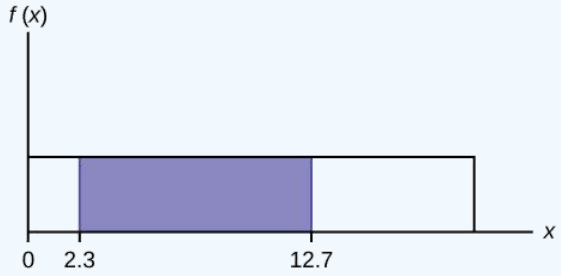



Figure 6.1.6

$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52 \quad (6.1.4)$$

## 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 $P(a < x < b)$.

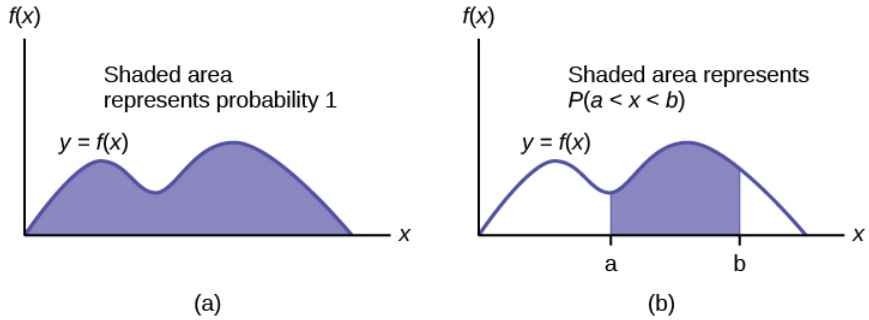



Figure 6.1.8

The cumulative distribution function (cdf) of $X$ is defined by $P(X \leq x)$. 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): $P(X \leq x)$

## WeBWorK Problems

## 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 6.1: Probability Density 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.

- **5.2: Continuous Probability Functions** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 6.2: The Uniform and Other Simple Continuous Distributions

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 6.2.1

The data in Table **6.2.1** are 55 smiling times, in seconds, of an eight-week-old baby.

Table **6.2.1**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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.

## Example 6.2.2

a. Refer to Example 6.2.1. What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

**Answer**

a. Find $P(2 < x < 18)$.

$P(2 < x < 18) = (\text{base})(\text{height}) = (18–2)\left(\frac{1}{23}\right) = \left(\frac{16}{23}\right)$.

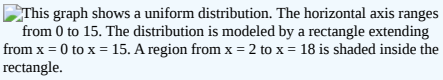



Figure 6.2.1

## Example 6.2.3

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

## Example 6.2.4

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 30$^{th}$ 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 $f(x) = 0.4$

$P(x > 2) = (\text{base})(\text{height}) = (4–2)(0.4) = 0.8$

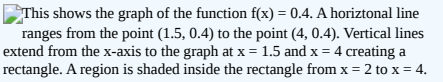



Figure 6.2.3. Uniform Distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time $x$ is greater than two

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.

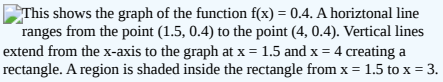



Figure 6.2.4. Uniform Distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time $x$ is less than three

c.

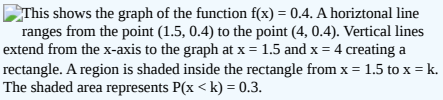



Figure 6.2.5. Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

$P(x < k) = 0.30$
$P(x < k) = (\text{base})(\text{height}) = (k–1.5)(0.4)$
$0.3 = (k–1.5)(0.4)$; 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.

Figure 6.2.6. 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

## 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.

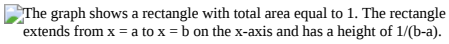



Figure 6.2.8.

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 $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d-c)\left(\frac{1}{b-a}\right)$

## WeBWorK Problems

## 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 $X \sim U(1.5, 4.5)$.

## Glossary

**Conditional Probability**

the likelihood that an event will occur given that another event has already occurred

## 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 6.2: The Uniform and Other Simple Continuous 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.

- **5.3: The Uniform Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 6.3: The Standard Normal Distribution

## Introduction to Normal Distributions

The normal distribution is the most important of all the probability 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.

In the remainder of this chapter, you will study the normal distributions and applications associated with them. A normal distribution has two parameters (two numerical descriptive measures), the mean ($\mu$) and the standard deviation ($\sigma$). If $X$ is a quantity to be measured that has a normal distribution with mean ($\mu$) and standard deviation ($\sigma$), we designate this by writing

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.3.1}$$

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.

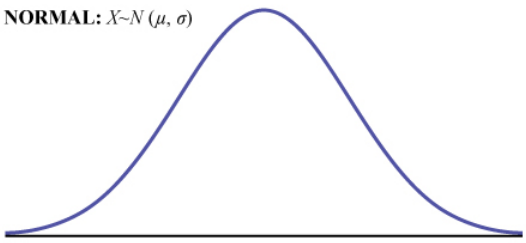



Figure **6.3.2**: The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, $\mu$. In theory, the mean is the same as the median, because the graph is symmetric about $\mu$. 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, $\sigma$, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on $\sigma$. A change in $\mu$ 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

## 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:

$$z = \frac{x - \mu}{\sigma} \tag{6.3.2}$$

**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, $\mu$.** 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 ??? produces the distribution $Z \sim N(0, 1)$. The value $x$ comes from a normal distribution with mean $\mu$ and standard deviation $\sigma$.

> *A z-score is measured in units of the standard deviation.*

**Example 6.3.1**

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 ???):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two** standard deviations ($2\sigma$) 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 ($-0.67\sigma$) 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 $\mu$ and when $z$ is negative, $x$ is to the left of or below $\mu$. Or, when $z$ is positive, $x$ is greater than $\mu$, and when $z$ is negative $x$ is less than $\mu$.

**Example 6.3.2**

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**.

## The Empirical Rule

If $X$ is a random variable and has a normal distribution with mean $\mu$ and standard deviation $\sigma$, then the *Empirical Rule* says the following:

- About 68% of the $x$ values lie between $-1\sigma$ and $+1\sigma$ of the mean $\mu$ (within one standard deviation of the mean).
- About 95% of the $x$ values lie between $-2\sigma$ and $+2\sigma$ of the mean $\mu$ (within two standard deviations of the mean).
- About 99.7% of the $x$ values lie between $-3\sigma$ and $+3\sigma$ of the mean $\mu$ (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 $-1\sigma$ are +1 and –1, respectively.
- The $z$-scores for $+2\sigma$ and $-2\sigma$ are +2 and –2, respectively.
- The $z$-scores for $+3\sigma$ and $-3\sigma$ are +3 and –3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.

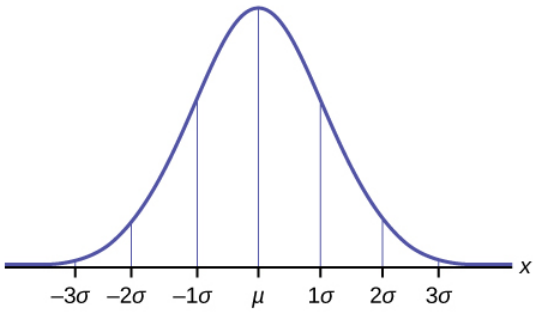



Figure **6.3.1**

**Example 6.3.3**

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

## Example 6.3.4

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 ???) 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.

## Example 6.3.5

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.

## Example 6.3.6

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.

## 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) $\mu$.

## Formula Review

$Z \sim N(0,1)$

$z = a$ standardized value ($z$-score)

mean = 0; standard deviation = 1

To find the $K^{th}$ 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

$Z \sim N(0,1)$

## WebWork Problems

## 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 $\mu$ and standard deviation $\sigma$, 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).

## 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 6.3: 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.
- **6.1: Prelude to The Normal Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 6.4: Applications of Finding Normal Probabilities

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 $P(X < x)$.

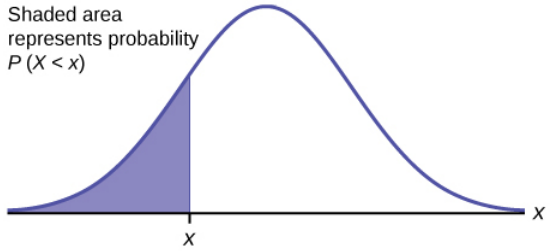



Figure 6.4.1.

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 involving normal distributions are usually calculated using technology such as calculators or spreadsheets. The can also be calculated using probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

**Example 6.4.2**

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 90$^{th}$ 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 70$^{th}$ 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 $P(x > 65)$.

$$P(x > 65) = 0.3446$$

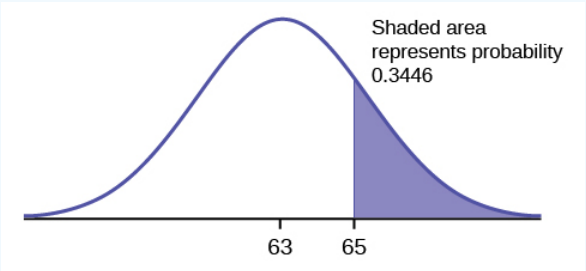



Figure 6.4.2.

The probability that any student selected at random scores more than 65 is 0.3446.

## Example 6.4.3

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 $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

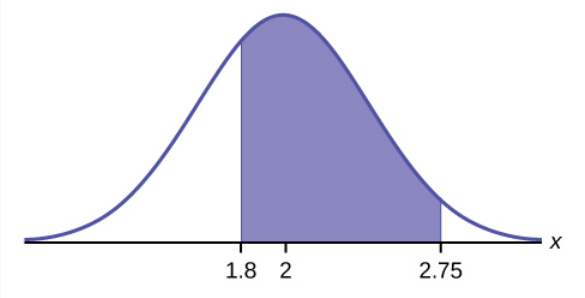



Figure 6.4.4.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

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 $P(x < k) = 0.25$.

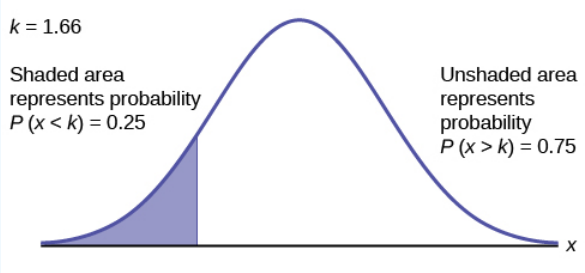



Figure 6.4.5.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

## Example 6.4.4

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. $\text{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$
b. $\text{normalcdf}(-10^{99}, 50.8, 36.9, 13.9) = 0.8413$
c. $\text{invNorm}(0.80, 36.9, 13.9) = 48.6$

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.

**Example 6.4.5**

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 $Q_3 =$ 75th percentile and $Q_1 =$ 25th percentile.

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.

## WeBWorK Problems

## 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 $\mu$ and the standard deviation $\sigma$. 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.

## 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 6.4: Applications of Finding Normal Probabilities 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.

# CHAPTER OVERVIEW

## 7: Sampling Distributions

In this chapter, you will study means and the **central limit theorem**, which is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size $n$ from a population with a known mean, $\mu$, and a known standard deviation, $\sigma$. The first alternative says that if we collect samples of size $n$ with a "large enough $n$," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size $n$ that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.


7.1: The Sample Mean and Sources of Error

7.2: The Sum Distribution



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: Sampling 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.

# 7.1: The Sample Mean and Sources of Error

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

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right). \tag{7.1.1}$$

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.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)} \tag{7.1.2}$$

- $\mu_x$ 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.

**Example 7.1.1**

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.

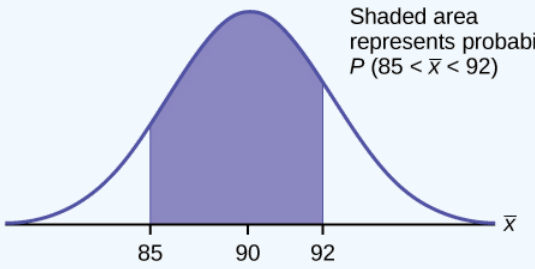



Figure 7.1.1.

normalcdf (lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, $\mu$, $\frac{\sigma}{\sqrt{n}}$)

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 $n$.

## Example 7.1.2

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.

## Example 7.1.3

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:

$$P(\text{X} > 30) = \text{normalcdf}(30, E99, 34, 1.5) = 0.9962$$

d. Let $k$ = the 95th percentile.

$$k = \text{invNorm}\left(0.95, 34, \frac{15}{\sqrt{100}}\right) = 36.5$$

## Exercise 7.1.3

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.

$$P(29 < \bar{x} < 35) = \text{normalcdf}\left(29, 35, 28, \frac{4.8}{\sqrt{100}}\right) = 0.0186 \quad (7.1.3)$$

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 7.1.4

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. $P(8 < \bar{x} < 8.5) = \text{normalcdf}\left(8, 8.5, 8.2, \frac{1}{\sqrt{60}}\right) = 0.9293$

## WeBWorK Problems

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 $\bar{X} : \sigma_x$
- 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 $\mu$ and known standard deviation, $\sigma$, 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 $\mu$ is the mean of the distribution and $\sigma$ 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 7.1: The Sample Mean and Sources of Error 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.

- **7.2: The Central Limit Theorem for Sample Means (Averages)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 7.2: The Sum Distribution

Suppose $X$ is a random variable with a distribution that may be known or unknown (it can be any distribution) and suppose:

- $\mu_x$ = the mean of $X$
- $\sigma_x$ = the standard deviation of $X$

If you draw random samples of size $n$, then as $n$ increases, the random variable $\sum X$ consisting of sums tends to be normally distributed and

$$\sum X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x)). \tag{7.2.1}$$

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.

The random variable $\sum X$ has the following $z$-score associated with it:

a. $\sum x$ is one sum.
b. $z = \frac{\sum x - (n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
   i. $(n)(\mu_x)$= the mean of $\sum X$
   ii. $(\sqrt{n})(\sigma_x)$= standard deviation of $\sum X$

**Example 7.2.1**

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500.
b. Find the sum that is 1.5 standard deviations above the mean of the sums.

**Answer**

Let $X =$ one value from the original unknown population. The probability question asks you to find a probability for the sum (or total of) 80 values.

$\sum X =$ the sum or total of 80 values. Since $\mu_x = 90$, $\sigma_x = 15$, and $n = 80$, $\sum X \sim N((80)(90), (\sqrt{80})(15))$

- mean of the sums $= (n)(\mu_x) = (80)(90) = 7,200$
- standard deviation of the sums $= (\sqrt{n})(\sigma_x) = (\sqrt{80})(15) = (80)(15)$
- sum of 80 values $= \sum X = 7,500$

a. Find $P(\sum X > 7,500)$

$P(\sum X > 7,500) = 0.0127$

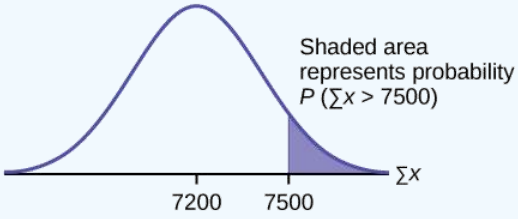



Figure 7.2.1.

`normalcdf` (lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated $(lower, upper, (n)(\mu_x, (\sqrt{n}(\sigma_x))$

`normalcdf` $(7500, 1E99, (80)(90), (\sqrt{80})(15)) = 0.0127$

REMINDER

**1E99 = $10^{99}$.**

Press the EE key for E.

b. Find $\sum x$ where $z = 1.5$.

$\sum x = (n)(\nu_x) + (z)(\sqrt{n})(\sigma_x) = (80)(90) + (1.5)(\sqrt{80})(15) = 7,401.2$

**Example 7.2.2**

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. The sample of size is 50.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
b. Find the probability that the sum of the ages is between 1,500 and 1,800 years.
c. Find the 80$^{th}$ percentile for the sum of the 50 ages.

**Answer**

a. $\mu_x - n\mu_x = 1,700$ and $\sigma_{\sum X} = \sqrt{n}\sigma_X = (\sqrt{50})(15) = 106.01$
The distribution is normal for sums by the central limit theorem.
b. $P(1500 < \sum X < 1800) = (1,500, 1,800, (50)(34), (\sqrt{50})(15)) = 0.7974$
c. Let $k$ = the 80$^{th}$ percentile.
$k = (0.80, (50)(34), (\sqrt{50})(15)) = 1,789.3$

**Example 7.2.3**

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 size 70.

a. What are the mean and standard deviation for the sums?
b. Find the 95$^{th}$ percentile for the sum of the sample. Interpret this value in a complete sentence.
c. Find the probability that the sum of the sample is at least ten hours.

**Answer**

a. $\mu_{\sum X} = n\mu_X = 70(8.2) = 574$ minutes and $\sigma_{\sum X}(\sqrt{n})(\sigma_x) = (\sqrt{70})(1) = 8.37$ minutes
b. Let $k$ = the 95$^{th}$ percentile.
$k = \text{invNorm}(0.95, (70)(8.2), (\sqrt{70})(1)) = 587.76$ minutes
Ninety five percent of the app engagement times are at most 587.76 minutes.
c. ten hours = 600 minutes
$P(\sum X \geq 600) = \text{normalcdf}(600, E99, (70)(8.2), (\sqrt{70})(1)) = 0.0009$

## WeBWorK Problems

## References

1. Farago, Peter. “The Truth About Cats and Dogs: Smartphone vs Tablet Usage Differences.” The Flurry Blog, 2013. Posted October 29, 2012. Available online at blog.flurry.com (accessed May 17, 2013).

## Review

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if

the original population has a mean of $\mu_x$ and a standard deviation of $\sigma_x$, the mean of the sums is $n\mu_x$ and the standard deviation is ( $\sqrt{n}$)($\sigma_x$) where $n$ is the sample size.

## Formula Review

- The Central Limit Theorem for Sums: $\sum X\ N[(n)(\mu_x, (\sqrt{n})(\sigma_x))]$
- Mean for Sums $(\sum X) : (n)(\mu_x)$
- The Central Limit Theorem for Sums $z$-score and standard deviation for sums: $z$ **for the sample mean** $= \frac{\sum x - (n)(\mu_x)}{(\sqrt{n})(\sigma_x)}$
- Standard deviation for Sums $(\sum X) : (\sqrt{n})(\sigma_x)$

---

This page titled 7.2: The Sum 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.

- **7.3: The Central Limit Theorem for Sums** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 8: Confidence Intervals

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.


8.1: Estimating Population Means

8.2: The t-distribution

8.3: Estimating Proportions

8.4: Confidence Intervals


### 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: 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: Estimating Population Means

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 $\mu$, where the population standard deviation is known, we need $\bar{x}$ as an estimate for $\mu$ 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 $\mu$.

The confidence interval estimate will have the form:

or, in symbols,

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 ($\alpha$). $\alpha$ is related to the confidence level, $CL$. $\alpha$ is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

**Example 8.1.1**

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."

