SMC: SOCIAL SCIENCES STATISTICS



Bettina Spencer
Saint Mary's College



Saint Mary's College

SMC: Social Sciences Statistics

Bettina Spencer

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 04/16/2026

TABLE OF CONTENTS

Licensing

1: Introduction to Statistics

- 1.1: Basic Definitions and Concepts
- 1.2: Overview
- 1.3: Presentation of Data
- 1.4: Introduction to Statistics (Exercises)

2: Sampling and Data

- 2.1: Introduction
- 2.2: Definitions of Statistics, Probability, and Key Terms
- 2.3: Data, Sampling, and Variation in Data and Sampling
- 2.4: Frequency, Frequency Tables, and Levels of Measurement
- 2.5: Experimental Design and Ethics
- 2.6: Data Collection Experiment (Worksheet)
- 2.7: Sampling Experiment (Worksheet)
- 2.8: Sampling and Data (Exercises)

3: Descriptive Statistics

- 3.1: Prelude to Descriptive Statistics
- 3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
- 3.3: Histograms, Frequency Polygons, and Time Series Graphs
- 3.4: Measures of the Location of the Data
  - Measures of the Location of the Data (Exercises)
- 3.5: Box Plots
- 3.6: Measures of the Center of the Data
- 3.7: Skewness and the Mean, Median, and Mode
- 3.8: Measures of the Spread of the Data
- 3.9: Descriptive Statistics (Worksheet)
- 3.10: Descriptive Statistics (Exercises)

4: Graphing Distributions

- 4.1: Graphing Qualitative Variables
- 4.2: Quantitative Variables
- 4.3: Stem and Leaf Displays
- 4.4: Histograms
- 4.5: Frequency Polygons
- 4.6: Box Plots
- 4.7: Box Plot Demo
- 4.8: Bar Charts
- 4.9: Line Graphs
- 4.10: Dot Plots
- 4.11: Statistical Literacy
- 4.12: Graphing Distributions (Exercises)

5: Probability Topics

- 5.1: Introduction
- 5.2: Terminology
- 5.3: Independent and Mutually Exclusive Events
- 5.4: Two Basic Rules of Probability
- 5.5: Contingency Tables
- 5.6: Tree and Venn Diagrams
- 5.7: Probability Topics (Worksheet)
- 5.8: Probability Topics (Exericses)

6: The Normal Distribution

- 6.1: Prelude to The Normal Distribution
- 6.2: The Standard Normal Distribution
  - The Standard Normal Distribution (Exercises)
- 6.3: Using the Normal Distribution
- 6.4: Normal Distribution - Lap Times (Worksheet)
- 6.5: Normal Distribution - Pinkie Length (Worksheet)
- 6.6: The Normal Distribution (Exercises)

7: Sampling Distributions

- 7.1: The Mean and Standard Deviation of the Sample Mean
- 7.2: The Sampling Distribution of the Sample Mean
- 7.3: The Sample Proportion
- 7.4: Sampling Distributions (Exercises)

8: Power

- 8.1: Introduction to Power
- 8.2: Example Calculations
- 8.3: Power Demo
- 8.4: Power Demo II
- 8.5: Factors Affecting Power
- 8.6: Statistical Literacy
- 8.7: Power (Exercises)

9: Effect Size

- 9.1: Prelude to Effect Size
- 9.2: Proportions
- 9.3: Difference Between Two Means
- 9.4: Proportion of Variance Explained
- 9.5: Statistical Literacy
- 9.6: Effect Size (Exercises)

10: Confidence Intervals

- 10.1: Prelude to Confidence Intervals
- 10.2: A Single Population Mean using the Normal Distribution
  - A Single Population Mean using the Normal Distribution (Exercises)

- 10.3: A Single Population Mean using the Student t-Distribution
- 10.4: A Population Proportion
- 10.5: Confidence Interval - Home Costs (Worksheet)
- 10.6: Confidence Interval -Place of Birth (Worksheet)
- 10.7: Confidence Interval -Women's Heights (Worksheet)
- 10.8: Confidence Intervals (Exercises)
- 10.9: Confidence Intervals (Summary)

11: Hypothesis Testing with One Sample

- 11.1: Prelude to Hypothesis Testing
- 11.2: Null and Alternative Hypotheses
  - Null and Alternative Hypotheses (Exercises)
- 11.3: Outcomes and the Type I and Type II Errors
  - Outcomes and the Type I and Type II Errors (Exercises)
- 11.4: Distribution Needed for Hypothesis Testing
  - Distribution Needed for Hypothesis Testing (Exercises)
- 11.5: Rare Events, the Sample, Decision and Conclusion
  - Rare Events, the Sample, Decision and Conclusion (Exercises)
- 11.6: Additional Information and Full Hypothesis Test Examples
- 11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)
- 11.8: Hypothesis Testing with One Sample (Exercises)

12: Hypothesis Testing with Two Samples

- 12.1: Prelude to Hypothesis Testing with Two Samples
- 12.2: Two Population Means with Unknown Standard Deviations
- 12.3: Two Population Means with Known Standard Deviations
- 12.4: Comparing Two Independent Population Proportions
- 12.5: Matched or Paired Samples
- 12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet)
- 12.7: Hypothesis Testing with Two Samples (Exercises)