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.

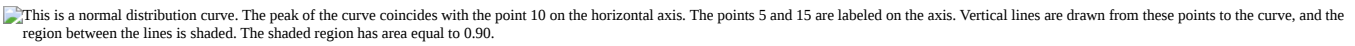

This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

Figure 8.1.1

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 $\sigma$.

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 $\sigma$ 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 $\sigma$.
- 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 $\sigma$, 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 $\alpha$ 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 $\mu$ when the population standard deviation $\sigma$ 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.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$

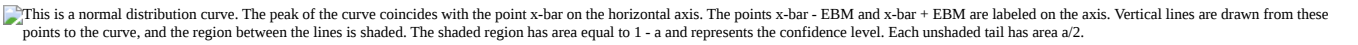

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has 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 8.1.2**

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.
- solution is shown step-by-step (Solution A).

**Solution**

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.

$$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225).**

**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.

## Example 8.1.3: 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**

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, $CL = 0.98$.

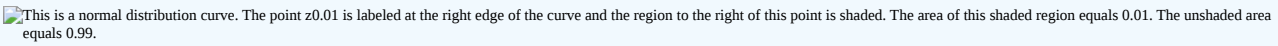



Figure 8.2.3.

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.

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 8.1.4**

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.

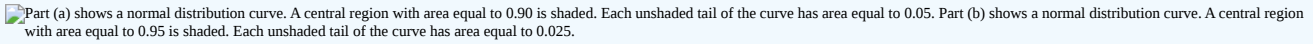

Part (a) shows a normal distribution curve. A central region with area equal to 0.90 is shaded. Each unshaded tail of the curve has area equal to 0.05. Part (b) shows a normal distribution curve. A central region with area equal to 0.95 is shaded. Each unshaded tail of the curve has area equal to 0.025.

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.

**Example 8.1.5**

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.

## 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 8.1.6**

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: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

## 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 8.1.7**

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.

## WeBWorK Problems

## 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.

## 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: Estimating Population 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.

- **8.2: 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.

# 8.2: The t-distribution

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 $\sigma$ 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 $\sigma$ 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 $\sigma$. If you draw a simple random sample of size $n$ from a population that has an approximately a normal distribution with mean $\mu$ and unknown population standard deviation $\sigma$ and calculate the $t$-score 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 $\mu$. 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.*

- 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 $\mu$ and unknown population standard deviation $\sigma$. 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. However for confidence intervals, we need to use **inverse** probability to find the value of $t$ when we know the probability.

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).

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.

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).

**Solution**

To find the confidence interval, you need the sample mean, $\bar{x}$, and the $EBM$.

$$\bar{x} = 8.2267$$

$$s = 1.6722\, n = 15$$

$$df = 15–1 = 14 CL so \alpha = 1–CL = 1–0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 t_{\frac{\alpha}{2}} = t_{0.025}$$

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 ???:

$$\bar{x}–EBM = 8.2267–0.9240 = 7.3$$
$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

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.

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.

**Example 8.2.2: 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 8.2.1 shows how many of the targeted chemicals were found in each infant's cord blood.

Table 8.2.1

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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**

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

$$\alpha = 1–CL = 1–0.90 = 0.10 \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} \tag{8.2.1}$$

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.

## Example 8.2.3

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 8.2.2. 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 8.2.2

| | | | | |
|---|---|---|---|---|
| 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
- $df = 15–1 = 14$.

$\frac{\alpha}{2} = 0.01 t_{\frac{\alpha}{2}} = t_{0.01} 2.624$

$\bar{x}–EBM = 6.133–3.736 = 2.397$

$$\bar{x} + EBM = 6.133 + 3.736 = 9.869$$

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.

## WeBWorK Problems

## 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 $\mu$ is the mean of the distribution and $\sigma$ 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 $\sigma$ 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.

## 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: The t-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.

- **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.

# 8.3: Estimating 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 $n$.)

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

$P'$ follows a normal distribution for proportions:

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

then the $z$-score formula is

$$z = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \tag{8.3.1}$$

### Example 8.3.1

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**

- The first solution is step-by-step (Solution A).

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 $= 421$


https://stats.libretexts.org/@go/page/26087

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

- $p' = 0.842$ 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.

**Exercise 8.3.1**

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 8.3.2**

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).

**Solution**

- $x = 300$ 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

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \left(\frac{\alpha}{2}\right) = 0.05 \tag{8.3.2}$$

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

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.

**Exercise 8.3.2**

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.

## "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 8.3.3**

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**

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.

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.

The confidence interval is (0.113, 0.439).

**Exercise 8.3.3**

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**

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 $\frac{\alpha}{2} = 0.02$.

$$z_{0.02} = 2.054$$

$$p' - EPB = 0.478 - 0.124 = 0.354$$

$$p' + EPB = 0.478 + 0.124 = 0.602$$

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

The confidence interval is (0.355, 0.602).

**Example 8.3.4**

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 8.3.4**

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**

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.

- 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 8.3.5**

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.

## 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 $P(X = x = \binom{n}{x}) p^x q^{n-x}$.

**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.

## 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: Estimating 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.

# 8.4: Confidence Intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter.

In this section and in Section 4.3, we will emphasize the special case where the point estimate is a sample mean and the parameter is the population mean. In Section 4.5, we generalize these methods for a variety of point estimates and population parameters that we will encounter in Chapter 5 and beyond.

## Capturing the population Parameter

A plausible range of values for the population parameter is called a **confidence interval**. Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like shing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values - a confidence interval - we have a good shot at capturing the parameter.

**Exercise 4.7**

If we want to be very certain we capture the population parameter, should we use a wider interval or a smaller interval?[8]

## An Approximate 95% confidence interval

Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate. The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

The standard error represents the standard deviation associated with the estimate, and roughly 95% of the time the estimate will be within 2 standard errors of the parameter. If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% con dent that we have captured the true parameter:

$$\text{point estimate} \pm 2 \times SE \tag{8.4.1}$$

But what does "95% confident" mean? Suppose we took many samples and built a confidence interval from each sample using Equation 8.4.1. Then about 95% of those intervals would contain the actual mean, $\mu$. Figure 4.8 shows this process with 25 samples, where 24 of the resulting confidence intervals contain the average time for all the runners, $\mu = 94.52$ minutes, and one does not.

**Exercise 4.9**

In Figure 4.8, one interval does not contain 94.52 minutes. Does this imply that the mean cannot be 94.52? [9]

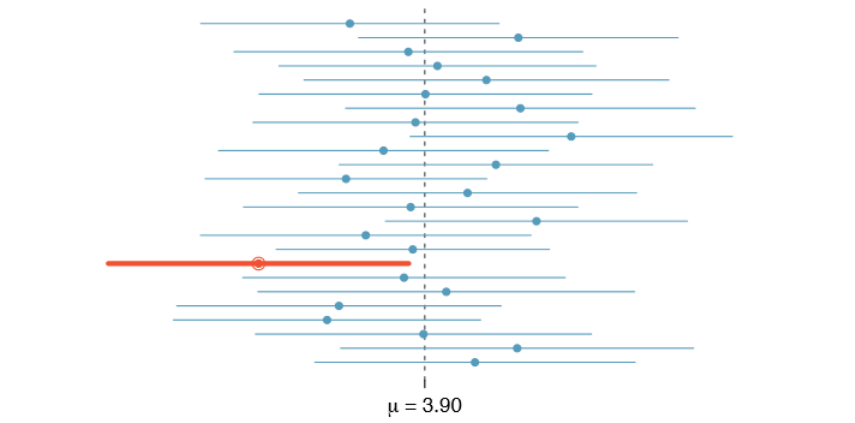



Figure 4.8: Twenty- ve samples of size n = 100 were taken from the run10 data set. For each sample, a confidence interval was created to try to capture the average 10 mile time for the population. Only 1 of these 25 intervals did not capture the true mean, $\mu = 94.52$ minutes.

[8]*If we want to be more certain we will capture the sh, we might use a wider net. Likewise, we use a wider confidence interval if we want to be more certain that we capture the parameter.*

[9]*Just as some observations occur more than 2 standard deviations from the mean, some point estimates will be more than 2 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.*

The rule where about 95% of observations are within 2 standard deviations of the mean is only approximately true. However, it holds very well for the normal distribution. As we will soon see, the mean tends to be normally distributed when the sample size is sufficiently large.

**Example 4.10**

If the sample mean of times from run10Samp is 95.61 minutes and the standard error, as estimated using the sample standard deviation, is 1.58 minutes, what would be an approximate 95% confidence interval for the average 10 mile time of all runners in the race? Apply the standard error calculated using the sample standard deviation ($SE = \frac{15.78}{\sqrt{100}} = 1:58$), which is how we usually proceed since the population standard deviation is generally unknown.

**Solution**

We apply Equation 8.4.1:

$$95.61 \pm 2 \times 1.58 \rightarrow (92.45, 98.77) \tag{8.4.2}$$

Based on these data, we are about 95% con dent that the average 10 mile time for all runners in the race was larger than 92.45 but less than 98.77 minutes. Our interval extends out 2 standard errors from the point estimate, $\bar{x}$.

**Exercise 4.11**

he sample data suggest the average runner's age is about 35.05 years with a standard error of 0.90 years (estimated using the sample standard deviation, 8.97). What is an approximate 95% confidence interval for the average age of all of the runners?[10]

[10]*Again apply Equation 8.4.1:* $35.05 \pm 2 \times 0.90 \rightarrow (33.25, 36.85)$. *We interpret this interval as follows: We are about 95% con dent the average age of all participants in the 2012 Cherry Blossom Run was between 33.25 and 36.85 years.*

## A Sampling Distribution for the Mean

In Section 4.1.3, we introduced a sampling distribution for $\bar{x}$, the average run time for samples of size 100. We examined this distribution earlier in Figure 4.7. Now we'll take 100,000 samples, calculate the mean of each, and plot them in a histogram to get an especially accurate depiction of the sampling distribution. This histogram is shown in the left panel of Figure 4.9.

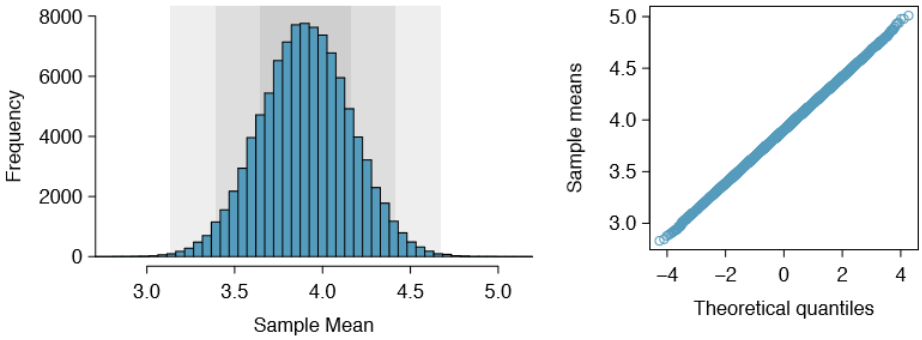



Figure 4.9: The left panel shows a histogram of the sample means for 100,000 different random samples. The right panel shows a normal probability plot of those sample means.

Does this distribution look familiar? Hopefully so! The distribution of sample means closely resembles the normal distribution (see Section 3.1). A normal probability plot of these sample means is shown in the right panel of Figure 4.9. Because all of the points closely fall around a straight line, we can conclude the distribution of sample means is nearly normal. This result can be explained by the Central Limit Theorem.

**Central Limit Theorem, informal description**

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

We will apply this informal version of the Central Limit Theorem for now, and discuss its details further in Section 4.4.

The choice of using 2 standard errors in Equation8.4.1 was based on our general guideline that roughly 95% of the time, observations are within two standard deviations of the mean. Under the normal model, we can make this more accurate by using 1.96 in place of 2.

$$\text{point estimate} \pm 1.96 \times SE \tag{8.4.3}$$

If a point estimate, such as $\bar{x}$, is associated with a normal model and standard error SE, then we use this more precise 95% confidence interval.

## Changing the confidence level

Suppose we want to consider confidence intervals where the confidence level is somewhat higher than 95%: perhaps we would like a confidence level of 99%. Think back to the analogy about trying to catch a sh: if we want to be more sure that we will catch the fish, we should use a wider net. To create a 99% confidence level, we must also widen our 95% interval. On the other hand, if we want an interval with lower confidence, such as 90%, we could make our original 95% interval slightly slimmer.

The 95% confidence interval structure provides guidance in how to make intervals with new confidence levels. Below is a general 95% confidence interval for a point estimate that comes from a nearly normal distribution:

$$\text{point estimate} \pm 1.96 \times SE \tag{8.4.4}$$

There are three components to this interval: the point estimate, "1.96", and the standard error. The choice of 1:96 SE was based on capturing 95% of the data since the estimate is within 1.96 standard deviations of the parameter about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

**Exercise 4.14**

If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[11]

To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2:58. Exercise 4.14 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean. This approach - using the Z scores

in the normal model to compute confidence levels - is appropriate when $\bar{x}$ is associated with a normal distribution with mean $\mu$ and standard deviation $SE_{\bar{x}}$. Thus, the formula for a 99% confidence interval is

$$\bar{x} \pm 2.58 \times SE\bar{x} \tag{8.4.5}$$

The normal approximation is crucial to the precision of these confidence intervals. Section 4.4 provides a more detailed discussion about when the normal model can safely be applied. When the normal model is not a good fit, we will use alternative distributions that better characterize the sampling distribution.

**Conditions for $\bar{x}$ being nearly normal and SE being accurate**

Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The distribution of sample observations is not strongly skewed.

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

[11] *This is equivalent to asking how often the Z score will be larger than -2.58 but less than 2.58. (For a picture, see Figure 4.10.) To determine this probability, look up -2.58 and 2.58 in the normal probability table (0.0049 and 0.9951). Thus, there is a* $0.9951 - 0.0049 \approx 0.99$ *probability that the unobserved random variable X will be within 2.58 standard deviations of* $\mu$.

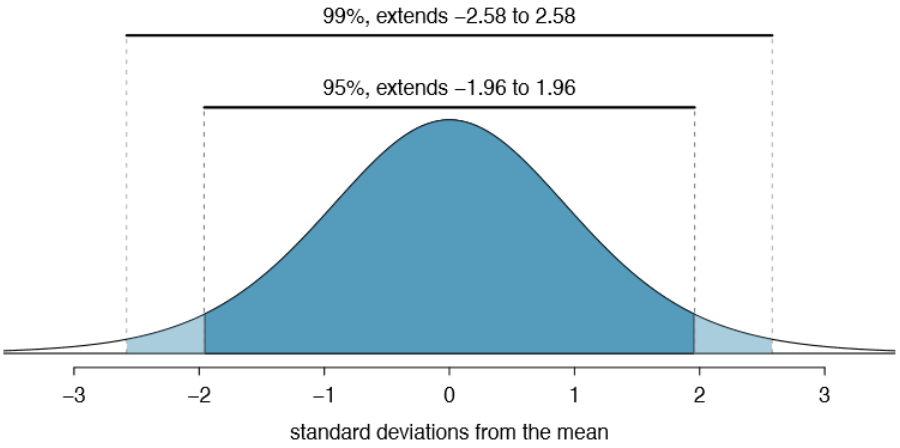



Figure 4.10: The area between -z* and z* increases as |z*| becomes larger. If the confidence level is 99%, we choose z* such that 99% of the normal curve is between -z* and z*, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: z* = 2:58.

Verifying independence is often the most difficult of the conditions to check, and the way to check for independence varies from one situation to another. However, we can provide simple rules for the most common scenarios.

**TIP: How to verify sample observations are independent**

Observations in a simple random sample consisting of less than 10% of the population are independent.

**Caution: Independence for random processes and experiments**

If a sample is from a random process or experiment, it is important to verify the observations from the process or subjects in the experiment are nearly independent and maintain their independence throughout the process or experiment. Usually subjects are considered independent if they undergo random assignment in an experiment.

**Exercise 4.16**

Create a 99% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run. The point estimate is $\bar{y} = 35.05$ and the standard error is $SE_{\bar{y}} = 0.90$.[12]

[12]*The observations are independent (simple random sample, < 10% of the population), the sample size is at least 30 (n = 100), and the distribution is only slightly skewed (Figure 4.4); the normal approximation and estimate of SE should be reasonable. Apply the 99% confidence interval formula:* $\bar{y} \pm 2.58 \times SE_{\bar{y}} \rightarrow (32.7, 37.4)$. *We are 99% confident that the average age of all runners is between 32.7 and 37.4 years.*

**Confidence interval for any confidence level**

If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z * SE \tag{8.4.6}$$

where z* corresponds to the confidence level selected.

Figure 4.10 provides a picture of how to identify z* based on a confidence level. We select z* so that the area between -z* and z* in the normal model corresponds to the confidence level.

**Margin of error**

In a confidence interval, z* SE is called the **margin of error**.

**Exercise 4.17** Use the data in Exercise 4.16 to create a 90% confidence interval for the average age of all runners in the 2012 Cherry Blossom Run.[13]

## Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

$$\text{We are XX ...} \tag{8.4.7}$$

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability. This is one of the most common errors: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.

Another especially important consideration of confidence intervals is that they only try to capture the population parameter. Our intervals say nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates. confidence intervals only attempt to capture population parameters.

## Nearly normal population with known SD (special topic)

In rare circumstances we know important characteristics of a population. For instance, we might know a population is nearly normal and we may also know its parameter values. Even so, we may still like to study characteristics of a random sample from the population. Consider the conditions required for modeling a sample mean using the normal distribution:

1. The observations are independent.
2. The sample size n is at least 30.
3. The data distribution is not strongly skewed.

[13]We first find z* such that 90% of the distribution falls between -z* and z* in the standard normal model, $N(mu = 0, \sigma = 1)$. We can look up -z* in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail), thus z* = 1.65. The 90% confidence interval can then be computed as $\bar{y} \pm 1.65 \times SE_{\bar{y}} \rightarrow (33.6, 36.5)$. (We had already verified conditions for normality and the standard error.) That is, we are 90% con dent the average age is larger than 33.6 but less than 36.5 years.

These conditions are required so we can adequately estimate the standard deviation and so we can ensure the distribution of sample means is nearly normal. However, if the population is known to be nearly normal, the sample mean is always nearly normal (this is a special case of the Central Limit Theorem). If the standard deviation is also known, then conditions (2) and (3) are not necessary for those data.

Example

**Example 4.18** The heights of male seniors in high school closely follow a normal distribution $N(\mu = 70.43, \sigma = 2.73)$, where the units are inches.[14] If we randomly sampled the heights of ve male seniors, what distribution should the sample mean follow?

**Solution**

The population is nearly normal, the population standard deviation is known, and the heights represent a random sample from a much larger population, satisfying the independence condition. Therefore the sample mean of the heights will follow a nearly normal distribution with mean $\mu = 70.43$ inches and standard error $SE = \frac{\sigma}{\sqrt{n}} = \frac{2.73}{\sqrt{5}} = 1.22$ inches.