13: The Chi-Square Distribution

- 13.1: Prelude to The Chi-Square Distribution
- 13.2: Facts About the Chi-Square Distribution
- 13.3: Goodness-of-Fit Test
- 13.4: Test of Independence
- 13.5: Test for Homogeneity
- 13.6: Comparison of the Chi-Square Tests
- 13.7: Test of a Single Variance
- 13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet)
- 13.9: Lab 2: Chi-Square Test of Independence (Worksheet)
- 13.10: The Chi-Square Distribution (Exercises)

14: Linear Regression and Correlation

- 14.1: Prelude to Linear Regression and Correlation
- 14.2: Linear Equations

- Linear Equations (Exercises)
- 14.3: Scatter Plots
  - Scatter Plots (Exercises)
- 14.4: The Regression Equation
  - The Regression Equation (Exercise)
- 14.5: Testing the Significance of the Correlation Coefficient
  - Testing the Significance of the Correlation Coefficient (Exercises)
- 14.6: Prediction
  - Prediction (Exercises)
- 14.7: Outliers
  - Outliers (Exercises)
- 14.8: Regression - Distance from School (Worksheet)
- 14.9: Regression - Textbook Cost (Worksheet)
- 14.10: Regression - Fuel Efficiency (Worksheet)
- 14.11: Linear Regression and Correlation (Exercises)

15: F Distribution and One-Way ANOVA

- 15.1: Prelude to F Distribution and One-Way ANOVA
- 15.2: One-Way ANOVA
- 15.3: The F Distribution and the F-Ratio
- 15.4: Facts About the F Distribution
- 15.5: Test of Two Variances
- 15.6: Lab: One-Way ANOVA
- 15.7: F Distribution and One-Way ANOVA (Exercises)

16: Advanced Graphs

- 16.1: Q-Q Plots
- 16.2: Contour Plots
- 16.3: 3D Plots
- 16.4: Statistical Literacy
- 16.5: Advanced Graphs (Exercises)

Index

Glossary

Detailed Licensing

Licensing

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

CHAPTER OVERVIEW

1: Introduction to Statistics

Template:TopShafer

In this chapter we will introduce some basic terminology and lay the groundwork for the course. We will explain in general terms what statistics and probability are and the problems that these two areas of study are designed to solve.

1.1: Basic Definitions and Concepts

1.2: Overview

1.3: Presentation of Data

1.4: Introduction to Statistics (Exercises)

1: Introduction to Statistics is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

1.1: Basic Definitions and Concepts

Skills to Develop

- To learn the basic definitions used in statistics and some of its key concepts.

We begin with a simple example. There are millions of passenger automobiles in the United States. What is their average value? It is obviously impractical to attempt to solve this problem directly by assessing the value of every single car in the country, add up all those values, then divide by the number of values, one for each car. In practice the best we can do would be to *estimate* the average value. A natural way to do so would be to randomly select *some* of the cars, say 200 of them, ascertain the value of each of those cars, and find the average of those 200 values. The set of all those millions of vehicles is called the *population* of interest, and the number attached to each one, its value, is a *measurement*. The average value is a *parameter*: a number that describes a characteristic of the population, in this case monetary worth. The set of 200 cars selected from the population is called a *sample*, and the 200 numbers, the monetary values of the cars we selected, are the *sample data*. The average of the data is called a *statistic*: a number calculated from the sample data. This example illustrates the meaning of the following definitions.

Definitions: populations and samples

A **population** is any specific collection of objects of interest. A **sample** is any subset or subcollection of the population, including the case that the sample consists of the whole population, in which case it is termed a census.

Definitions: measurements and Sample Data

A **measurement** is a number or attribute computed for each member of a population or of a sample. The measurements of sample elements are collectively called the **sample data**.

Definition: parameters

A **parameter** is a number that summarizes some aspect of the population as a whole. A **statistic** is a number computed from the sample data.

Continuing with our example, if the average value of the cars in our sample was $\$8,357$, then it seems reasonable to conclude that the average value of all cars is about $\$8,357$. In reasoning this way we have drawn an inference about the *population* based on information obtained from the *sample*. In general, *statistics* is a study of data: describing properties of the data, which is called ***descriptive statistics***, and drawing conclusions about a population of interest from information extracted from a sample, which is called ***inferential statistics***. Computing the single number $\$8,357$ to summarize the data was an operation of descriptive statistics; using it to make a statement about the population was an operation of inferential statistics.

Definition: Statistics

Statistics is a collection of methods for collecting, displaying, analyzing, and drawing conclusions from data.

Definition: Descriptive statistics

Descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data.

Definition: Inferential statistics

Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population.

The measurement made on each element of a sample need not be numerical. In the case of automobiles, what is noted about each car could be its color, its make, its body type, and so on. Such data are categorical or qualitative, as opposed to numerical or quantitative data such as value or age. This is a general distinction.

Definition: Qualitative data

Qualitative data are measurements for which there is no natural numerical scale, but which consist of attributes, labels, or other non-numerical characteristics.