Alternative conditions for applying the normal distribution to model the sample mean

If the population of cases is known to be nearly normal and the population standard deviation $\sigma$ is known, then the sample mean $\bar{x}$ will follow a nearly normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$ if the sampled observations are also independent.

Sometimes the mean changes over time but the standard deviation remains the same. In such cases, a sample mean of small but nearly normal observations paired with a known standard deviation can be used to produce a confidence interval for the current population mean using the normal distribution.

Example 4.19

Is there a connection between height and popularity in high school? Many students may suspect as much, but what do the data say? Suppose the top 5 nominees for prom king at a high school have an average height of 71.8 inches. Does this provide strong evidence that these seniors' heights are not representative of all male seniors at their high school?

**Solution**

If these five seniors are height-representative, then their heights should be like a random sample from the distribution given in Example 4.18, $N(\mu = 70.43, \sigma = 2.73)$, and the sample mean should follow $N(\mu = 70.43, \frac{\sigma}{\sqrt{n}} = 1.22)$. Formally we are conducting what is called a hypothesis test, which we will discuss in greater detail during the next section. We are weighing two possibilities:

- $\mathbf{H_0}$: The prom king nominee heights are representative; $\bar{x}$ will follow a normal distribution with mean 70.43 inches and standard error 1.22 inches.
- $\mathbf{H_A}$: The heights are not representative; we suspect the mean height is different from 70.43 inches.

If there is strong evidence that the sample mean is not from the normal distribution provided in $H_0$, then that suggests the heights of prom king nominees are not a simple random sample (i.e. $H_A$ is true). We can look at the Z score of the sample mean to tell us how unusual our sample is. If $H_0$ is true:

[14] *These values were computed using the USDA Food Commodity Intake Database.*

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{71.8 - 70.43}{1.22} = 1.12 \tag{8.4.8}$$

A Z score of just 1.12 is not very unusual (we typically use a threshold of $\pm 2$ to decide what is unusual), so there is not strong evidence against the claim that the heights are representative. This does not mean the heights are actually representative, only that this very small sample does not necessarily show otherwise.

**TIP: Relaxing the nearly normal condition**

As the sample size becomes larger, it is reasonable to slowly relax the nearly normal assumption on the data when dealing with small samples. By the time the sample size reaches 30, the data must show strong skew for us to be concerned about the normality of the sampling distribution.

---

This page titled 8.4: Confidence Intervals 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.

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

# CHAPTER OVERVIEW

## 9: Hypothesis Testing for a Single Variable and Population

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.


9.1: Hypothesis Tests- An Introduction

9.2: Type I and Type II Errors

9.3: Hypothesis Tests about μ- p-value Approach

9.4: Hypothesis Tests about μ- Critical Region Approach

9.5: Hypothesis Tests for a Proportion


### 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: Hypothesis Testing for a Single Variable and Population 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: Hypothesis Tests- An Introduction

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 **9.1.1**: 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 9.1.1**

- $H_0$: 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. $p > 30$

**Example 9.1.2**

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:

- $H_0 : \mu = 2.0$
- $H_a : \mu \neq 2.0$

**Example 9.1.3**

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:

- $H_0 : \mu \geq 66$
- $H_a : \mu < 66$

**Example 9.1.4**

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.

- $H_0 : p \leq 0.066$
- $H_a : p > 0.066$

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 $H_0$.

$\alpha$ 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.

## WeBWorK Problems

## 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 $H_a$).

## 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 9.1: Hypothesis Tests- An 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.

- **9.2: Null and Alternative Hypotheses** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 9.2: 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 $\alpha$ and $\beta$ represent the probabilities.

- $\alpha =$ probability of a Type I error $= P(\text{Type I error}) =$ probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta =$ probability of a Type II error $= P(\text{Type II error}) =$ probability of not rejecting the null hypothesis when the null hypothesis is false.

$\alpha$ and $\beta$ 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 9.2.1: 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.

$\alpha =$ **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.

$\beta =$ **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.)

**Example 9.2.2**

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.

$\alpha =$ **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive $= P(\text{Type I error})$.

$\beta =$ **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead $= P(\text{Type II error})$.

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.)

**Example 9.2.3**

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, $\alpha$.
- **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, $\beta$.

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.

**Example 9.2.4**

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.

## 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 $\alpha$ and $\beta$, 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

- $\alpha$ = probability of a Type I error $= P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.
- $\beta$ = probability of a Type II error $= P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

## WeBWorK Problems

## 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.

## 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 9.2: 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.

- **9.3: Outcomes and the Type I and Type II Errors** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 9.3: Hypothesis Tests about μ- p-value Approach

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*:

$$\bar{X} - N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{9.3.1}$$

or

$$t_{df} \tag{9.3.2}$$

The population parameter is $\mu$. The estimated value (point estimate) for $\mu$ is $\bar{x}$, the sample mean.

If you are testing a single population proportion, the distribution for the test is for proportions or percentages:

$$P' - N\left(p, \sqrt{\frac{p-q}{n}}\right) \tag{9.3.3}$$

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** $\mu$ 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** $\mu$ 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 $\alpha$, 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 $\mu$ is the mean of the distribution, and $\sigma$ 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 $\sigma$ 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.

## 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 9.3: Hypothesis Tests about μ- p-value Approach 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.

# 9.4: Hypothesis Tests about μ- Critical Region Approach

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 9.4.1

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 $\sigma$ **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.

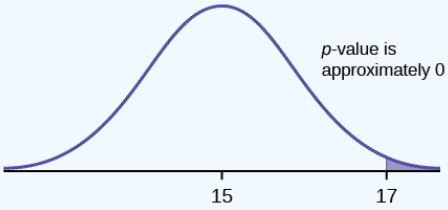



Figure **9.4.1**

$p$-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 9.4.1

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 $= 0.0013$

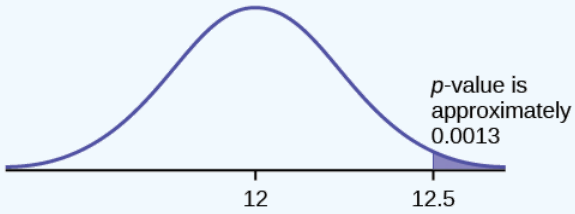



Figure **9.4.2**

## 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 $\alpha$ (also called a "**significance level**"). A preset $\alpha$ 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 9.4.2**

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**.

## 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

## WeBWorK Problems

## Glossary

**Level of Significance of the Test**

probability of a Type I error (reject the null hypothesis when it is true). Notation: $\alpha$. In hypothesis testing, the Level of Significance is called the preconceived $\alpha$ or the preset $\alpha$.

**$p$-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.

## 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 9.4: Hypothesis Tests about μ- Critical Region Approach 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.

- **9.5: Rare Events, the Sample, Decision and Conclusion** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 9.5: Hypothesis Tests for a Proportion

## WeBWorK Problems

This page titled 9.5: Hypothesis Tests for a Proportion 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

## 10: Hypothesis Testing for Paired and Unpaired Data

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.


10.1: Two Population Means

10.2: Two Independent Population Proportions

10.3: Matched or Paired Samples

10.4: Two Population Means with Known Standard Deviations

10.5: Difference of Two Means


### 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: Hypothesis Testing for Paired and Unpaired 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.

# 10.1: Two Population Means

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:
   - if the sample sizes are small, the distributions are important (should be normal)
   - if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch $t$-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\bar{X}_1 - \bar{X}_2$, and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\bar{X}_1 - \bar{X}_2$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{10.1.1}$$

The test statistic ($t$-score) is calculated as follows:

$$\frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.1.2}$$

where:

- $s_1$ and $s_2$, the sample standard deviations, are estimates of $\sigma_1$ and $\sigma_1$, respectively.
- $\sigma_1$ and $\sigma_2$ are the unknown population standard deviations.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means. $\mu_1$ and $\mu_2$ are the population means.

The number of *degrees of freedom* ($df$) requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The $df$ are not always a whole number. The test statistic calculated previously is approximated by the Student's $t$-distribution with $df$ as follows:

**Degrees of freedom**

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{10.1.3}$$

When both sample sizes $n_1$ and $n_2$ are five or larger, the Student's $t$ approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

It is not necessary to compute the degrees of freedom by hand. A calculator or computer easily computes it.

## Example 10.1.1: Independent groups

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in Table **10.1.1**. Each populations has a normal distribution.

Table **10.1.1**

| | Sample Size | Average Number of Hours Playing Sports Per Day | Sample Standard Deviation |
|---|---|---|---|
| Girls | 9 | 2 | 0.8660.866 |
| Boys | 16 | 3.2 | 1.00 |

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

**Answer**

The population standard deviations are not known. Let $g$ be the subscript for girls and $b$ be the subscript for boys. Then, $\mu_g$ is the population mean for girls and $\mu_b$ is the population mean for boys. This is a test of two independent groups, two population means.

Random variable: $\bar{X}_g - \bar{X}_b =$ difference in the sample mean amount of time girls and boys play sports each day.

- $H_0 : \mu_g = \mu_b$
- $H_0 : \mu_g - \mu_b = 0$
- $H_a : \mu_g \neq \mu_b$
- $H_a : \mu_g - \mu_b \neq 0$

The words "**the same**" tell you $H_0$ has an "=". Since there are no other words to indicate $H_a$, assume it says "**is different.**" This is a two-tailed test.

**Distribution for the test:** Use $t_{df}$ where $df$ is calculated using the $df$ formula for independent groups, two population means. Using a calculator, $df$ is approximately 18.8462. **Do not pool the variances.**

**Calculate the $p$-value using a Student's $t$-distribution:** $p\text{-value} = 0.0054$

**Graph:**

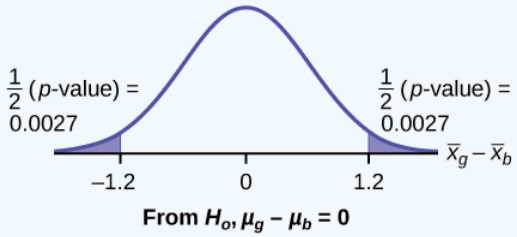



Figure **10.1.1**: Normal distribution curve representing the difference in the average amount of time girls and boys play sports all day

$$s_g = 0.866 \tag{10.1.4}$$

$$s_b = 1 \tag{10.1.5}$$

So,

Half the $p$-value is below –1.2 and half is above 1.2.

**Make a decision:** Since $\alpha > p\text{-value}$, reject $H_0$. This means you reject $\mu_g = \mu_b$. The means are different.

Press STAT . Arrow over to TESTS and press 4:2-SampTTest . Arrow over to Stats and press ENTER . Arrow down and enter 2 for the first sample mean, $\sqrt{0.866}$ for Sx1, 9 for n1, 3.2 for the second sample mean, 1 for Sx2, and 16 for n2. Arrow down to μ1: and arrow to does not equal μ2. Press ENTER . Arrow down to Pooled: and No . Press ENTER . Arrow down to Calculate and press ENTER . The $p$-value is $p = 0.0054$, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

**Conclusion:** At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different (mean number of hours boys aged seven to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged seven to 11 play sports per day is greater than the mean number of hours played by boys).

Two samples are shown in Table. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

Table **10.1.2**

| | Sample Size | Sample Mean | Sample Standard Deviation |
|---|---|---|---|
| Population A | 25 | 5 | 1 |
| Population B | 16 | 4.7 | 1.2 |

**Answer**

The $p$-value is 0.4125, which is much higher than 0.05, so we decline to reject the null hypothesis. There is not sufficient evidence to conclude that the means of the two populations are not the same.

When the sum of the sample sizes is larger than $30 (n_1 + n_2 > 30)$ you can use the normal distribution to approximate the Student's $t$.

**Example 10.1.2**

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes,** on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?
b. Are the populations standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the $p$-value?
h. Do you reject or not reject the null hypothesis?

**Solutions**

a. two means
b. unknown
c. Student's $t$
d. $\bar{X}_A - \bar{X}_B$
e. $H_0 : \mu_A \leq \mu_B$ and $H_a : \mu_A > \mu_B$

f.

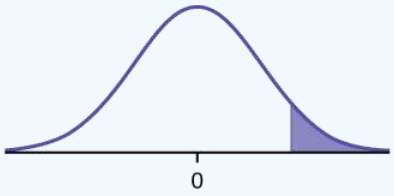



***Figure* 10.1.2.**

right

g. g. 0.1928
h. h. Do not reject.
i. i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

## Example 10.1.3

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in Table **10.1.3** and Table **10.1.4**.

Table **10.1.3**: Online Class

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 67.6 | 41.2 | 85.3 | 55.9 | 82.4 | 91.2 | 73.5 | 94.1 | 64.7 | 64.7 |
| 70.6 | 38.2 | 61.8 | 88.2 | 70.6 | 58.8 | 91.2 | 73.5 | 82.4 | 35.5 |
| 94.1 | 88.2 | 64.7 | 55.9 | 88.2 | 97.1 | 85.3 | 61.8 | 79.4 | 79.4 |

Table **10.1.4**: Face-to-face Class

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 77.9 | 95.3 | 81.2 | 74.1 | 98.8 | 88.2 | 85.9 | 92.9 | 87.1 | 88.2 |
| 69.4 | 57.6 | 69.4 | 67.1 | 97.6 | 85.9 | 88.2 | 91.8 | 78.8 | 71.8 |
| 98.8 | 61.2 | 92.9 | 90.6 | 97.6 | 100 | 95.3 | 83.5 | 92.9 | 89.4 |

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?
g. What is the ***p*-value**?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

(See the conclusion in Example, and write yours in a similar fashion)

*Be careful not to mix up the information for Group 1 and Group 2!*

**Answer**

a. two means
b. unknown
c. Student's $t$
d. $\bar{X}_1 - \bar{X}_2$
e. i. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
   ii. $H_a : \mu_1 < \mu_2$ Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
f. left-tailed
g. $p$-value $= 0.0011$

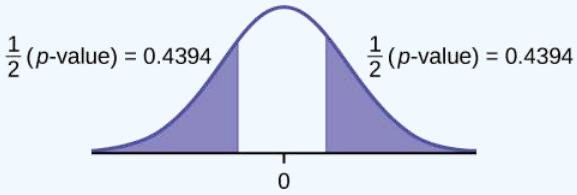



***Figure* 10.1.3.**

h. Reject the null hypothesis
i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.

   At the 5% level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to $\mu_1$: and arrow to $\neq \mu_1$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

## Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's $d$ is a measure of effect size based on the differences between two means. Cohen's $d$, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Table 10.1.5: Cohen's Standard Effect Sizes

| Size of effect | $d$ |
|---|---|
| Small | 0.2 |
| medium | 0.5 |
| Large | 0.8 |

Cohen's $d$ is the measure of the difference between two means divided by the pooled standard deviation: $d = \frac{\bar{x}_2 - \bar{x}_2}{s_{\text{pooled}}}$ where $s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

## Example 10.1.4

Calculate Cohen's $d$ for Example. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

**Answer**

$$\mu_1 = 4 s_1 = 1.5 n_1 = 11$$

$$\mu_2 = 3.5 s_2 = 1 n_2 = 9$$

$$d = 0.384$$

The effect is small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small indicating that there is not a significant difference between them.

## Example 10.1.5

Calculate Cohen's $d$ for Example. Is the size of the effect small, medium or large? Explain what the size of the effect means for this problem.

**Answer**

$d = 0.834$; Large, because 0.834 is greater than Cohen's 0.8 for a large effect size. The size of the differences between the means of the Final Exam scores of online students and students in a face-to-face class is large indicating a significant difference.

## Example 10.2.6

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013 are listed in Table and Table, respectively.

Northeast

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 94.2 | 75.2 | 69.6 | 52.0 | 48.0 | 41.9 | 36.4 | 33.4 | 31.5 | 27.6 |
| 77.3 | 71.9 | 67.5 | 50.6 | 46.2 | 38.4 | 35.2 | 33.0 | 28.7 | 26.5 |
| 76.3 | 71.7 | 56.3 | 48.7 | 43.2 | 37.6 | 33.7 | 31.8 | 28.5 | 26.0 |

West

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 126.0 | 70.6 | 65.2 | 51.4 | 45.5 | 37.0 | 33.0 | 29.6 | 23.7 | 22.6 |
| 116.1 | 70.6 | 58.2 | 51.2 | 43.2 | 36.0 | 31.4 | 28.7 | 23.5 | 21.6 |
| 78.2 | 68.2 | 55.6 | 50.3 | 39.0 | 34.1 | 31.0 | 25.3 | 23.4 | 21.5 |

Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?

g. What is the $p$-value?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
j. Calculate Cohen's $d$ and interpret it.

**Answer**

a. two means
b. unknown
c. Student's-t
d. $\bar{X}_1 - \bar{X}_2$
e. i. $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the weighted alphas are equal.
ii. $H_a : \mu_1 \neq \mu_2$ Alternative hypothesis : the means of the weighted alphas are not equal.
f. two-tailed
g. $p$-value $= 0.8787$
h. Do not reject the null hypothesis
i. This indicates that the trends in stocks are about the same in the top 30 banks in each region.

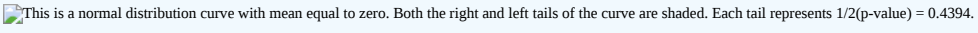



***Figure* 10.1.4.**

5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean weighted alphas for the banks in the northeast and the west are different

j. $d = 0.040$, Very small, because 0.040 is less than Cohen's value of 0.2 for small effect size. The size of the difference of the means of the weighted alphas for the two regions of banks is small indicating that there is not a significant difference between their trends in stocks.

## References


1. Data from Graduating Engineer + Computer Careers. Available online at www.graduatingengineer.com
2. Data from *Microsoft Bookshelf*.
3. Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).
4. "List of current United States Senators by Age." Wikipedia. Available online at en.Wikipedia.org/wiki/List_of...enators_by_age (accessed June 17, 2013).
5. "Sectoring by Industry Groups." Nasdaq. Available online at www.nasdaq.com/markets/barcha...&base=industry (accessed June 17, 2013).
6. "Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).
7. "World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).


## Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\bar{X}_1 - \bar{X}_2 =$ the difference of the sampling means
- Distribution: Student's $t$-distribution with degrees of freedom (variances not pooled)

## Formula Review

**Standard error:**

$$SE = \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}} \quad (10.1.6)$$

**Test statistic** ($t$-score):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{10.1.7}$$

**Degrees of freedom:**

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{10.1.8}$$

where:

- $s_1$ and $s_2$ are the sample standard deviations, and $n_1$ and $n_2$ are the sample sizes.
- $x_1$ and $x_2$ are the sample means.

**Cohen's $d$ is the measure of effect size:**

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{10.1.9}$$

where

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{10.1.10}$$

## WeBWorK Problems

## Glossary

**Degrees of Freedom ($df$)**

the number of objects in a sample that are free to vary.