Definition: Quantitative data

Quantitative data are numerical measurements that arise from a natural numerical scale.

Qualitative data can generate numerical sample statistics. In the automobile example, for instance, we might be interested in the proportion of all cars that are less than six years old. In our same sample of 200 cars we could note for each car whether it is less than six years old or not, which is a qualitative measurement. If 172 cars in the sample are less than six years old, which is 0.86 or 86%, then we would estimate the parameter of interest, the population proportion, to be about the same as the sample statistic, the sample proportion, that is, about 0.86.

The relationship between a population of interest and a sample drawn from that population is perhaps the most important concept in statistics, since everything else rests on it. This relationship is illustrated graphically in Figure 1.1.1. The circles in the large box represent elements of the population. In the figure there was room for only a small number of them but in actual situations, like our automobile example, they could very well number in the millions. The solid black circles represent the elements of the population that are selected at random and that together form the sample. For each element of the sample there is a measurement of interest, denoted by a lower case x (which we have indexed as $x_1, \ldots, x_n$ to tell them apart); these measurements collectively form the sample data set. From the data we may calculate various statistics. To anticipate the notation that will be used later, we might compute the sample mean $\bar{x}$ and the sample proportion $\hat{p}$, and take them as approximations to the population mean μ (this is the lower case Greek letter mu, the traditional symbol for this parameter) and the population proportion p, respectively. The other symbols in the figure stand for other parameters and statistics that we will encounter.

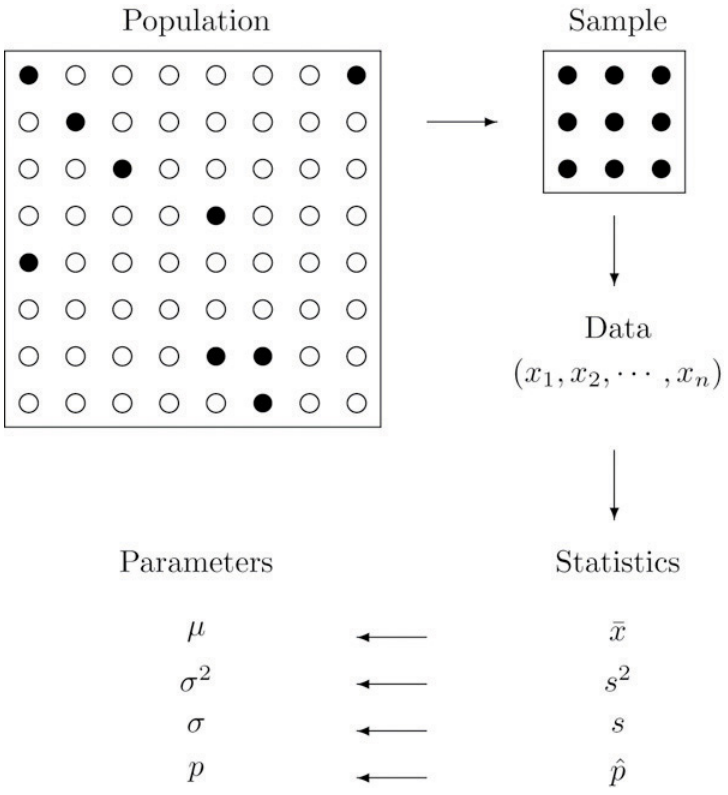


Figure* 1.1.1*: *The Grand Picture of Statistics*

Key Takeaway

- Statistics is a study of data: describing properties of data (descriptive statistics) and drawing conclusions about a population based on information in a sample (inferential statistics).
- The distinction between a population together with its parameters and a sample together with its statistics is a fundamental concept in inferential statistics.
- Information in a sample is used to make inferences about the population from which the sample was drawn.

Contributor

- Template:ContribShaferZhang

1.1: Basic Definitions and Concepts is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

1.2: Overview

Skills to Develop

- To obtain an overview of the material in the text.

The example we gave in the first section seems fairly simple, but it illustrates some significant problems. We supposed that the 200 cars of the sample had an average value of $\$8,357$ (a number that is precisely known), and concluded that the population has an average of about the same amount, although its precise value is still unknown. What would happen if someone else were to take another sample of exactly the same size from exactly the same population? Would he or she get the same sample average as we did, $\$8,357$? Almost surely not. In fact, if the investigator who took the second sample reported precisely the same value, we would immediately become suspicious of his result. The sample average is an example of what is called a *random variable*: a number that varies from trial to trial of an experiment (in this case, from sample to sample), and does so in a way that cannot be predicted precisely. Random variables will be a central object of study for us, beginning in Chapter 4.