**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 $\sigma$ for population standard deviation.

**Variable (Random Variable)**

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 $x, y, z,....$ 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 $x$ of the random variable $X$ takes only after performing the experiment.

## 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 10.1: Two Population 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.

- **10.2: Two Population Means with Unknown Standard Deviations** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 10.2: Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_0 : p_A = p_B$. To conduct the test, we use a pooled proportion, $p_c$.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{10.2.1}$$

The distribution for the differences is:

$$P\_A - P_b'N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.2.2}$$

The test statistic (z-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.2.3}$$

**Example 10.2.1**

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

**Answer**

The problem asks for a difference in proportions, making it a test of two proportions.

Let $A$ and $B$ be the subscripts for medication A and medication B, respectively. Then $p_A$ and $p_B$ are the desired population proportions.

Random Variable: $P'_A - P'_B =$ difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

$H_0 : p_A = p_B$

$p_A - p_B = 0$

$H_a : p_A \neq p_B$

$p_A - p_B \neq 0$

The words "**is a difference**" tell you the test is two-tailed.

**Distribution for the test:** Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 8001 - p_c = 0.92 \tag{10.2.4}$$

$$P'_A - P'_B - N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right] \tag{10.2.5}$$

$P'_A - P'_B$ follows an approximate normal distribution.

**Calculate the $p$-value using the normal distribution:** $p\text{-value} = 0.1404$.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

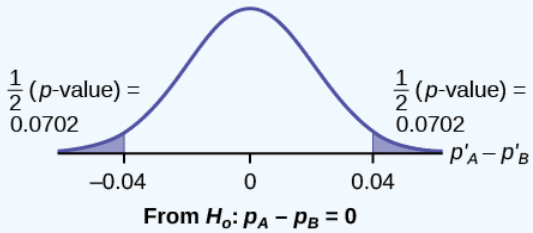



Figure 10.4.1.

$P'_A - P'_B = 0.1–0.06 = 0.04$.

Half the $p$-value is below -0.04, and half is above 0.04.

Compare $\alpha$ and the $p$-value : $\alpha = 0.01$ and the $p$-value $= 0.1404$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject $H_0$.

**Conclusion:** At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication *A* and medication *B*.

Press `STAT` . Arrow over to `TESTS` and press `6:2-PropZTest` . Arrow down and enter `20` for x1, `200` for n1, `12` for x2, and `200` for n2. Arrow down to `p1` : and arrow to `not equal p2` . Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The *p*-value is $p$ = 0.1404 and the test statistic is 1.47. Do the procedure again, but instead of `Calculate` do `Draw` .

## Exercise 10.2.1

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

**Answer**

The $p$-value is 0.0379, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

## Example 10.2.2: Sexting

A research study was conducted about gender differences in “sexting.” The researcher believed that the proportion of girls involved in “sexting” is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in Table. Is the proportion of girls sending sexts less than the proportion of boys “sexting?” Test at a 1% level of significance.

| | Males | Females |
|---|---|---|
| Sent “sexts” | 183 | 156 |
| Total number surveyed | 2231 | 2169 |

**Answer**

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then $p_M$ and $p_F$ are the desired population proportions.

Random variable: $p'_F - p'_M$ = difference in the proportions of males and females who sent “sexts.”

$H_a : p_F = p_m \quad H_0 : p_F - p_M = 0$

$H_a : p_F < p_m \quad H_a : p_F - p_M < 0$

The words **"less than"** tell you the test is left-tailed.

**Distribution for the test:** Since this is a test of two population proportions, the distribution is normal:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \tag{10.2.6}$$

$$1 - p_C = 0.923 \tag{10.2.7}$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right) \tag{10.2.8}$$

$p'_F - p'_M$ follows an approximate normal distribution.

**Calculate the $p$-value using the normal distribution:**

$p$-value $= 0.1045$

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082

Graph:

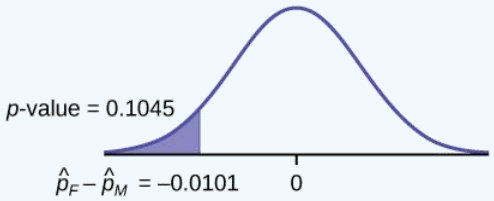



Figure 10.4.2.

**Decision:** Since $\alpha < p$-value, 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 proportion of girls sending “sexts” is less than the proportion of boys sending “sexts.”

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press ENTER . Arrow down to Calculate and press ENTER. The $p$-value is $P = 0.1045$ and the test statistic is $z = -1.256$.

Example 10.2.3

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

**Answer**

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then $p_W$ and $p_A$ are the desired population proportions.

Random variable: $p'_W - p'_A$ = difference in the proportions of Android and iPhone users.

$H_0 : p_W = p_A \quad H_0 : p_W - p_A = 0$

$H_a : p_W > p_A \quad H_a : p_W - p_A < 0$

The words "more popular" indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \tag{10.2.9}$$

$$1 - p_C = 0.9073 \tag{10.2.10}$$

Therefore,

$$p'_W - p'_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right) \tag{10.2.11}$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the $p$-value using the normal distribution:

$p$-value $= 0.0077$

Estimated proportion for group A: 0.10

Estimated proportion for group B: 0.05

**Graph:**

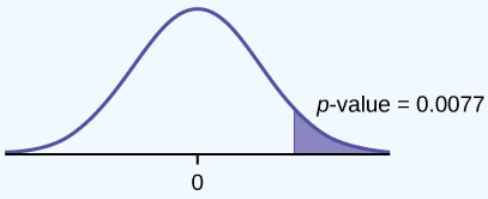



Figure 10.4.3.

**Decision:** Since $\alpha > p$-value, reject the $H_0$.

**Conclusion:** At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The P-value is P = 0.0092 and the test statistic is Z = 2.33.

Example 10.2.3

A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Which distribution do you use to perform the test?
c. What is the random variable?
d. What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.
e. Is this test right-, left-, or two-tailed?
f. What is the $p$-value?
g. Do you reject or not reject the null hypothesis?
h. At the ___ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

**Solutions**

a. two proportions

b. normal for two proportions

c. Subscripts: 1 = 2010, 2 = 2011 $P'_1 - P'_2$

d. Subscripts: 1 = 2010, 2 = 2011 $H_0 : p_1 = p_2 H_0 : p_1 - p_2 = 0\ H_0 : p_1 \neq p_2 H_0 : p_1 - p_2 \neq 0$

e. two-tailed

f. $p$-value $= 0.00086$

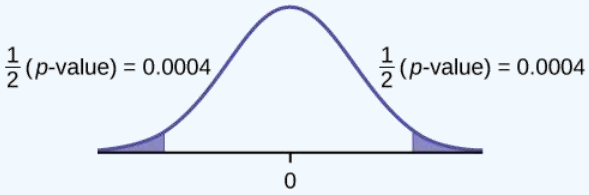



Figure 10.4.4.

g. Reject the $H_0$.

h. At the 5% significance level, from the sample data, there is sufficient evidence to conclude that there is a difference between the proportion of forcible rapes in 2011 and 2010.

## References


1. Data from *Educational Resources*, December catalog.
2. Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).
3. Data from Hyatt Hotels. Available online at hyatt.com (accessed June 17, 2013).
4. Data from Statistics, United States Department of Health and Human Services.
5. Data from Whitney Exhibit on loan to San Jose Museum of Art.
6. Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).
7. Data from the Chancellor's Office, California Community Colleges, November 1994.
8. "State of the States." Gallup, 2013. Available online at www.gallup.com/poll/125066/St...ef=interactive (accessed June 17, 2013).
9. "West Nile Virus." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

## Review

- Test of two population proportions from independent samples.
- Random variable: $\hat{p}_A - \hat{p}_B =$ difference between the two estimated proportions
- Distribution: normal distribution

## Formula Review

**Pooled Proportion:**

$$p_c = \frac{x_F + x_M}{n_F + n_M} \tag{10.2.12}$$

**Distribution for the differences:**

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.2.13}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

**Test Statistic ($z$-score):**

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.2.14}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

and

- $p'_A$ and $p'_B$ are the sample proportions, $p_A$ and $p_B$ are the population proportions,
- $P_c$ is the pooled proportion, and $n_A$ and $n_B$ are the sample sizes.

## Glossary

**Pooled Proportion**

estimate of the common value of $p_1$ and $p_2$.

## 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 10.2: Two Independent Population 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.

- **10.4: Comparing Two Independent Population Proportions** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 10.3: Matched or Paired Samples

When using a hypothesis test for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, $\mu_d$, is then tested using a Student's $t$-test for a single population mean with $n - 1$ degrees of freedom, where $n$ is the number of differences.

The test statistic ($t$-score) is:

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in Table. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

| Subject: | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Before | 6.6 | 6.5 | 9.0 | 10.3 | 11.3 | 8.1 | 6.3 | 11.6 |
| After | 6.8 | 2.4 | 7.4 | 8.5 | 8.1 | 6.1 | 3.4 | 2.0 |

**Answer**

Corresponding "before" and "after" values form matched pairs. (Calculate "after" – "before.")

| After Data | Before Data | Difference |
|---|---|---|
| 6.8 | 6.6 | 0.2 |
| 2.4 | 6.5 | -4.1 |
| 7.4 | 9 | -1.6 |
| 8.5 | 10.3 | -1.8 |
| 8.1 | 11.3 | -3.2 |
| 6.1 | 8.1 | -2 |
| 3.4 | 6.3 | -2.9 |
| 2 | 11.6 | -9.6 |

The data for the test are the differences: $\{0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6\}$

The sample mean and sample standard deviation of the differences are: $\bar{x}_d = -3.13$ and $s_d = 2.91$ Verify these values.

Let $\mu_d$ be the population mean for the differences. We use the subscript dd to denote "differences."

**Random variable:**

$\bar{X}_d =$ the mean difference of the sensory measurements

$$H_0 : \mu_d \geq 0 \tag{10.3.1}$$

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after hypnotism. That means the subject shows no improvement. $\mu_d$ is the population mean of the differences.

$$H_a : \mu_d < 0 \tag{10.3.2}$$

The alternative hypothesis is negative, meaning there is less pain felt after hypnotism. That means the subject shows improvement. The score should be lower after hypnotism, so the difference ought to be negative to indicate improvement.

**Distribution for the test:**

The distribution is a Student's *t* with $df = n - 1 = 8 - 1 = 7$. Use $t_7$. *(Notice that the test is for a single population mean.)*

**Calculate the *p*-value using the Student's-t distribution:**

$$p\text{-value} = 0.0095 \tag{10.3.3}$$

**Graph:**

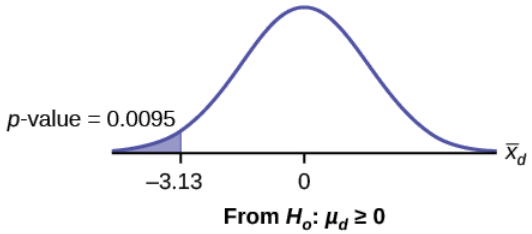



Figure 10.5.1.

$\bar{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\bar{x}_d = -3.13$

$s_d = 2.91$

**Compare $\alpha$ and the *p*-value**

$\alpha = 0.05$ and $p\text{-value} = 0.0095$. $\alpha > p\text{-value}$

**Make a decision**

Since $\alpha > p\text{-value}$, reject $H_0$. This means that $\mu_d < 0$ and there is improvement.

**Conclusion**

At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

> For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction, and you will have the differences in the third list.

Use your list of differences as the data. Press `STAT` and arrow over to `TESTS` . Press `2:T-Test` . Arrow over to `Data` and press `ENTER` . Arrow down and enter `0` for $\mu_0$, the name of the list where you put the data, and `1` for Freq:. Arrow down to $\mu$: and arrow over to `<` $\mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The *p*-value is 0.0094, and the test statistic is -3.04. Do these instructions again except, arrow to `Draw` (instead of `Calculate` ). Press `ENTER` .

## Exercise 10.3.1

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

| Subject | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Before | 209 | 210 | 205 | 198 | 216 | 217 | 238 | 240 | 222 |
| After | 199 | 207 | 189 | 209 | 217 | 202 | 211 | 223 | 201 |

**Answer**

The ***p*-value** is 0.0130, so we can reject the null hypothesis. There is enough evidence to suggest that the diet lowers cholesterol.

## Example 10.3.2

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

| Weight (in pounds) | Player 1 | Player 2 | Player 3 | Player 4 |
|---|---|---|---|---|
| Amount of weight lifted prior to the class | 205 | 241 | 338 | 368 |
| Amount of weight lifted after the class | 295 | 252 | 330 | 360 |

**The coach wants to know if the strength development class makes his players stronger, on average.**

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: $\{90, 11, -8, -8\}$. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$$\bar{x}_d = 21.3 \tag{10.3.4}$$

and

$$s_d = 46.7 \tag{10.3.5}$$

> The data given here would indicate that the distribution is actually right-skewed. The difference 90 may be an extreme outlier? It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are actually negative.

Using the difference data, this becomes a test of a single _____ (fill in the blank).

**Define the random variable:** $\bar{X}$ mean difference in the maximum lift per player.

The distribution for the hypothesis test is $t_3$.

- $H_0 : \mu_d \leq 0$,
- $H_a : \mu_d > 0$

**Graph:**

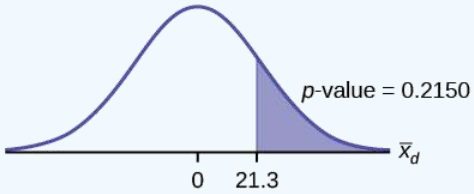



Figure 10.5.2.

**Calculate the $p$-value:** The $p$-value is 0.2150

**Decision:** If the level of significance is 5%, the decision is not to reject the null hypothesis, because $\alpha < p$-value.

**What is the conclusion?**

At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

## Exercise 10.3.2

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data recorded in Table. Are the scores, on average, higher after the class? Test at a 5% level.

| SAT Scores | Student 1 | Student 2 | Student 3 | Student 4 |
|---|---|---|---|---|
| Score before class | 1840 | 1960 | 1920 | 2150 |
| Score after class | 1920 | 2160 | 2200 | 2100 |

**Answer**

The $p$-value is 0.0874, so we decline to reject the null hypothesis. The data do not support that the class improves SAT scores significantly.

## Example 10.3.3

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded in Table.

| Distance (in feet) using | Student 1 | Student 2 | Student 3 | Student 4 | Student 5 | Student 6 | Student 7 |
|---|---|---|---|---|---|---|---|
| Dominant Hand | 30 | 26 | 34 | 17 | 19 | 26 | 20 |
| Weaker Hand | 28 | 14 | 27 | 18 | 17 | 26 | 16 |

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the **differences** data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: $\{2, 12, 7, -1, 2, 0, 4\}$. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation. $\bar{x} = 3.71$, $s_d = 4.5$.

**Random variable:** $\bar{X} =$ mean difference in the distances between the hands.

**Distribution for the hypothesis test:** $t_6$

$H_0 : \mu_d = 0 \quad H_a : \mu_d \neq 0$

**Graph:**

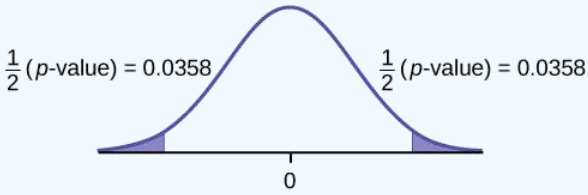



Figure 10.5.3.

**Calculate the *p*-value:** The $p$-value is 0.0716 (using the data directly).

(test statistic = 2.18. $p$-value $= 0.0719$ using ($\bar{x}_d = 3.71, s_d = 4.5$.

**Decision:** Assume $\alpha = 0.05$. Since $\alpha < p$-value, 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 is a difference in the children's weaker and dominant hands to push the shot-put.

**Exercise 10.3.3**

Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in Table. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

| | Player 1 | Player 2 | Player 3 | Player 4 | Player 5 |
|---|---|---|---|---|---|
| Dominant Hand | 120 | 111 | 135 | 140 | 125 |
| Off-hand | 105 | 109 | 98 | 111 | 99 |

**Answer**

The $p$-level is 0.0230, so we can reject the null hypothesis. The data show that the players do not throw the same distance with their off-hands as they do with their dominant hands.

## Review

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: $x_d =$ mean of the differences
- Distribution: Student's t-distribution with $n - 1$ degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

## Formula Review

**Test Statistic (*t*-score):**

$$t = \frac{\bar{x}_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{10.3.6}$$

where:

$\bar{x}_d$ is the mean of the sample differences. $\mu_d$ is the mean of the population differences. $s_d$ is the sample standard deviation of the differences. $n$ is the sample size.

## 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: Matched or Paired 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.

- **10.5: Matched or Paired Samples** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 10.4: Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal and both populations must be normal. The random variable is $\bar{X}_1 - \bar{X}_2$. The normal distribution has the following format:

Normal distribution is:

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \tag{10.4.1}$$

The test statistic (z-score) is:

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have a normal distributions. The data are recorded in Table.

| Wax | Sample Mean Number of Months Floor Wax Lasts | Population Standard Deviation |
|---|---|---|
| 1 | 3 | 0.33 |
| 2 | 2.9 | 0.36 |

Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

**Answer**

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X}_1 - \bar{X}_2 =$ difference in the mean number of months the competing floor waxes last.

- $H_0 : \mu_1 \leq \mu_2$
- $H_a : \mu_1 > \mu_2$

The words **"is more effective"** says that **wax 1 lasts longer than wax 2**, on average. "Longer" is a “>” symbol and goes into $H_a$. Therefore, this is a right-tailed test.

**Distribution for the test:** The population standard deviations are known so the distribution is normal. Using Equation ???, the distribution is:

Since $\mu_1 \leq \mu_2$ then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

**Calculate the $p$-value using the normal distribution:** $p\text{-value} = 0.1799$

**Graph:**

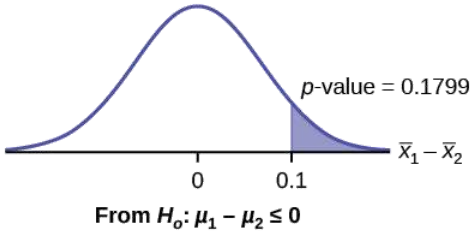



Figure 10.3.1.

$\bar{X}_1 - \bar{X}_2 = 3–2.9 = 0.1$

**Compare $\alpha$ and the $p$-value:** $\alpha = 0.05$ and $p\text{-value} = 0.1799$. Therefore, $\alpha < p\text{-value}$.

**Make a decision:** Since $\alpha < p\text{-value}$, do not reject $H_0$.

**Conclusion:** At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

Press `STAT` . Arrow over to `TESTS` and press `3:2-SampZTest` . Arrow over to `Stats` and press `ENTER` . Arrow down and enter `.33` for sigma1, `.36` for sigma2, `3` for the first sample mean, `20` for n1, `2.9` for the second sample mean, and `20` for n2. Arrow down to $\mu 1$: and arrow to $> \mu_2$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The $p$-value is $p = 0.1799$ and the test statistic is 0.9157. Do the procedure again, but instead of `Calculate` do `Dra` .

## Exercise 10.4.1

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. Table shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5% level of significance.

| Engine | Sample Mean Number of RPM | Population Standard Deviation |
|---|---|---|
| 1 | 1,500 | 50 |
| 2 | 1,600 | 60 |

**Answer**

The $p$-value is almost 0, so we reject the null hypothesis. There is sufficient evidence to conclude that Engine 2 runs at a higher RPM than Engine 1.

## Example 10.4.2: Age of Senators

An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years.

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance.

**Answer**

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is 30 + 30 = 60, which is greater than 30, so we can use the normal approximation to the Student's-t distribution. Subscripts: 1: Democratic senators 2: Republican senators

**Random variable:** $\bar{X}_1 - \bar{X}_2 =$ difference in the mean age of Democratic and Republican U.S. senators.

- $H_0 : \mu_1 \leq \mu_2 \quad H_0 : \mu_1 - \mu_2 \leq 0$
- $H_a : \mu_1 > \mu_2 \quad H_a : \mu_1 - \mu_2 > 0$

The words "older than" translates as a ">" symbol and goes into $H_a$. Therefore, this is a right-tailed test.

**Distribution for the test:** The distribution is the normal approximation to the Student's *t* for means, independent groups. Using the formula, the distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}\right] \tag{10.4.2}$$

Since $\mu_1 \leq \mu_2, \mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

**(Calculating the *p*-value using the normal distribution gives $p$-value $= 0.4040$)**

**Graph:**

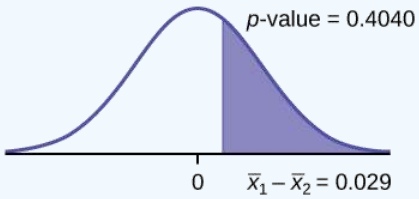



Figure 10.3.2.

**Compare $\alpha$ and the $p$-value:** $\alpha = 0.05$ and $p$-value $= 0.4040$. Therefore, $\alpha < p$-value.

**Make a decision:** Since $\alpha < p$-value, do not reject $H_0$.

**Conclusion:** At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

## References


1. Data from the United States Census Bureau. Available online at www.census.gov/prod/cen2010/b...c2010br-02.pdf
2. Hinduja, Sameer. "Sexting Research and Gender Differences." Cyberbulling Research Center, 2013. Available online at cyberbullying.us/blog/sexting...r-differences/ (accessed June 17, 2013).
3. "Smart Phone Users, By the Numbers." Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).
4. Smith, Aaron. "35% of American adults own a Smartphone." Pew Internet, 2013. Available online at www.pewinternet.org/~/media/F...martphones.pdf (accessed June 17, 2013).
5. "State-Specific Prevalence of Obesity AmongAduls—Unites States, 2007." MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).
6. "Texas Crime Rates 1960–1012." FBI, Uniform Crime Reports, 2013. Available online at: http://www.disastercenter.com/crime/txcrime.htm (accessed June 17, 2013).


## Review

A hypothesis test of two population means from independent samples where the population standard deviations are known will have these characteristics:

- Random variable: $\overline{X}_1 - \overline{X}_2 =$ the difference of the means
- Distribution: normal distribution

## Formula Review

**Normal Distribution:**

$$\bar{X}_1 - \bar{X}_2 =\sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{10.4.3}$$

Generally $\mu_1 - \mu_2 = 0$.

**Test Statistic ($z$-score):**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{10.4.4}$$

Generally $\mu_1 - \mu_2 = 0$.

where:

$\sigma_1$ and $\sigma_2$ are the known population standard deviations. $n_1$ and $n_1$ are the sample sizes. $\bar{x}_1$ and $\bar{x}_2$ are the sample means. $\mu_1$ and $\mu_2$ are the population means

## 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 10.4: Two Population Means with Known Standard Deviations 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: Two Population Means with Known Standard Deviations** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 10.5: Difference of Two Means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. The methods are similar in theory but different in the details. Just as with a single sample, we identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal. Next we introduce a formula for the standard error, which allows us to apply our general tools from Section 4.5.

We apply these methods to two examples: participants in the 2012 Cherry Blossom Run and newborn infants. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

## Point Estimates and Standard Errors for Differences of Means

We would like to estimate the average difference in run times for men and women using the run10Samp data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run. Table **10.5.2** presents relevant summary statistics, and box plots of each sample are shown in Figure 5.6.

Table **10.5.2**: Summary statistics for the run time of 100 participants in the 2009 Cherry Blossom Run.

| | men | women |
|---|---|---|
| $\bar{x}$ | 87.65 | 102.13 |
| $s$ | 12.5 | 15.2 |
| $n$ | 45 | 55 |

The two samples are independent of one-another, so the data are not paired. Instead a point estimate of the difference in average 10 mile times for men and women, $\mu_w - \mu_m$, can be found using the two sample means:

$$\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48 \tag{10.5.1}$$

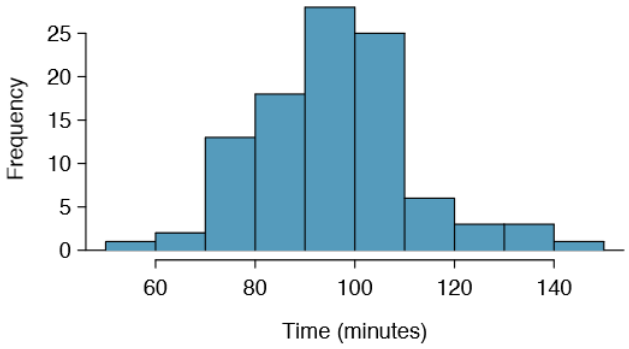



Figure **10.5.1**: *A histogram of time for the sample Cherry Blossom Race data.*

Because we are examining two simple random samples from less than 10% of the population, each sample contains at least 30 observations, and neither distribution is strongly skewed, we can safely conclude the sampling distribution of each sample mean is nearly normal. Finally, because each sample is independent of the other (e.g. the data are not paired), we can conclude that the difference in sample means can be modeled using a normal distribution. (Probability theory guarantees that the difference of two independent normal random variables is also normal. Because each sample mean is nearly normal and observations in the samples are independent, we are assured the difference is also nearly normal.)

**Conditions for normality of $\bar{x}_1 - \bar{x}_2$**

If the sample means, $\bar{x}_1$ and $\bar{x}_2$, each meet the criteria for having nearly normal sampling distributions and the observations in the two samples are independent, then the difference in sample means, $\bar{x}_1 - \bar{x}_2$, will have a sampling distribution that is nearly normal.

We can quantify the variability in the point estimate, $\bar{x}_w - \bar{x}_m$, using the following formula for its standard error:

$$SE_{\bar{x}_w - \bar{x}_m} = \sqrt{\frac{\sigma_w^2}{n_w} + \frac{\sigma_m^2}{n_m}} \tag{10.5.2}$$

We usually estimate this standard error using standard deviation estimates based on the samples:

$$SE_{\bar{x}_w - \bar{x}_m} \approx \sqrt{\frac{s_w^2}{n_w} + \frac{s_m^2}{n_m}} \tag{10.5.3}$$

$$= \sqrt{\frac{15.2^2}{55} + \frac{12.5^2}{45}} \tag{10.5.4}$$

$$= 2.77 \tag{10.5.5}$$

Because each sample has at least 30 observations ($n_w = 55$ and $n_m = 45$), this substitution using the sample standard deviation tends to be very good.

**Distribution of a difference of sample means**

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, is nearly normal with mean $\mu_1 - \mu_2$ and estimated standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.5.6}$$

when each sample mean is nearly normal and all observations are independent.

## Confidence Interval for the Difference

When the data indicate that the point estimate $\bar{x}_1 - \bar{x}_2$ comes from a nearly normal distribution, we can construct a confidence interval for the difference in two means from the framework built in Chapter 4. Here a point estimate, $\bar{x}_w - \bar{x}_m = 14.48$, is associated with a normal model with standard error SE = 2.77. Using this information, the general confidence interval formula may be applied in an attempt to capture the true difference in means, in this case using a 95% confidence level:

$$\text{point estimate} \pm z^* SE \rightarrow 14.48 \pm 1.96 \times 2.77 = (9.05, 19.91) \tag{10.5.7}$$

Based on the samples, we are 95% confident that men ran, on average, between 9.05 and 19.91 minutes faster than women in the 2012 Cherry Blossom Run.

**Exercise 10.5.1**

What does 95% confidence mean?

**Solution**

If we were to collected many such samples and create 95% confidence intervals for each, then about 95% of these intervals would contain the population difference, $\mu_w - \mu_m$.

**Exercise 10.5.2**

We may be interested in a different confidence level. Construct the 99% confidence interval for the population difference in average run times based on the sample data.

**Solution**

The only thing that changes is z*: we use z* = 2:58 for a 99% confidence level. (If the selection of $z^*$ is confusing, see Section 4.2.4 for an explanation.) The 99% confidence interval:

$$14.48 \pm 2.58 \times 2.77 \rightarrow (7.33, 21.63). \quad (10.5.8)$$

We are 99% confident that the true difference in the average run times between men and women is between 7.33 and 21.63 minutes.

## Hypothesis tests Based on a Difference in Means

A data set called baby smoke represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table **10.5.2**. We are particularly interested in two variables: weight and smoke. The weight variable represents the weights of the newborns and the smoke variable describes which mothers smoked during pregnancy. We would like to know if there is convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure **10.5.2**.

Table **10.5.2**: Four cases from the baby smoke data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

| | fAge | mAge | weeks | weight | sexBaby | smoke |
|---|---|---|---|---|---|---|
| 1 | NA | 13 | 37 | 5.00 | female | nonsmoker |
| 2 | NA | 14 | 36 | 5.88 | female | nonsmoker |
| 3 | 19 | 15 | 41 | 8.13 | male | smoker |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 150 | 45 | 50 | 36 | 9.25 | female | nonsmoker |

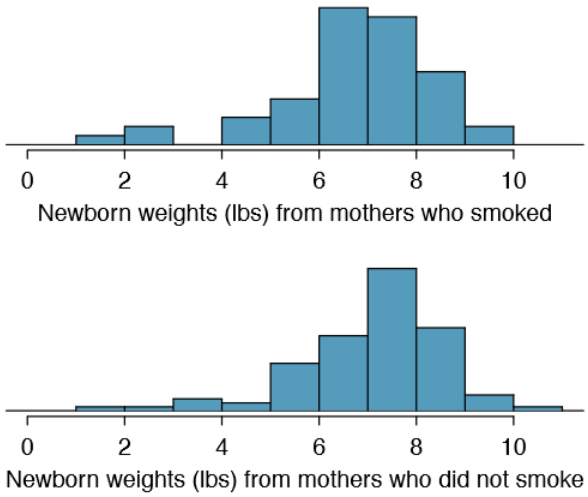



Figure **10.5.2**: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. Both distributions exhibit strong skew.

**Example 10.5.1**

Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

**Solution**

The null hypothesis represents the case of no difference between the groups.

- $H_0$: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where $\mu_n$ represents non-smoking mothers and $\mu_s$ represents mothers who smoked.
- $H_A$: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

Summary statistics are shown for each sample in Table **10.5.3**. Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, each group's sample size is at least 30 and the skew in each sample distribution is strong (Figure **10.5.2**). However, this skew is reasonable for these sample sizes of 50 and 100. Therefore, each sample mean is associated with a nearly normal distribution.

Table **10.5.3**: Summary statistics for the baby smoke data set.

| | smoker | nonsmoker |
|---|---|---|
| mean | 6.78 | 7.18 |
| st. dev. | 1.43 | 1.60 |
| samp. size | 50 | 100 |

**Exercise 10.5.3**

a. What is the point estimate of the population difference, $\mu_n - \mu_s$?
b. Can we use a normal distribution to model this difference?
c. Compute the standard error of the point estimate from part (a)

**Solution**

(a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$.

(b) Because the samples are independent and each sample mean is nearly normal, their difference is also nearly normal.

(c) The standard error of the estimate can be estimated using Equation 10.5.6:

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26 \tag{10.5.9}$$

The standard error estimate should be sufficiently accurate since the conditions were reasonably satisfied.

**Example 10.5.2**

If the null hypothesis from Exercise 5.8 was true, what would be the expected value of the point estimate? And the standard deviation associated with this estimate? Draw a picture to represent the p-value.

**Solution**

If the null hypothesis was true, then we expect to see a difference near 0. The standard error corresponds to the standard deviation of the point estimate: 0.26. To depict the p-value, we draw the distribution of the point estimate as though $H_0$ was true and shade areas representing at least as much evidence against $H_0$ as what was observed. Both tails are shaded because it is a two-sided test.

**Example 10.5.3**

Compute the p-value of the hypothesis test using the figure in Example 5.9, and evaluate the hypotheses using a signi cance level of $\alpha = 0.05$.

**Solution**

Since the point estimate is nearly normal, we can nd the upper tail using the Z score and normal probability table:

$$Z = \frac{0.40 - 0}{0.26} = 1.54 \rightarrow \text{upper tail} = 1 - 0.938 = 0.062 \tag{10.5.10}$$

Because this is a two-sided test and we want the area of both tails, we double this single tail to get the p-value: 0.124. This p-value is larger than the signi cance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say

there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

**Exercise 10.5.4**

Does the conclusion to Example 5.10 mean that smoking and average birth weight are unrelated?

**Solution**

Absolutely not. It is possible that there is some difference but we did not detect it. If this is the case, we made a Type 2 Error.

**Exercise 10.5.5**

If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect such a difference?

**Solution**

We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

## Summary for inference of the difference of two means

When considering the difference of two means, there are two common cases: the two samples are paired or they are independent. (There are instances where the data are neither paired nor independent.) The paired case was treated in Section 5.1, where the one-sample methods were applied to the differences from the paired observations. We examined the second and more complex scenario in this section.

When applying the normal model to the point estimate $\bar{x}_1 - \bar{x}_2$ (corresponding to unpaired data), it is important to verify conditions before applying the inference framework using the normal model. First, each sample mean must meet the conditions for normality; these conditions are described in Chapter 4 on page 168. Secondly, the samples must be collected independently (e.g. not paired data). When these conditions are satisfied, the general inference tools of Chapter 4 may be applied.

For example, a confidence interval may take the following form:

$$\text{point estimate} \pm z^* SE \tag{10.5.11}$$

When we compute the confidence interval for $\mu_1 - \mu_2$, the point estimate is the difference in sample means, the value $z^*$ corresponds to the confidence level, and the standard error is computed from Equation 10.5.6. While the point estimate and standard error formulas change a little, the framework for a confidence interval stays the same. This is also true in hypothesis tests for differences of means.

In a hypothesis test, we apply the standard framework and use the specific formulas for the point estimate and standard error of a difference in two means. The test statistic represented by the Z score may be computed as

$$Z = \frac{\text{point estimate - null value}}{SE} \tag{10.5.12}$$

When assessing the difference in two means, the point estimate takes the form $\bar{x}_1 - \bar{x}_2$, and the standard error again takes the form of Equation 10.5.6. Finally, the null value is the difference in sample means under the null hypothesis. Just as in Chapter 4, the test statistic Z is used to identify the p-value.

## Examining the Standard Error Formula

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}} \tag{10.5.13}$$

where $s_1$ and $n_1$ represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.5.14}$$

This special relationship follows from probability theory.

**Exercise 10.5.6**

Prerequisite: Section 2.4. We can rewrite Equation 10.5.14 in a different way:

$$SE_{\bar{x}_1 - \bar{x}_2}^2 = SE_{\bar{x}_1}^2 + SE_{barx_2}^2 \tag{10.5.15}$$

Explain where this formula comes from using the ideas of probability theory.[10]

---

This page titled 10.5: Difference of Two Means 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.

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

# CHAPTER OVERVIEW

## 11: Linear Regression and Hypothesis Testing

11.1: Testing the Hypothesis that $\beta = 0$

---

11: Linear Regression and Hypothesis Testing is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

# 11.1: Testing the Hypothesis that β = 0

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 $\rho$, 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 $\rho$ 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.

> - 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 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

> 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

To calculate the $p$-value using LinRegTTEST:

On the LinRegTTEST input screen, on the line prompt for $\beta$ or $\rho$, 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 11.1.1

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.

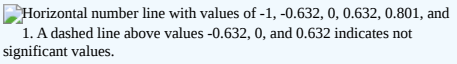



Figure 11.1.1. $r$ is not significant between $-0.632$ and $+0.632$. $r = 0.801 > +0.632$. Therefore, $r$ is significant.

### Exercise 11.1.1

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 11.1.2

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 11.1.2. $r = –0.624 – 0.532$. Therefore, $r$ is significant.

### Exercise 11.1.2

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 11.1.3

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.

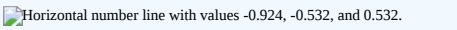



Figure 11.1.3. $–0.811 < r = 0.776 < 0.811$. Therefore, $r$ is not significant.

### Exercise 11.1.3

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 11.1.4**

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 11.1.4**

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.

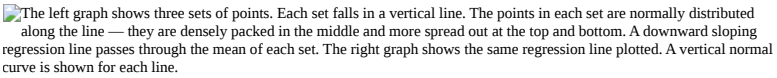



Figure 11.1.4. The $y$ values for each $x$ value are normally distributed about the line with the same standard deviation. For each $x$ value, the mean of the $y$ values lies on the regression line. More $y$ values lie near the line than are scattered further away from the line.

## 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 $\beta$ and intercept $\alpha$ of the population (true) regression line. To estimate the population standard deviation of $y$, $\sigma$, use the standard deviation of the residuals, $s$. $s = \sqrt{\frac{SEE}{n-2}}$. The variable $\rho$ (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:

$$\hat{y} = a + bx \tag{11.1.1}$$

where

$$a = y\text{-intercept} \tag{11.1.2}$$

$$b = \text{slope} \tag{11.1.3}$$

Standard deviation of the residuals:

$$s = \sqrt{\frac{SEE}{n-2}} \tag{11.1.4}$$

where

$$SSE = \text{sum of squared errors} \tag{11.1.5}$$

$$n = \text{the number of data points} \tag{11.1.6}$$

This page titled 11.1: Testing the Hypothesis that β = 0 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.

- **12.5: Testing the Significance of the Correlation Coefficient** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 12: The Chi-Square Distribution

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.


12.1: The Chi-Square Distribution
12.2: A Goodness-of-Fit Test
12.3: A Test of Independence or Homogeneity
12.4: Test of a Single Variance
12.5: Test for Homogeneity
12.6: Comparison of the Chi-Square Tests



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 12: 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.