Another issue that arises is that different samples have different levels of reliability. We have supposed that our sample of size 200 had an average of $\$8,357$. If a sample of size $1,000$ yielded an average value of $\$7,832$, then we would naturally regard this latter number as probably a better estimate of the average value of all cars, since it came from a larger sample. How can this be expressed? An important idea developed in Chapter 7 is the *confidence interval*: from the data we will construct an interval of values using a process that has a certain chance, say a 95% chance, of generating an interval that contains the true population average. Thus, instead of reporting a single estimate, $\$8,357$, for the population mean we might say that, based on our sample data, we are 95% certain that the true average is within $\$100$ of our sample mean, that is, we are 95% certain that the true average is the between $\$8,257$ and $\$8,457$. The number $\$100$ will be computed from the sample data just as the sample mean $\$8,357$ was. This "95% confidence interval" will automatically indicate the reliability of the estimate that we obtained from the sample. Moreover, to obtain the same chance of containing the unknown parameter, a large sample will typically produce a shorter interval than a small sample will. Thus large samples usually give more accurate results. Unless we perform a *census*, which is a "sample" that includes the entire population, we can never be completely sure of the exact average value of the population. The best that we can do is to make statements of *probability*, an important concept that we will begin to study formally in Chapter 3.

Sampling may be done not only to estimate a population parameter, but to test a claim that is made about that parameter. Suppose a food package asserts that the amount of sugar in one serving of the product is 14 grams. A consumer group might suspect that it actually contains more. How would they test the competing claims about the amount of sugar, "14 grams" versus "more than 14 grams"? They might take a random sample of perhaps 20 food packages, measure the amount of sugar in one serving of each one, and average those amounts. They are not interested in measuring the average amount of sugar in a serving for its own sake; their interest is simply whether the claim about the true amount is accurate. Stated another way, they are sampling not in order to estimate the average amount of sugar in one serving, but to see whether that amount, whatever it may be, is larger than 14 grams. Again because one can have certain knowledge only by taking a census, ideas of probability enter into the analysis. We will examine tests of hypotheses beginning in Chapter 8 .

Several times in this introduction we have used the term “random sample.” Generally the value of our data is only as good as the sample that produced it. For example, suppose we wish to estimate the proportion of all students at a large university who are females, which we denote by p. If we select 50 students at random and 27 of them are female, then a natural estimate is $p \approx \hat{p} = 27/50 = 0.54$ or 54%. How much confidence we can place in this estimate depends not only on the size of the sample, but on its quality, whether or not it is truly random, or at least truly representative of the whole population. If all 50 students in our sample were drawn from a College of Nursing, then the proportion of female students in the sample is likely higher than that of the entire campus. If all 50 students were selected from a College of Engineering Sciences, then the proportion of students in the entire student body who are females could be underestimated. In either case, the estimate would be distorted or biased. In statistical practice an unbiased sampling scheme is important but in most cases not easy to produce. For this introductory course we will assume that all samples are either random or at least representative.

Key Takeaway

- Statistics computed from samples vary randomly from sample to sample. Conclusions made about population parameters are statements of probability.

Contributor

- Template:ContribShaferZhang

1.2: Overview is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

1.3: Presentation of Data

Skills to Develop

- To learn two ways that data will be presented in the text.

In this book we will use two formats for presenting data sets. The first is a data list, which is an explicit listing of all the individual measurements, either as a display with space between the individual measurements, or in set notation with individual measurements separated by commas.

Example 1.3.1

The data obtained by measuring the age of 21 randomly selected students enrolled in freshman courses at a university could be presented as the data list:

$$\begin{array}{cccccccccccc} 18 & 18 & 19 & 19 & 19 & 18 & 22 & 20 & 18 & 18 & 17 \\ 19 & 18 & 24 & 18 & 20 & 18 & 21 & 20 & 17 & 19 & \end{array} \tag{1.3.1}$$

or in set notation as:

$$\{18, 18, 19, 19, 19, 18, 22, 20, 18, 18, 17, 19, 18, 24, 18, 20, 18, 21, 20, 17, 19\} \tag{1.3.2}$$

A data set can also be presented by means of a data frequency table, a table in which each distinct value x is listed in the first row and its frequency f, which is the number of times the value x appears in the data set, is listed below it in the second row.

A data set can also be presented by means of a data frequency table, a table in which each distinct value x is listed in the first row and its frequency f, which is the number of times the value x appears in the data set, is listed below it in the second row.

Example 1.3.2

The data set of the previous example is represented by the data frequency table

x	17	18	19	20	21	22	24
f	2	8	5	3	1	1	1

(1.3.3)

The data frequency table is especially convenient when data sets are large and the number of distinct values is not too large.

Key Takeaway

- Data sets can be presented either by listing all the elements or by giving a table of values and frequencies.

Contributor

- Template:ContribShaferZhang

1.3: Presentation of Data is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

1.4: Introduction to Statistics (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang. Complementary General Chemistry question banks can be found for other Textmaps and can be accessed here. In addition to these publicly available questions, access to private problems bank for use in exams and homework is available to faculty only on an individual basis; please contact Delmar Larsen for an account with access permission.