# 12.1: The Chi-Square Distribution

The notation for the chi-square distribution is:

$$\chi \sim \chi^2_{df} \tag{12.1.1}$$

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 $\chi^2$ distribution, the population mean is $\mu = df$ and the population standard deviation is

$$\sigma = \sqrt{2(df)}. \tag{12.1.2}$$

The random variable is shown as $\chi^2$, 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.

$$\chi^2 = (Z_1)^2 + \ldots + (Z_k)^2 \tag{12.1.3}$$

a. The curve is nonsymmetrical and skewed to the right.
b. There is a different chi-square curve for each $df$.

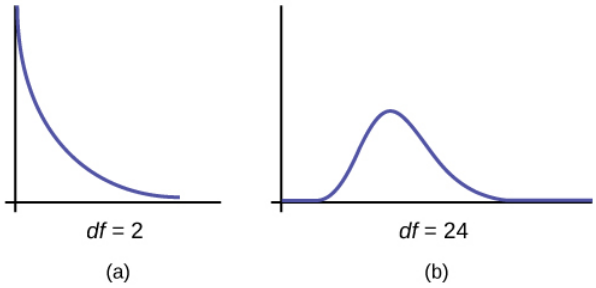



Figure 12.1.1

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, $\mu$, is located just to the right of the peak.

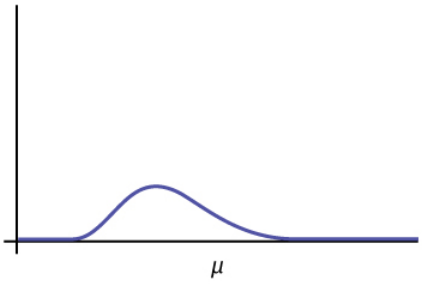



Figure 12.1.2

## WeBWorK Problems

## 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

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_{df})^2 \quad (12.1.4)$$

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

---

This page titled 12.1: 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: Facts About the Chi-Square Distribution** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 12.2: A 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:

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 Table.

| 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 Table 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.

**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 Table. 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 $\chi^2$ 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 $\chi^2$ 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 $\chi^2$cdf. 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.)

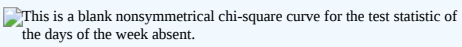



Figure 12.2.1.

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.

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:** $\chi^2_2$ where $df = 3 - 1 = 2$.

**Calculate the test statistic:** $\chi^2 = 2.14$

**Graph:**

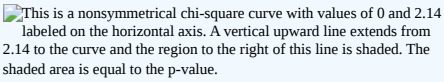



Figure 12.2.3.

**Probability statement:** $p\text{-value} = P(\chi^2 > 2.14) = 0.3430$

**Compare $\alpha$ 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.

## WeBWorK Problems

## 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

---

This page titled 12.2: A 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.

# 12.3: A Test of Independence or Homogeneity

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:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{12.3.1}$$

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 12.3.1**

Suppose $A =$ a speeding violation in the last year and $B =$ a cell phone user while driving. If $A$ and $B$ are independent then $P(A \text{ AND } B) = P(A)P(B)$. $A$ 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** ($E$):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

.

**Example 12.3.2**

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 12.3.1 is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table 12.3.1: 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 12.3.2.

Table 12.3.2: 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:** $\chi^2_4$

$$df = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$$

**Graph:**

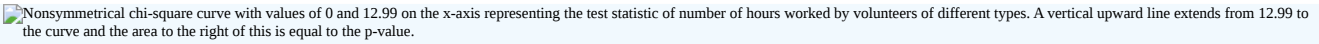



Figure **12.3.1**.

**Probability statement:** $p\text{-value} = P(\chi^2 > 12.99) = 0.0113$

**Compare $\alpha$ and the $p$-value:** Since no $\alpha$ 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?

## Example 12.3.3

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.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09 \quad (12.3.2)$$

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

## WeBWorK Problems

## 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 $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

## 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 12.3: A Test of Independence or 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.

- **11.4: Test of Independence** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 12.4: Test of a Single Variance

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternative hypotheses are stated in terms of the population variance (or population standard deviation). The test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{12.4.1}$$

where:

- $n$ is the the total number of data
- $s^2$ is the sample variance
- $\sigma^2$ is the population variance

You may think of $s$ as the random variable in this test. The number of degrees of freedom is $df = n - 1$. **A test of a single variance may be right-tailed, left-tailed, or two-tailed.** The next example will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

**Example 12.4.1**

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

**Answer**

Even though we are given the population standard deviation, we can set up the test using the population variance as follows.

- $H_0 : \sigma^2 = 5^2$
- $H_a : \sigma^2 > 5^2$

**Example 12.4.2**

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5%, test the claim that **a single line causes lower variation among waiting times (shorter waiting times) for customers**.

**Answer**

Since the claim is that a single line causes less variation, this is a test of a single variance. The parameter is the population variance, $\sigma^2$, or the population standard deviation, $\sigma$.

**Random Variable:** The sample standard deviation, $s$, is the random variable. Let $s =$ **standard deviation for the waiting times.**

- $H_0 : \sigma^2 = 7.2^2$
- $H_a : \sigma^2 < 7.2^2$

The word **"less"** tells you this is a left-tailed test.

**Distribution for the test:** $\chi^2_{24}$, where:

- $n =$ **the number of customers sampled**
- $df = n - 1 = 25 - 1 = 24$

**Calculate the test statistic (Equation 12.4.1):**

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

**Graph:**

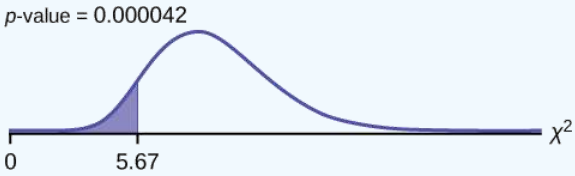



Figure 12.4.1.

**Probability statement:** $p\text{-value} = P(\chi^2 < 5.67) = 0.000042$

**Compare $\alpha$ and the $p$-value:**

$$\alpha = 0.05 (p\text{-value} = 0.000042 \alpha > p\text{-value}$$

**Make a decision:** Since $\alpha > p\text{-value}$, reject $H_0$. This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes; you think the variation in waiting times is less.

**Conclusion:** At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times **or** with a single line, the customer waiting times vary less than 7.2 minutes.

## References


1. "AppleInsider Price Guides." Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).
2. Data from the World Bank, June 5, 2012.


## Review

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

## Formula Review

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ Test of a single variance statistic where:

$n$ : sample size

$s$ : sample standard deviation

$\sigma$ : population standard deviation

$df = n$–1Degrees of freedom

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom is the number of samples − 1.
- The test statistic is $\frac{(n-1)\cdot s^2}{\sigma^2}$, where $n$ = the total number of data, $s^2$ = sample variance, and $\sigma^2$ = population variance.
- The test may be left-, right-, or two-tailed.

*Use the following information to answer the next three exercises:* An archer's standard deviation for his hits is six (data is measured in distance from the center of the target). An observer claims the standard deviation is less.


This page titled 12.4: Test of a Single Variance 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.

- **11.7: Test of a Single Variance** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

https://stats.libretexts.org/@go/page/26119

# 12.5: 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 $\chi^2$ 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 12.5.1**

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 **12.5.1**. Do male and female college students have the same distribution of living arrangements?

Table **12.5.1**: 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:** $\chi^2_3$

**Calculate the test statistic:** $\chi^2 = 10.1287$ (calculator or computer)

**Probability statement:** $p\text{-value} = P(\chi^2 > 10.1287) = 0.0175$

**Compare $\alpha$ and the $p$-value:** Since no $\alpha$ is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0175$. $\alpha > p\text{-value}$.

**Make a decision:** Since $\alpha > p\text{-value}$, 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.

### 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:** $\chi^2_2$

**Calculate the test statistic**: $\chi^2 = 3.2603$ (calculator or computer)

**Probability statement:** $p\text{-value} = P(\chi^2 > 3.2603) = 0.1959$

**Compare $\alpha$ 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.

## 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

---

This page titled 12.5: 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.

- **11.5: Test for Homogeneity** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 12.6: Comparison of the Chi-Square Tests

You have seen the $\chi^2$ test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which $\chi^2$ 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.

This page titled 12.6: 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.

- **11.6: Comparison of the Chi-Square Tests** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# CHAPTER OVERVIEW

## 13: F Distribution and One-Way ANOVA

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

**Topic hierarchy**


13.1: Prelude to F Distribution and One-Way ANOVA

13.2: One-Way ANOVA

13.3: The F Distribution and the F-Ratio

13.4: Facts About the F Distribution

13.5: Test of Two Variances


### 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 13: 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.

# 13.1: 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 **13.1.1**: 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 13.1: 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.

- **13.1: Prelude to F Distribution and One-Way ANOVA** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 13.2: 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).

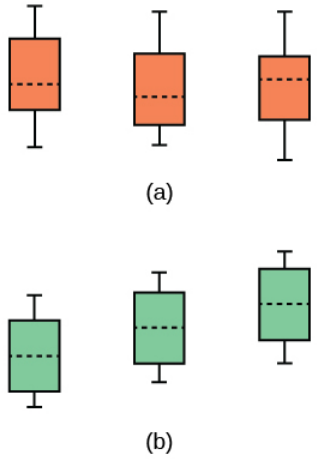



Figure 13.2.1: (a) $H_0$ is true. All means are the same; the differences are due to random variation. (b) $H_0$ is not true. All means are not the same; the differences are too large to be due to random variation.

## 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.

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 13.2: 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.

- **13.2: One-Way ANOVA** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# 13.3: 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 $\sigma^2$ 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 $\sigma^2$ 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.**

- $SS_{\text{between}}$ = the sum of squares that represents the variation among the different samples
- $SS_{\text{within}}$ = 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." $MS_{\text{between}}$ is the variance between groups, and $MS_{\text{within}}$ 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 $j^{th}$ group}
- $s_j$ = the sum of the values in the $j^{th}$ group
- $n$ = total number of all the values combined (total sample size):

$$n = \sum n_j \tag{13.3.1}$$

- $x$ = one value:

$$\sum x = \sum s_j \tag{13.3.2}$$

- Sum of squares of all values from every group combined:

$$\sum x^2 \tag{13.3.3}$$

- Between group variability:

$$SS_{\text{total}} = \sum x^2 - \frac{(\sum x^2)}{n} \tag{13.3.4}$$

- Total sum of squares:

$$\sum x^2 - \frac{(\sum x)^2}{n} \tag{13.3.5}$$

- Explained variation: sum of squares representing variation among the different samples:

$$SS_{\text{between}} = \sum \left[ \frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n} \tag{13.3.6}$$

- Unexplained variation: sum of squares representing variation within samples due to chance:

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}} \tag{13.3.7}$$

- $df$'s for different groups ($df$'s for the numerator):

$$df = k - 1 \tag{13.3.8}$$

- Equation for errors within samples ($df$'s for the denominator):

$$df_{\text{within}} = n - k \tag{13.3.9}$$

- Mean square (variance estimate) explained by the different groups:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} \tag{13.3.10}$$

- Mean square (variance estimate) that is due to chance (unexplained):

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} \tag{13.3.11}$$

$MS_{\text{between}}$ and $MS_{\text{within}}$ can be written as follows:

The one-way ANOVA test depends on the fact that $MS_{\text{between}}$ can be influenced by population differences among means of the several groups. Since $MS_{\text{within}}$ compares values of each group to its own group mean, the fact that group means might be different does not affect $MS_{\text{within}}$.

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, $MS_{\text{between}}$ and $MS_{\text{within}}$ 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

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{13.3.12}$$

If $MS_{\text{between}}$ and $MS_{\text{within}}$ 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, $MS_{\text{between}}$ consists of the population variance plus a variance produced from the differences between the samples. $MS_{\text{within}}$ is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, $MS_{\text{between}}$ will generally be larger than $MS_{\text{within}}$.Then the $F$-ratio will be larger than one. However, if the population effect is small, it is not unlikely that $MS_{\text{within}}$ 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

$$F = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2} \tag{13.3.13}$$

where ...

- $n$ = the sample size

- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- $s^2_{\text{pooled}} =$ 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) | $SS(\text{Factor})$ | | | $F = \frac{MS(\text{Factor})}{MS(\text{Error})}$ |
| Error (Within) | $SS(\text{Error})$ | | | |
| Total | $SS(\text{Total})$ | | | |

**Example 13.3.1**

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 | |

$$s_1 = 16.5, s_2 = 15, s_3 = 15.7 \tag{13.3.14}$$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

where $n_1 = 4, n_2 = 3, n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$ so

$$SS(\text{between}) = \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} = \frac{(16.5 + 15 + 15.5)^2}{10} \tag{13.3.15}$$

$$= 2.2458 \tag{13.3.16}$$

$$S(\text{total}) = \sum x^2 - \frac{(\sum x)^2}{n} \tag{13.3.17}$$

$$= (5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2) \tag{13.3.18}$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10} \tag{13.3.19}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9 \tag{13.3.20}$$

$$= 23.1 \tag{13.3.21}$$

$$SS(\text{within}) = SS(\text{total}) - SS(\text{between}) \tag{13.3.22}$$

$$= 23.1 - 2.2458 \tag{13.3.23}$$

$$= 20.8542 \tag{13.3.24}$$

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) | $SS(\text{Factor}) = SS(\text{Between}) = 2.2458$ | | $MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)} = \frac{2.2458}{2} = 1.1229$ | $F = \frac{MS(\text{Factor})}{MS(\text{Error})} = \frac{1.1229}{2.9792} = 0.3769$ |
| Error (Within) | $SS(\text{Error}) = SS(\text{Within}) = 20.8542$ | | $MS(\text{Error})) = \frac{SS(\text{Error})}{(n-k)} = \frac{20.8542}{7} = 2.9792$ | |
| Total | $SS(\text{Total}) = 2.2458 + 20.8542 = 23.1$ | | | |

## Exercise 13.3.1

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 | | $\frac{36,648,561}{4} = 9,162,140$ | $\frac{9,162,140}{2,044,672.6} = 4.4810$ |
| Error (Within) | 20,446,726 | | $\frac{20,446,726}{10} = 2,044,672.6$ | |
| Total | 57,095,287 | | | |

*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 $j^{th}$ group
- $s_j =$ the sum of the values in the $j^{th}$ group
- $n =$ the total number of all values (observations) combined
- $x =$ one value (one observation) from the data
- $s_{\bar{x}}^2 =$ 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 13.3: 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.

# 13.4: 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.

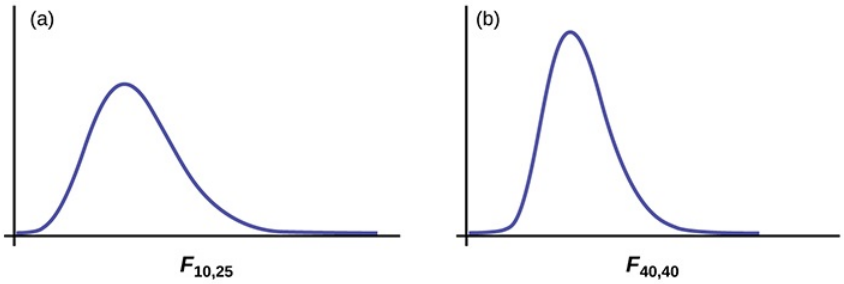



Figure **13.4.1**

**Example 13.4.1**

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 | | $\frac{36,648,561}{4} = 9,162,140$ | $\frac{9,162,140}{2,044,672.6} = 4.4810$ |
| Error (Within) | 20,446,726 | | $\frac{20,446,726}{10} = 2,044,672.6$ | |
| Total | 57,095,287 | | | |

**Distribution for the test:** $F_{4,10}$

$$df(\text{num}) = 5 - 1 = 4 \quad (13.4.1)$$

$$df(\text{denom}) = 15 - 5 = 10 \quad (13.4.2)$$

**Test statistic:** $F = 4.4810$

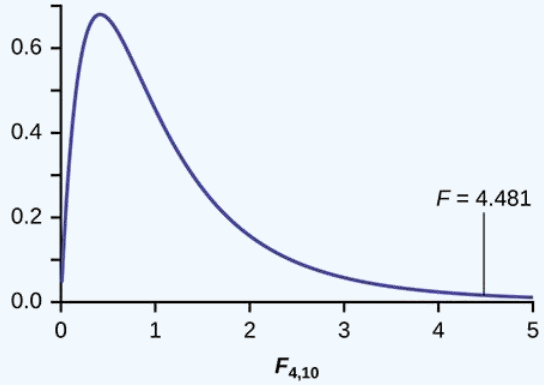



Figure 13.4.2

**Probability Statement:** $p\text{-value} = P(F > 4.481) = 0.0248$.

**Compare $\alpha$ 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 13.4.1**

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:

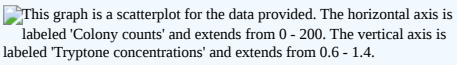



Figure **13.4.3**

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 **13.4.1**), 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 **13.4.1**

| Source of Variation | Sum of Squares ($SS$) | Degrees of Freedom ( $df$) | Mean Square ($MS$) | $F$ |
|---|---|---|---|---|
| Factor (Between) | 10,233 | | $\frac{10,233}{4} = 2,558.25$ | $\frac{2,558.25}{4,194.9} = 0.6099$ |
| Error (Within) | 41,949 | | | |
| Total | 52,182 | | $\frac{41,949}{10} = 4,194.9$ | |

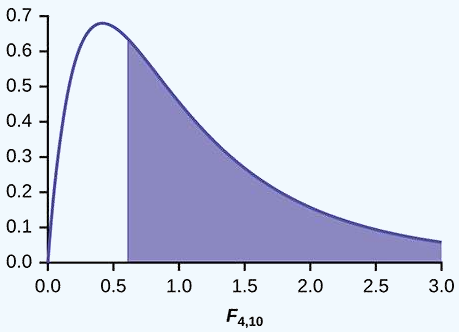



***Figure* 13.4.2**

**Distribution for the test:** $F_{4,10}$

**Probability Statement:** $p\text{-value} = P(F > 0.6099) = 0.6649$.

**Compare $\alpha$ 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 13.4.2

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in Table.

Figure **13.4.1**: 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:**

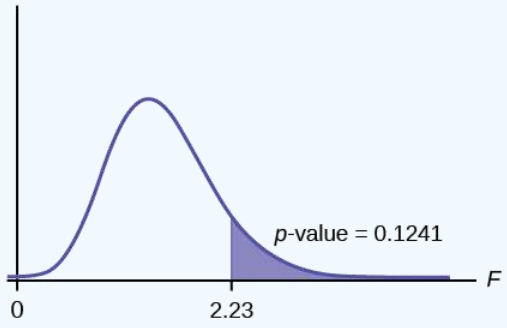



Figure **13.4.5**

**Probability statement:** $p\text{-value} = P(F > 2.23) = 0.1241$

**Compare $\alpha$ 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 13.4.2

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 13.4.3

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 13.4.3.

Table 13.4.3

| 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

$$F' = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2}. \tag{13.4.3}$$

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 13.4.3

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 13.4.3.

**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!

## 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 13.4: 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.

# 13.5: Test of Two Variances

Another of the uses of the $F$ distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

**to perform a $F$ test of two variances, it is important that the following are true:**

- The populations from which the two samples are drawn are *normally* distributed.
- The two populations are *independent* of each other.

Unlike most other tests in this book, the $F$ test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher $p$-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let $\sigma_1^2$ and $\sigma_2^2$ be the population variances and $s_1^2$ and $s_2^2$ be the sample variances. Let the sample sizes be $n_1$ and $n_2$. Since we are interested in comparing the two sample variances, we use the $F$ ratio:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} \tag{13.5.1}$$

$F$ has the distribution

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the $F$ Ratio becomes

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}. \tag{13.5.2}$$

The $F$ ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on $H_a$ and on which sample variance is larger.

If the two populations have equal variances, then $s_1^2$ and $s_2^2$ are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to one. But if the two population variances are very different, $s_1^2$ and $s_2^2$ tend to be very different, too. Choosing $s_1^2$ as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than one. If $s_1^2$ and $s_2^2$ are far apart, then

$$F = \frac{(s_1)^2}{(s_2)^2} \tag{13.5.3}$$

is a large number.

Therefore, if $F$ is close to one, the evidence favors the null hypothesis (the two population variances are equal). But if $F$ is much larger than one, then the evidence is against the null hypothesis. A test of two variances may be left, right, or two-tailed.

> *A test of two variances may be left, right, or two-tailed.*

**Example 13.5.1**

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a

variance of 89.9. Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

**Answer**

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

- $n_1 = n_2 = 30$.
- $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_a : \sigma_1^2 < \sigma_2^2$

**Calculate the test statistic:** By the null hypothesis $\sigma_1^2 = \sigma_2^2$), the $F$ statistic is:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(s_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818 \quad (13.5.4)$$

**Distribution for the test:** $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

**Graph: This test is left tailed.**

Draw the graph labeling and shading appropriately.

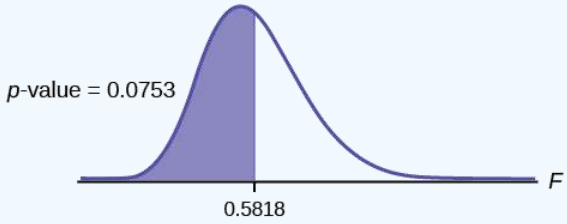



Figure **13.5.1**

**Probability statement:** $p\text{-value} = P(F < 0.5818) = 0.0753$

**Compare $\alpha$ and the $p$-value:** $\alpha = 0.10 \alpha > p\text{-value}$.

**Make a decision:** Since $\alpha > p\text{-value}$, reject $H_0$.

**Conclusion:** With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Press `STAT` and arrow over to `TESTS` . Arrow down to `D:2-SampFTest` . Press `ENTER` . Arrow to `Stats` and press `ENTER` . For `Sx1` , `n1` , `Sx2` , and `n2` , enter `(52.3)-----√(52.3)` , `30` , `(89.9)-----√(89.9)` , and `30` . Press `ENTER` after each. Arrow to `σ1:` and `<σ2` . Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . $F$ = 0.5818 and $p$-value = 0.0753. Do the procedure again and try `Draw` instead of `Calculate` .

## Exercise 13.5.1

The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

| Tenor1 | Bass2 | Tenor 1 | Bass 2 | Tenor 1 | Bass 2 |
|---|---|---|---|---|---|
| 69 | 72 | 67 | 72 | 68 | 67 |
| 72 | 75 | 70 | 74 | 67 | 70 |
| 71 | 67 | 65 | 70 | 64 | 70 |
| 66 | 75 | 72 | 66 | | 69 |
| 76 | 74 | 70 | 68 | | 72 |
| 74 | 72 | 68 | 75 | | 71 |
| 71 | 72 | 64 | 68 | | 74 |
| 66 | 74 | 73 | 70 | | 75 |
| 68 | 72 | 66 | 72 | | |

**Answer**

The histograms are not as normal as one might like. Plot them to verify. However, we proceed with the test in any case.

Subscripts: $\mathbf{T1} =$ tenor 1 and $\mathbf{B2} =$ bass 2

The standard deviations of the samples are $s_{\mathrm{T1}} = 3.3302$ and $s_{\mathrm{B2}} = 2.7208$.

The hypotheses are

$H_0 : \sigma^2_{\mathrm{T1}} = \sigma^2_{\mathrm{B2}}$ and $H_0 : \sigma^2_{\mathrm{T1}} \neq \sigma^2_{\mathrm{B2}}$ (two tailed test)

The $F$ statistic is $1.4894$ with 20 and 25 degrees of freedom.

The $p$-value is $0.3430$. If we assume alpha is 0.05, then we cannot reject the null hypothesis.

We have no good evidence from the data that the heights of Tenor1 and Bass2 singers have different variances (despite there being a significant difference in mean heights of about 2.5 inches.)

## References

1. "MLB Vs. Division Standings – 2012." Available online at http://espn.go.com/mlb/standings/_/y...ion/order/true.

## Review

The $F$ test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an $F$ statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two groups. A **test of two variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the $F$ distribution with two different degrees of freedom.

**Assumptions:**

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

## Formula Review

$F$ has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \frac{s_1^2}{s_2^2}$

## 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 13.5: Test of Two Variances 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.

- **13.5: Test of Two Variances** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# Index


## A

Adding probabilities
- 4.4: Counting Basics- the Multiplication and Addition Rules

anova
- 13.2: One-Way ANOVA

## B

bar graph
- 2.3: Stem-and-Leaf Displays

Bernoulli trial
- 5.4: The Binomial Distribution
- 5.5: The Geometric Distribution

binomial probability distribution
- 5.4: The Binomial Distribution
- 8.3: Estimating Proportions

box plots
- 2.6: Box Plots

## C

central limit theorem for sums
- 7.2: The Sum Distribution

cluster sampling
- 1.3: Populations and Samples

coefficient of determination
- 3.3: Simple Linear Regression

Cohen's Standards
- 10.1: Two Population Means

Comparing two population means
- 10.4: Two Population Means with Known Standard Deviations

Comparing Two Population Proportions
- 10.2: Two Independent Population Proportions

complement
- 4.1: Probability Experiments and Sample Spaces
- 4.3: Conditional Probability and Independence

conditional probability
- 4.1: Probability Experiments and Sample Spaces

contingency table
- 4.6: Joint and Marginal Probabilities and Contingency Tables
- 12.3: A Test of Independence or Homogeneity

continuous data
- 1.3: Populations and Samples

## D

Decision
- 9.4: Hypothesis Tests about μ- Critical Region Approach

degrees of freedom
- 10.1: Two Population Means

direction of a relationship between the variables
- 3.2: Scatter Plots

discrete data
- 1.3: Populations and Samples

Distribution for the differences
- 10.2: Two Independent Population Proportions

## E

Equal variance
- 11.1: Testing the Hypothesis that β = 0

event
- 4.1: Probability Experiments and Sample Spaces

expected value
- 5.3: Expectation, Variance and Standard Deviation

extrapolation
- 3.4: Prediction

## F

F distribution
- 13.1: Prelude to F Distribution and One-Way ANOVA

Frequency Polygons
- 2.2: Organizing and Graphing Quantitative Data

## G

geometric distribution
- 5.5: The Geometric Distribution

goodness of fit
- 12.2: A Goodness-of-Fit Test

## H

histograms
- 2.2: Organizing and Graphing Quantitative Data

homogeneity
- 12.5: Test for Homogeneity

Hypergeometric Distribution
- 5.6: The Hypergeometric Distribution

Hypothesis Testing
- 9.1: Hypothesis Tests- An Introduction
- 9.3: Hypothesis Tests about μ- p-value Approach

## I

independent events
- 4.3: Conditional Probability and Independence
- 4.4: Counting Basics- the Multiplication and Addition Rules
- 12.3: A Test of Independence or Homogeneity

interpolation
- 3.4: Prediction

interval of interest
- 5.7: The Poisson Distribution

## L

line graph
- 2.3: Stem-and-Leaf Displays

linear correlation coefficient
- 3.3: Simple Linear Regression
- 11.1: Testing the Hypothesis that β = 0

linear equations
- 3.1: Linear Equations

LINEAR REGRESSION MODEL
- 3.3: Simple Linear Regression

## M

margin of error
- 8.1: Estimating Population Means

matched samples
- 10.3: Matched or Paired Samples

mean
- 2.8: Skewness and the Mean, Median, and Mode
- 5.3: Expectation, Variance and Standard Deviation

mean for sums
- 7.2: The Sum Distribution

median
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.5: Measures of Position- Percentiles and Quartiles
- 2.8: Skewness and the Mean, Median, and Mode

mode
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.8: Skewness and the Mean, Median, and Mode

Multiplying probabilities
- 4.4: Counting Basics- the Multiplication and Addition Rules

mutually exclusive
- 4.3: Conditional Probability and Independence
- 4.4: Counting Basics- the Multiplication and Addition Rules

## N

normal distribution
- 6.4: Applications of Finding Normal Probabilities
- 7.1: The Sample Mean and Sources of Error

## O

outcome
- 4.1: Probability Experiments and Sample Spaces

outliers
- 2.5: Measures of Position- Percentiles and Quartiles
- 3.5: Outliers

## P

paired samples
- 10.3: Matched or Paired Samples

parameter
- 1.2: Key Terms and Definitions

Pareto chart
- 1.3: Populations and Samples

Poisson distribution
- 5.7: The Poisson Distribution

Pooled Proportion
- 10.2: Two Independent Population Proportions

pooled variance
- 13.3: The F Distribution and the F-Ratio

population
- 1.2: Key Terms and Definitions

population mean
- 2.4: Measures of Central Tendency- Mean, Median and Mode

Population Standard Deviation
- 2.7: Measures of Spread- Variance and Standard Deviation

power of the test
- 9.2: Type I and Type II Errors

prediction
- 3.4: Prediction

probability
1.2: Key Terms and Definitions
probability distribution function
5.2: The Probability Distribution Function
6.4: Applications of Finding Normal Probabilities

## Q

qualitative data
1.3: Populations and Samples
quantitative data
1.3: Populations and Samples
quartiles
2.5: Measures of Position- Percentiles and Quartiles

## R

rare events
9.4: Hypothesis Tests about μ- Critical Region Approach

## S

sample mean
2.4: Measures of Central Tendency- Mean, Median and Mode
sample space
4.1: Probability Experiments and Sample Spaces
sample standard deviation
2.7: Measures of Spread- Variance and Standard Deviation
sampling
1: Sampling and Data
sampling bias
1.3: Populations and Samples
sampling distribution of the mean
7.1: The Sample Mean and Sources of Error
Sampling Error
1.3: Populations and Samples
sampling with replacement
1.3: Populations and Samples
4.3: Conditional Probability and Independence
4.5: Intersection and Union of Events and Venn Diagrams
sampling without replacement
1.3: Populations and Samples
4.3: Conditional Probability and Independence
4.5: Intersection and Union of Events and Venn Diagrams
scatter plot
3.2: Scatter Plots
significance level
9.4: Hypothesis Tests about μ- Critical Region Approach
Skewed
2.6: Box Plots
2.8: Skewness and the Mean, Median, and Mode
slope
3.1: Linear Equations
standard deviation
2.7: Measures of Spread- Variance and Standard Deviation
5.3: Expectation, Variance and Standard Deviation
Standard deviation for Sums
7.2: The Sum Distribution
standard error
10.1: Two Population Means
standard error of the mean
7.1: The Sample Mean and Sources of Error
standard normal distribution
6.3: The Standard Normal Distribution
statistic
1.2: Key Terms and Definitions
stemplot
2.3: Stem-and-Leaf Displays
strength of a relationship between the variables
3.2: Scatter Plots

## T

test for homogeneity
12.5: Test for Homogeneity
test statistic
10.3: Matched or Paired Samples
The alternative hypothesis
9.1: Hypothesis Tests- An Introduction
The AND Event
4.1: Probability Experiments and Sample Spaces
the central limit theorem
7: Sampling Distributions
The null hypothesis
9.1: Hypothesis Tests- An Introduction
The Or Event
4.1: Probability Experiments and Sample Spaces
The OR of Two Events
4.3: Conditional Probability and Independence
Time Series Graphs
2.2: Organizing and Graphing Quantitative Data
tree diagram
4.5: Intersection and Union of Events and Venn Diagrams
Type I error
9.2: Type I and Type II Errors
Type II error
9.2: Type I and Type II Errors

## U

uniform distribution
6.2: The Uniform and Other Simple Continuous Distributions

## V

variable
1.2: Key Terms and Definitions
variation due to error or unexplained variation
13.3: The F Distribution and the F-Ratio
variation due to treatment or explained variation
13.3: The F Distribution and the F-Ratio
Venn diagram
4.5: Intersection and Union of Events and Venn Diagrams

# Glossary

## Analysis of Vardelmar

**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: (1) all populations of interest are normally distributed. (2) the populations have equal standard deviations. (3) samples (not necessarily of the same size) are randomly and independently selected from each population. (4) The test statistic for analysis of variance is the $F$-ratio. [OpenStax]

**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. [OpenStax]

**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). [OpenStax]

**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 $P(X = x) = \binom{n}{x} p^x q^{n-x}$. [OpenStax]

**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. [OpenStax]

**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 $P(X = x) = \binom{n}{x} p^x q^{n-x}$. [OpenStax]

**Blinding** | not telling participants which treatment a subject is receiving [OpenStax]

**Box plot** | a graph that gives a quick picture of the middle 50% of the data [OpenStax]

**Categorical Variable** | variables that take on values that are names or labels [OpenStax]

**Central Limit Theorem** | Given a random variable (RV) with known mean $\mu$ and known standard deviation, $\sigma$, 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. [OpenStax]

**Central Limit Theorem** | Given a random variable (RV) with known mean $\mu$ and known standard deviation $\sigma$. 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. [OpenStax]

**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. [OpenStax]

**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:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \quad (1)$$

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 $y$. [OpenStax]

**Conditional Probability** | the likelihood that an event will occur given that another event has already occurred [OpenStax]

**Confidence Interval (CI)** | an interval estimate for an unknown population parameter. This depends on: (1) The desired confidence level. (2) Information that is known about the distribution (for example, known standard deviation). (3) The sample and its size. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**Continuous Random Variable** | a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV. [OpenStax]

**Control Group** | a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups [OpenStax]

**Convenience Sampling** | a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data. [OpenStax]

**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. [OpenStax]

**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) [OpenStax]

**decay parameter** | The decay parameter describes the rate at which probabilities decay to zero for increasing values of $x$. It is the value $m$ in the probability density function $f(x) = me^{(-mx)}$ of an exponential random variable. It is also equal to $m = \frac{1}{\mu}$, where $\mu$ is the mean of the random variable. [OpenStax]

**Degrees of Freedom (df)** | the number of objects in a sample that are free to vary. [OpenStax]

**Dependent Events** | If two events are NOT independent, then we say that they are dependent. [OpenStax]

**Discrete Random Variable** | a random variable (RV) whose outcomes are counted [OpenStax]

**Double-blinding** | the act of blinding both the subjects of an experiment and the researchers who work with the subjects [OpenStax]

**Equally Likely** | Each outcome of an experiment has the same probability. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**Expected Value** | expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: $\mu$. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$. [OpenStax]

**Experiment** | a planned activity carried out under controlled conditions [OpenStax]

**Experimental Unit** | any individual or object to be measured [OpenStax]

**Explanatory Variable** | the independent variable in an experiment; the value controlled by researchers [OpenStax]

**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 $P(X \leq x) = 1 - e^{mx}$. [OpenStax]

**First Quartile** | the value that is the median of the of the lower half of the ordered data set [OpenStax]

**Frequency** | the number of times a value of the data occurs [OpenStax]

**Frequency Polygon** | looks like a line graph but uses intervals to display ranges of large amounts of data [OpenStax]

**Frequency Table** | a data representation in which grouped data is displayed along with the corresponding frequencies [OpenStax]

**Geometric Distribution** | a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable $X$ is defined as the number of trials until the first success. Notation: $X \sim G(p)$. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly $x$ failures before the first success is given by the formula: $P(X = x) = p(1-p)^{x-1}$. [OpenStax]

**Geometric Experiment** | a statistical experiment with the following properties: (1) There are one or more Bernoulli trials with all failures except the last one, which is a success. (2) In theory, the number of trials could go on forever. There must be at least one trial. (3) The probability, $p$, of a success and the probability, $q$, of a failure do not change from trial to trial [OpenStax]

**Hypergeometric Experiment** | a statistical experiment with the following properties: (1) You take samples from two groups. (2) You are concerned with a group of interest, called the first group. (3) You sample without replacement from the combined groups. (4) Each pick is not independent, since sampling is without replacement. (5) You are not dealing with Bernoulli Trials. [OpenStax]

**Hypergeometric Probability** | a discrete random variable (RV) that is characterized by: (1) A fixed number of trials. (2) The probability of success is not the same from trial to trial. We sample from two groups of items when we are interested in only one group. $X$ is defined as the number of successes out of the total number of items chosen. Notation: $X \sim H(r, b, n)$, where $r$ = the number of items in the group of interest, $b$ = the number of items in the group not of interest, and $n$ = the number of items chosen. [OpenStax]

**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 $H_a$). [OpenStax]

**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. [OpenStax]

**Independent Events** | The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true: (1) $P(\text{A}|\text{B}) = P(\text{A})$, (2) $P(\text{B}|\text{A}) = P(\text{B})$, (3) $P(\text{A AND B}) = P(\text{A})P(\text{B})$ [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**Institutional Review Board** | a committee tasked with oversight of research programs that involve human subjects [OpenStax]

**Interval** | also called a class interval; an interval represents a range of data and is used when displaying large data sets [OpenStax]

**Level of Significance of the Test** | probability of a Type I error (reject the null hypothesis when it is true). Notation: $\alpha$. In hypothesis testing, the Level of Significance is called the preconceived $\alpha$ or the preset $\alpha$. [OpenStax]

**Lurking Variable** | a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable [OpenStax]

**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 $\mu$) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$. [OpenStax]

**Mean of a Probability Distribution** | the long-term average of many trials of a statistical experiment [OpenStax]

**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. [OpenStax]

**memoryless property** | For an exponential random variable $X$, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that $X$ exceeds $x + k$, given that it has exceeded $x$, is the same as the probability that $X$ would exceed $k$ if we had no knowledge about it. In symbols we say that $P(X > x + k | X > x) = P(X > k)$ [OpenStax]