1.1: Basic Definitions and Concepts

Questions

1. Explain what is meant by the term *population*.
2. Explain what is meant by the term *sample*.
3. Explain how a sample differs from a population.
4. Explain what is meant by the term *sample data*.
5. Explain what a *parameter* is.
6. Explain what a *statistic* is.
7. Give an example of a population and two different characteristics that may be of interest.
8. Describe the difference between *descriptive statistics* and *inferential statistics*. Illustrate with an example.
9. Identify each of the following data sets as either a population or a sample:
    a. The grade point averages (GPAs) of all students at a college.
    b. The GPAs of a randomly selected group of students on a college campus.
    c. The ages of the nine Supreme Court Justices of the United States on $\textbf{January 1, 1842}$.
    d. The gender of every second customer who enters a movie theater.
    e. The lengths of Atlantic croakers caught on a fishing trip to the beach.
10. Identify the following measures as either quantitative or qualitative:
    a. The 30 high-temperature readings of the last 30 days.
    b. The scores of 40 students on an English test.
    c. The blood types of 120 teachers in a middle school.
    d. The last four digits of social security numbers of all students in a class.
    e. The numbers on the jerseys of 53 football players on a team.
11. Identify the following measures as either quantitative or qualitative:
    a. The genders of the first 40 newborns in a hospital one year.
    b. The natural hair color of 20 randomly selected fashion models.
    c. The ages of 20 randomly selected fashion models.
    d. The fuel economy in miles per gallon of 20 new cars purchased last month.
    e. The political affiliation of 500 randomly selected voters.
12. A researcher wishes to estimate the average amount spent per person by visitors to a theme park. He takes a random sample of forty visitors and obtains an average of $\$28$ per person.
    a. What is the population of interest?
    b. What is the parameter of interest?
    c. Based on this sample, do we know the average amount spent per person by visitors to the park? Explain fully.
13. A researcher wishes to estimate the average weight of newborns in South America in the last five years. He takes a random sample of 235 newborns and obtains an average of 3.27 kilograms.
    a. What is the population of interest?
    b. What is the parameter of interest?
    c. Based on this sample, do we know the average weight of newborns in South America? Explain fully.
14. A researcher wishes to estimate the proportion of all adults who own a cell phone. He takes a random sample of $1,572$ adults; $1,298$ of them own a cell phone, hence $1298 / 1572 \approx .83$ or about 83% own a cell phone.
    a. What is the population of interest?

b. What is the parameter of interest?
c. What is the statistic involved?
d. Based on this sample, do we know the proportion of all adults who own a cell phone? Explain fully.

15. A sociologist wishes to estimate the proportion of all adults in a certain region who have never married. In a random sample of $1,320$ adults, 145 have never married, hence $145 / 1320 \approx .11$ or about 11% have never married.
    a. What is the population of interest?
    b. What is the parameter of interest?
    c. What is the statistic involved?
    d. Based on this sample, do we know the proportion of all adults who have never married? Explain fully.
16. a. What must be true of a sample if it is to give a reliable estimate of the value of a particular population parameter?
    b. What must be true of a sample if it is to give *certain* knowledge of the value of a particular population parameter?

Answers

1. A population is the total collection of objects that are of interest in a statistical study.
2.
3. A sample, being a subset, is typically smaller than the population. In a statistical study, all elements of a sample are available for observation, which is not typically the case for a population.
4.
5. A parameter is a value describing a characteristic of a population. In a statistical study the value of a parameter is typically unknown.
6.
7. All currently registered students at a particular college form a population. Two population characteristics of interest could be the average GPA and the proportion of students over 23 years.
8.
9. a. Population.
   b. Sample.
   c. Population.
   d. Sample.
   e. Sample.
10.
11. a. Qualitative.
    b. Qualitative.
    c. Quantitative.
    d. Quantitative.
    e. Qualitative.
12.
13. a. All newborn babies in South America in the last five years.
    b. The average birth weight of all newborn babies in South America in the last five years.
    c. No, not exactly, but we know the approximate value of the average.
14.
15. a. All adults in the region.
    b. The proportion of the adults in the region who have never married.
    c. The proportion computed from the sample, 0.1.
    d. No, not exactly, but we know the approximate value of the proportion.

Contributor

- Template:ContribShaferZhang

1.4: Introduction to Statistics (Exercises) is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

2: 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."

2.1: Introduction
2.2: Definitions of Statistics, Probability, and Key Terms
2.3: Data, Sampling, and Variation in Data and Sampling
2.4: Frequency, Frequency Tables, and Levels of Measurement
2.5: Experimental Design and Ethics
2.6: Data Collection Experiment (Worksheet)
2.7: Sampling Experiment (Worksheet)
2.8: Sampling and Data (Exercises)

Contributors

Template:ContribOpenStax

This page titled 2: Sampling and Data is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.1: Introduction

Skills to Develop

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* 2.1.1*: *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

- Template:ContribOpenStax

This page titled 2.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.2: Definitions of Statistics, Probability, and Key Terms

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:

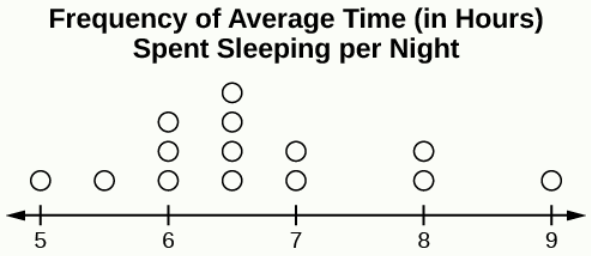


***Figure* 2.2.1**

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 2.2.1

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 2.2.1

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 2.2.2

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 2.2.3

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 2.2.4

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

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

Answer

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

COLLABORATIVE EXERCISE

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

References

1. The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories...stDummies.html (accessed May 1, 2013).

Practice

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

Researcher A:

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

Researcher B:

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

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

Exercise 1.2.2

population

Answer

AIDS patients.