**Midpoint** | the mean of an interval in a frequency table [OpenStax]

**Mode** | the value that appears most frequently in a set of data [OpenStax]

**Mutually Exclusive** | Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(\text{A AND B}) = 0$. [OpenStax]

**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. [OpenStax]

**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 $\mu$ is the mean of the distribution and $\sigma$ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called a standard normal distribution. [OpenStax]

**Numerical Variable** | variables that take on values that are indicated by numbers [OpenStax]

**One-Way ANOVA** | a method of testing whether or not the means of three or more populations are equal; the method is applicable if: (1) all populations of interest are normally distributed. (2) the populations have equal standard deviations. (3) samples (not necessarily of the same size) are randomly and independently selected from each population. (4) The test statistic for analysis of variance is the $F$-ratio. [OpenStax]

**Outcome** | a particular result of an experiment [OpenStax]

**Outlier** | an observation that does not fit the rest of the data [OpenStax]

**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. [OpenStax]

**Paired Data Set** | two data sets that have a one to one relationship so that: (1)both data sets are the same size, and (2) each data point in one data set is matched with exactly one point from the other set. [OpenStax]

**Parameter** | a number that is used to represent a population characteristic and that generally cannot be determined easily [OpenStax]

**Parameter** | a numerical characteristic of a population [OpenStax]

**Placebo** | an inactive treatment that has no real effect on the explanatory variable [OpenStax]

**Point Estimate** | a single number computed from a sample and used to estimate a population parameter [OpenStax]

**Poisson distribution** | If there is a known average of $\lambda$ events occurring per unit time, and these events are independent of each other, then the number of events $X$ occurring in one unit of time has the Poisson distribution. The probability of $k$ events occurring in one unit time is equal to $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. [OpenStax]

**Poisson Probability Distribution** | a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable: (1) The probability that the event occurs in a given interval is the same for all intervals. (2) The events occur with a known mean and independently of the time since the last event. The distribution is defined by the mean $\mu$ of the event in the interval. Notation: $X \sim P(\mu)$. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly $x$ successes in $r$ trials is $P(X = x) = (e^{-\mu})\frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when $n$ is "large" and $p$ is "small" (a general rule is that $n$ should be greater than or equal to 20 and $p$ should be less than or equal to 0.05). [OpenStax]

**Pooled Proportion** | estimate of the common value of $p_1$ and $p_2$. [OpenStax]

**Population** | all individuals, objects, or measurements whose properties are being studied [OpenStax]

**Probability** | a number between zero and one, inclusive, that gives the likelihood that a specific event will occur [OpenStax]

**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: (1) $0 \leq P(A) \leq 1$, (2) If A and B are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(B)$ and (3) $P(S) = 1$. [OpenStax]

**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. [OpenStax]

**Proportion** | the number of successes divided by the total number in the sample [OpenStax]

**Qualitative Data** | See Data. [OpenStax]

**Quantitative Data** | See Data. [OpenStax]

**Random Assignment** | the act of organizing experimental units into treatment groups using random methods [OpenStax]

**Random Sampling** | a method of selecting a sample that gives every member of the population an equal chance of being selected. [OpenStax]

**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. (1) 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}. (2) We can tell what specific value $x$ the random variable $X$ takes only after performing the experiment [OpenStax]

**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 [OpenStax]

**Representative Sample** | a subset of the population that has the same characteristics as the population [OpenStax]

**Response Variable** | the dependent variable in an experiment; the value that is measured for change at the end of an experiment [OpenStax]

**Sample** | a subset of the population studied [OpenStax]

**Sample Space** | the set of all possible outcomes of an experiment [OpenStax]

**Sampling Bias** | not all members of the population are equally likely to be selected [OpenStax]

**Sampling Distribution** | Given simple random samples of size $n$ from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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 $\sigma$ for population standard deviation. [OpenStax]

**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 [OpenStax]

**Standard Error of the Mean** | the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$. [OpenStax]

**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)\. [OpenStax]

**Statistic** | a numerical characteristic of the sample; a statistic estimates the corresponding population parameter. [OpenStax]

**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. [OpenStax]

**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: (1) It is continuous and assumes any real values. (2) The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution. (3) It approaches the standard normal distribution as $n$ gets larger. (4) 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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**The Complement Event** | The complement of event A consists of all outcomes that are NOT in A. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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. [OpenStax]

**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*. [OpenStax]

**Treatments** | different values or components of the explanatory variable applied in an experiment [OpenStax]

**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) [OpenStax]

**Type 1 Error** | The decision is to reject the null hypothesis when, in fact, the null hypothesis is true. [OpenStax]

**Type 2 Error** | The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false. [OpenStax]

**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 $P(X \leq x) = \frac{x-a}{b-a}$. [OpenStax]

**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 $P(X \leq x) = \frac{x-a}{b-a}$. [OpenStax]

**Variable** | a characteristic of interest for each person or object in a population [OpenStax]

**Variable (Random Variable)** | 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 $x, y, z,$.... 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. (1) 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}. (2) We can tell what specific value *x* of the random variable $X$ takes only after performing the experiment. [OpenStax]

**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. [OpenStax]

**Venn Diagram** | the visual representation of a sample space and events in the form of circles or ovals showing their intersections [OpenStax]

**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 $\mu$ and standard deviation $\sigma$, the result is called the $z$-score of $x$. The $z$-score allows us to compare data that are normally distributed but scaled differently. [OpenStax]


- **Glossary** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

# Index


## A

Adding probabilities
- 4.4: Counting Basics- the Multiplication and Addition Rules

anova
- 13.2: One-Way ANOVA

## B

bar graph
- 2.3: Stem-and-Leaf Displays

Bernoulli trial
- 5.4: The Binomial Distribution
- 5.5: The Geometric Distribution

binomial probability distribution
- 5.4: The Binomial Distribution
- 8.3: Estimating Proportions

box plots
- 2.6: Box Plots

## C

central limit theorem for sums
- 7.2: The Sum Distribution

cluster sampling
- 1.3: Populations and Samples

coefficient of determination
- 3.3: Simple Linear Regression

Cohen's Standards
- 10.1: Two Population Means

Comparing two population means
- 10.4: Two Population Means with Known Standard Deviations

Comparing Two Population Proportions
- 10.2: Two Independent Population Proportions

complement
- 4.1: Probability Experiments and Sample Spaces
- 4.3: Conditional Probability and Independence

conditional probability
- 4.1: Probability Experiments and Sample Spaces

contingency table
- 4.6: Joint and Marginal Probabilities and Contingency Tables
- 12.3: A Test of Independence or Homogeneity

continuous data
- 1.3: Populations and Samples

## D

Decision
- 9.4: Hypothesis Tests about μ- Critical Region Approach

degrees of freedom
- 10.1: Two Population Means

direction of a relationship between the variables
- 3.2: Scatter Plots

discrete data
- 1.3: Populations and Samples

Distribution for the differences
- 10.2: Two Independent Population Proportions

## E

Equal variance
- 11.1: Testing the Hypothesis that β = 0

event
- 4.1: Probability Experiments and Sample Spaces

expected value
- 5.3: Expectation, Variance and Standard Deviation

extrapolation
- 3.4: Prediction

## F

F distribution
- 13.1: Prelude to F Distribution and One-Way ANOVA

Frequency Polygons
- 2.2: Organizing and Graphing Quantitative Data

## G

geometric distribution
- 5.5: The Geometric Distribution

goodness of fit
- 12.2: A Goodness-of-Fit Test

## H

histograms
- 2.2: Organizing and Graphing Quantitative Data

homogeneity
- 12.5: Test for Homogeneity

Hypergeometric Distribution
- 5.6: The Hypergeometric Distribution

Hypothesis Testing
- 9.1: Hypothesis Tests- An Introduction
- 9.3: Hypothesis Tests about μ- p-value Approach

## I

independent events
- 4.3: Conditional Probability and Independence
- 4.4: Counting Basics- the Multiplication and Addition Rules
- 12.3: A Test of Independence or Homogeneity

interpolation
- 3.4: Prediction

interval of interest
- 5.7: The Poisson Distribution

## L

line graph
- 2.3: Stem-and-Leaf Displays

linear correlation coefficient
- 3.3: Simple Linear Regression
- 11.1: Testing the Hypothesis that β = 0

linear equations
- 3.1: Linear Equations

LINEAR REGRESSION MODEL
- 3.3: Simple Linear Regression

## M

margin of error
- 8.1: Estimating Population Means

matched samples
- 10.3: Matched or Paired Samples

mean
- 2.8: Skewness and the Mean, Median, and Mode
- 5.3: Expectation, Variance and Standard Deviation

mean for sums
- 7.2: The Sum Distribution

median
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.5: Measures of Position- Percentiles and Quartiles
- 2.8: Skewness and the Mean, Median, and Mode

mode
- 2.4: Measures of Central Tendency- Mean, Median and Mode
- 2.8: Skewness and the Mean, Median, and Mode

Multiplying probabilities
- 4.4: Counting Basics- the Multiplication and Addition Rules

mutually exclusive
- 4.3: Conditional Probability and Independence
- 4.4: Counting Basics- the Multiplication and Addition Rules

## N

normal distribution
- 6.4: Applications of Finding Normal Probabilities
- 7.1: The Sample Mean and Sources of Error

## O

outcome
- 4.1: Probability Experiments and Sample Spaces

outliers
- 2.5: Measures of Position- Percentiles and Quartiles
- 3.5: Outliers

## P

paired samples
- 10.3: Matched or Paired Samples

parameter
- 1.2: Key Terms and Definitions

Pareto chart
- 1.3: Populations and Samples

Poisson distribution
- 5.7: The Poisson Distribution

Pooled Proportion
- 10.2: Two Independent Population Proportions

pooled variance
- 13.3: The F Distribution and the F-Ratio

population
- 1.2: Key Terms and Definitions

population mean
- 2.4: Measures of Central Tendency- Mean, Median and Mode

Population Standard Deviation
- 2.7: Measures of Spread- Variance and Standard Deviation

power of the test
- 9.2: Type I and Type II Errors

prediction
- 3.4: Prediction

probability
1.2: Key Terms and Definitions
probability distribution function
5.2: The Probability Distribution Function
6.4: Applications of Finding Normal Probabilities

## Q

qualitative data
1.3: Populations and Samples
quantitative data
1.3: Populations and Samples
quartiles
2.5: Measures of Position- Percentiles and Quartiles

## R

rare events
9.4: Hypothesis Tests about μ- Critical Region Approach

## S

sample mean
2.4: Measures of Central Tendency- Mean, Median and Mode
sample space
4.1: Probability Experiments and Sample Spaces
sample standard deviation
2.7: Measures of Spread- Variance and Standard Deviation
sampling
1: Sampling and Data
sampling bias
1.3: Populations and Samples
sampling distribution of the mean
7.1: The Sample Mean and Sources of Error
Sampling Error
1.3: Populations and Samples
sampling with replacement
1.3: Populations and Samples
4.3: Conditional Probability and Independence
4.5: Intersection and Union of Events and Venn Diagrams
sampling without replacement
1.3: Populations and Samples
4.3: Conditional Probability and Independence
4.5: Intersection and Union of Events and Venn Diagrams
scatter plot
3.2: Scatter Plots
significance level
9.4: Hypothesis Tests about μ- Critical Region Approach
Skewed
2.6: Box Plots
2.8: Skewness and the Mean, Median, and Mode
slope
3.1: Linear Equations
standard deviation
2.7: Measures of Spread- Variance and Standard Deviation
5.3: Expectation, Variance and Standard Deviation
Standard deviation for Sums
7.2: The Sum Distribution
standard error
10.1: Two Population Means
standard error of the mean
7.1: The Sample Mean and Sources of Error
standard normal distribution
6.3: The Standard Normal Distribution
statistic
1.2: Key Terms and Definitions
stemplot
2.3: Stem-and-Leaf Displays
strength of a relationship between the variables
3.2: Scatter Plots

## T

test for homogeneity
12.5: Test for Homogeneity
test statistic
10.3: Matched or Paired Samples
The alternative hypothesis
9.1: Hypothesis Tests- An Introduction
The AND Event
4.1: Probability Experiments and Sample Spaces
the central limit theorem
7: Sampling Distributions
The null hypothesis
9.1: Hypothesis Tests- An Introduction
The Or Event
4.1: Probability Experiments and Sample Spaces
The OR of Two Events
4.3: Conditional Probability and Independence
Time Series Graphs
2.2: Organizing and Graphing Quantitative Data
tree diagram
4.5: Intersection and Union of Events and Venn Diagrams
Type I error
9.2: Type I and Type II Errors
Type II error
9.2: Type I and Type II Errors

## U

uniform distribution
6.2: The Uniform and Other Simple Continuous Distributions

## V

variable
1.2: Key Terms and Definitions
variation due to error or unexplained variation
13.3: The F Distribution and the F-Ratio
variation due to treatment or explained variation
13.3: The F Distribution and the F-Ratio
Venn diagram
4.5: Intersection and Union of Events and Venn Diagrams

# Glossary

**Sample Word 1** | Sample Definition 1

# Detailed Licensing

## Overview

**Title:** Introductory Statistics with Probability (CUNY)

**Webpages:** 87

**All licenses found:**

- CC BY 4.0: 88.5% (77 pages)
- Undeclared: 9.2% (8 pages)
- CC BY-SA 3.0: 2.3% (2 pages)

## By Page


- Introductory Statistics with Probability (CUNY) - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Sampling and Data - *CC BY 4.0*
    - 1.1: Introduction to Probability and Statistics - *CC BY 4.0*
    - 1.2: Key Terms and Definitions - *CC BY 4.0*
    - 1.3: Populations and Samples - *CC BY 4.0*
  - 2: Descriptive Statistics - *CC BY 4.0*
    - 2.1: Organizing and Graphing Qualitative Data - *CC BY 4.0*
    - 2.2: Organizing and Graphing Quantitative Data - *CC BY 4.0*
    - 2.3: Stem-and-Leaf Displays - *CC BY 4.0*
    - 2.4: Measures of Central Tendency- Mean, Median and Mode - *CC BY 4.0*
    - 2.5: Measures of Position- Percentiles and Quartiles - *CC BY 4.0*
    - 2.6: Box Plots - *CC BY 4.0*
    - 2.7: Measures of Spread- Variance and Standard Deviation - *CC BY 4.0*
    - 2.8: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
  - 3: Introduction to Linear Regression and Correlation - *CC BY 4.0*
    - 3.1: Linear Equations - *CC BY 4.0*
    - 3.2: Scatter Plots - *CC BY 4.0*
    - 3.3: Simple Linear Regression - *CC BY 4.0*
    - 3.4: Prediction - *CC BY 4.0*
    - 3.5: Outliers - *CC BY 4.0*
  - 4: Probability Theory - *CC BY 4.0*
    - 4.1: Probability Experiments and Sample Spaces - *CC BY 4.0*
    - 4.2: Experiments Having Equally Likely Outcomes - *Undeclared*
    - 4.3: Conditional Probability and Independence - *CC BY 4.0*
    - 4.4: Counting Basics- the Multiplication and Addition Rules - *CC BY 4.0*
    - 4.5: Intersection and Union of Events and Venn Diagrams - *CC BY 4.0*
    - 4.6: Joint and Marginal Probabilities and Contingency Tables - *CC BY 4.0*
    - 4.7: More Counting- Factorials, Combinations, and Permutations - *Undeclared*
  - 5: Discrete Random Variables - *CC BY 4.0*
    - 5.1: Introduction to Random Variables - *CC BY 4.0*
    - 5.2: The Probability Distribution Function - *CC BY 4.0*
    - 5.3: Expectation, Variance and Standard Deviation - *CC BY 4.0*
    - 5.4: The Binomial Distribution - *CC BY 4.0*
    - 5.5: The Geometric Distribution - *CC BY 4.0*
    - 5.6: The Hypergeometric Distribution - *CC BY 4.0*
    - 5.7: The Poisson Distribution - *CC BY 4.0*
  - 6: Continuous Random Variables - *CC BY 4.0*
    - 6.1: Probability Density Functions - *CC BY 4.0*
    - 6.2: The Uniform and Other Simple Continuous Distributions - *CC BY 4.0*
    - 6.3: The Standard Normal Distribution - *CC BY 4.0*
    - 6.4: Applications of Finding Normal Probabilities - *CC BY 4.0*
  - 7: Sampling Distributions - *CC BY 4.0*
    - 7.1: The Sample Mean and Sources of Error - *CC BY 4.0*
    - 7.2: The Sum Distribution - *CC BY 4.0*

- 8: Confidence Intervals - *CC BY 4.0*
  - 8.1: Estimating Population Means - *CC BY 4.0*
  - 8.2: The t-distribution - *CC BY 4.0*
  - 8.3: Estimating Proportions - *CC BY 4.0*
  - 8.4: Confidence Intervals - *CC BY-SA 3.0*
- 9: Hypothesis Testing for a Single Variable and Population - *CC BY 4.0*
  - 9.1: Hypothesis Tests- An Introduction - *CC BY 4.0*
  - 9.2: Type I and Type II Errors - *CC BY 4.0*
  - 9.3: Hypothesis Tests about μ- p-value Approach - *CC BY 4.0*
  - 9.4: Hypothesis Tests about μ- Critical Region Approach - *CC BY 4.0*
  - 9.5: Hypothesis Tests for a Proportion - *CC BY 4.0*
- 10: Hypothesis Testing for Paired and Unpaired Data - *CC BY 4.0*
  - 10.1: Two Population Means - *CC BY 4.0*
  - 10.2: Two Independent Population Proportions - *CC BY 4.0*
  - 10.3: Matched or Paired Samples - *CC BY 4.0*
  - 10.4: Two Population Means with Known Standard Deviations - *CC BY 4.0*
  - 10.5: Difference of Two Means - *CC BY-SA 3.0*
- 11: Linear Regression and Hypothesis Testing - *Undeclared*
  - 11.1: Testing the Hypothesis that β = 0 - *CC BY 4.0*
- 12: The Chi-Square Distribution - *CC BY 4.0*
  - 12.1: The Chi-Square Distribution - *CC BY 4.0*
  - 12.2: A Goodness-of-Fit Test - *CC BY 4.0*
  - 12.3: A Test of Independence or Homogeneity - *CC BY 4.0*
  - 12.4: Test of a Single Variance - *CC BY 4.0*
  - 12.5: Test for Homogeneity - *CC BY 4.0*
  - 12.6: Comparison of the Chi-Square Tests - *CC BY 4.0*
- 13: F Distribution and One-Way ANOVA - *CC BY 4.0*
  - 13.1: Prelude to F Distribution and One-Way ANOVA - *CC BY 4.0*
  - 13.2: One-Way ANOVA - *CC BY 4.0*
  - 13.3: The F Distribution and the F-Ratio - *CC BY 4.0*
  - 13.4: Facts About the F Distribution - *CC BY 4.0*
  - 13.5: Test of Two Variances - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
  - Index - *CC BY 4.0*
  - Glossary - *CC BY 4.0*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*