Exercise 1.2.3

sample

Exercise 1.2.4

parameter

Answer

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

Exercise 1.2.5

statistic

Exercise 1.2.6

variable

Answer

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

Glossary

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

Average

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

Categorical Variable

variables that take on values that are names or labels

Data

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

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

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

Population

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

Probability

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

Proportion

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

Representative Sample

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

Sample

a subset of the population studied

Statistic

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

Variable

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

Contributors

- Template:ContribOpenStax

This page titled 2.2: Definitions of Statistics, Probability, and Key Terms is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.3: Data, Sampling, and Variation in Data and Sampling

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

- Qualitative
- Quantitative

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

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

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

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

Sample of Quantitative Discrete Data

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

Exercise 2.3.1

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

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

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

Exercise 2.3.2

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

Answer

quantitative continuous data

Exercise 2.3.3

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

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

Solution

One Possible Solution:

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

Try to identify additional data sets in this example.

Sample of qualitative data

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

Exercise 2.3.4

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

Answer

qualitative data

COLLABORATIVE EXERCISE 2.3.1

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

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

Answer

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

Exercise 2.3.5

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

Answer

quantitative discrete

Exercise 2.3.6

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure **2.3.1**. What type of data does this graph show?

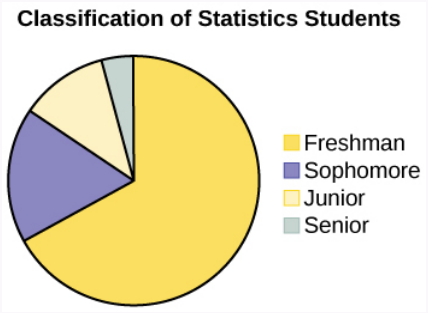


***Figure* 2.3.1**

Answer

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

Exercise 2.3.7

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

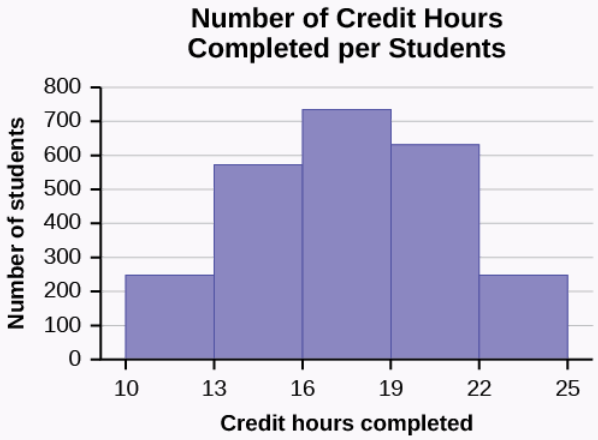


***Figure* 2.3.2**

What type of data does this graph show?

Answer

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

Qualitative Data Discussion

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

Table **2.3.1***: Fall Term 2007 (Census day)*

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

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

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

Look at Figures **2.3.3** and **2.3.4** and determine which graph (pie or bar) you think displays the comparisons better.

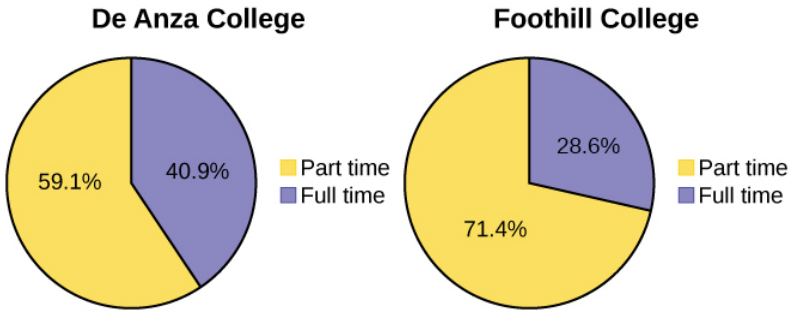


Figure* 2.3.3: Pie Charts*

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

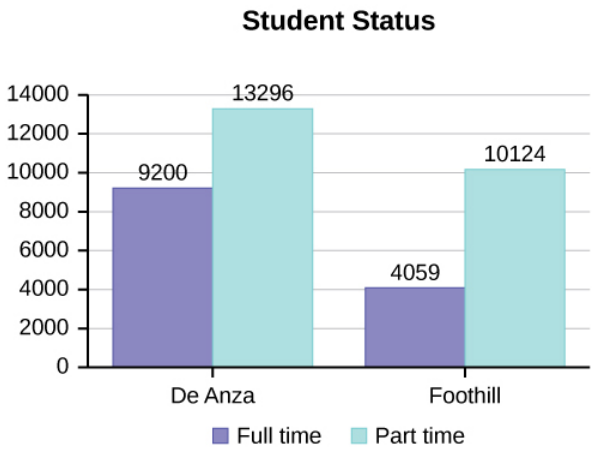


Figure* 2.3.4: Bar chart*

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

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

***Table* 2.3.2:** *De Anza College Spring 2010*

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

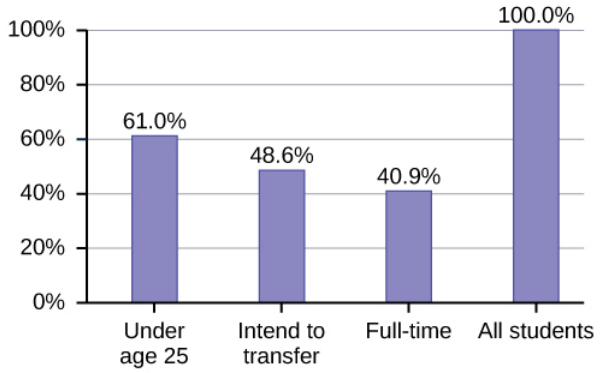


***Figure* 2.3.2**: *Bar chart of data in Table* **2.3.2**.

Omitting Categories/Missing Data

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

***Table* 2.3.2:** *Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)*

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

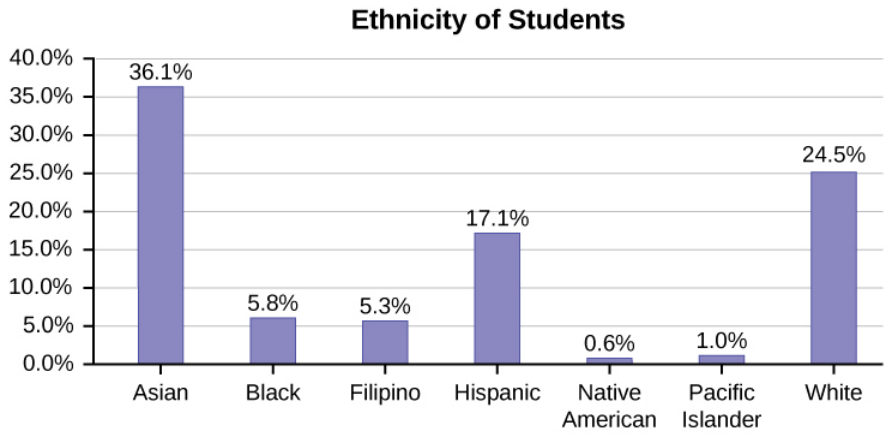


***Figure* 2.3.3:** *Enrollment of De Anza College (Spring 2010)*

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

This particular bar graph in Figure **2.3.4** can be difficult to understand visually. The graph in Figure **2.3.5** is a *Pareto chart*. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

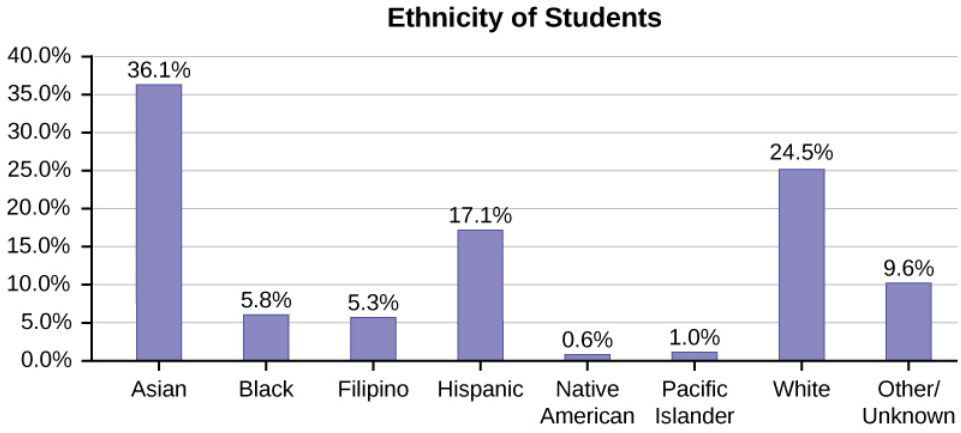


Figure **2.3.4***:* Bar Graph with Other/Unknown Category

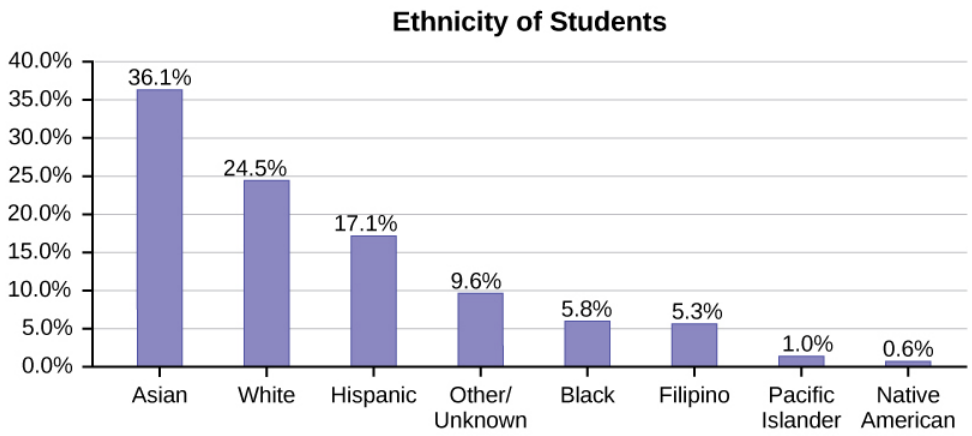


Figure* 2.3.5:* *Pareto Chart With Bars Sorted by Size*

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in Figure **2.3.6** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure **2.3.6**.

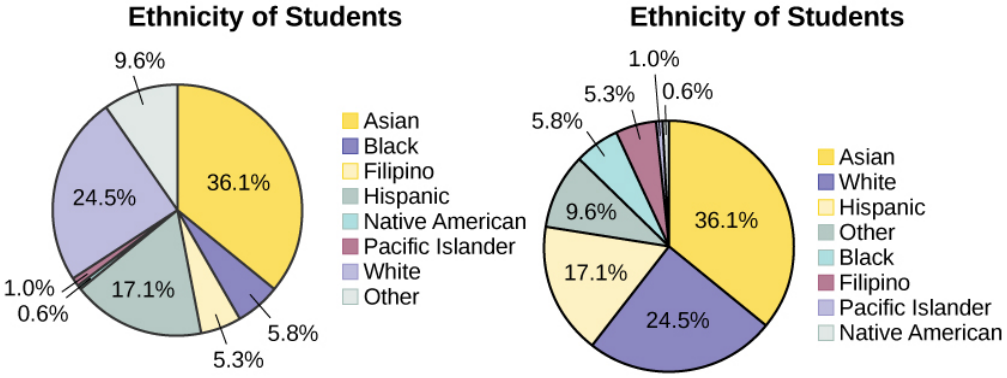


***Figure* 2.3.6.**

Sampling

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

Table* 2.3.3:* *Class Roster*

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

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

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

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

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

To generate random numbers:

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

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

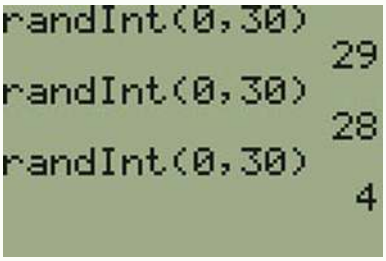


Figure 2.3.7

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

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

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

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

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

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

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

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

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

If you are sampling **without replacement**,

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

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

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

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

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

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

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

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

Exercise 2.3.8

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

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

Answer

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

Example 2.3.9: Calculator

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

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

Instructions: Use the Random Number Generator to pick samples.

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

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

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

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

Example 2.3.10

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

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.

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

Answer

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

Exercise 2.3.11

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

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

Answer

stratified

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

Example 2.3.12: Sampling

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

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

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

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

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

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

Answer

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

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

Answer

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

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator,

random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

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

c. Is the sample biased?

Answer

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

Exercise 2.3.12

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

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

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

Answer

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

COLLABORATIVE EXERCISE 2.3.8

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 2.3.8

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

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

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

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

Critical Evaluation

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

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.

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

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from http://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), http://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, http://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

Chapter 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: http://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, http://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

- Template:ContribOpenStax

This page titled 2.3: Data, Sampling, and Variation in Data and Sampling is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.4: Frequency, Frequency Tables, and Levels of Measurement

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

Answers and Rounding Off

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

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

Levels of Measurement

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

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

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

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

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

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

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

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

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

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

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

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

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* 2.4.1:** *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 Table **2.4.1**, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

Definition: Relative frequencies

A *relative frequency* is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

***Table* 2.4.2:** *Frequency Table of Student Work Hours with Relative Frequencies*

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

The sum of the values in the relative frequency column of Table **2.4.2** is $\frac{20}{20}$, or 1.

Definition: Cumulative relative frequency

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table **2.4.3**.

***Table* 2.4.3:** *Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies*

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	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.

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Table **2.4.4** represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

***Table* 2.4.4:** *Frequency Table of Soccer Player Height*

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

The data in this table have been **grouped** into the following intervals:

- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose

heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Exercise 2.4.1

From Table, find the percentage of heights that are less than 65.95 inches.

Answer

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Exercise 2.4.2

Table **2.4.5** shows the amount, in inches, of annual rainfall in a sample of towns.

***Table* 2.4.5:**

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97–6.99	7	$\frac{7}{50} = 0.14$	$0.12 + 0.14 = 0.26$
6.99–9.01	15	$\frac{15}{50} = 0.30$	$0.26 + 0.30 = 0.56$
9.01–11.03	8	$\frac{8}{50} = 0.16$	$0.56 + 0.16 = 0.72$
11.03–13.05	9	$\frac{9}{50} = 0.18$	$0.72 + 0.18 = 0.90$
13.05–15.07	5	$\frac{5}{50} = 0.10$	$0.90 + 0.10 = 1.00$
	Total = 50	Total = 1.00	

a. Find the percentage of rainfall that is less than 9.01 inches.
b. Find the percentage of heights that fall between 61.95 and 65.95 inches.
c. Find the percentage of rainfall that is between 6.99 and 13.05 inches.

Answer

a. 0.56 or 56
b. Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.
c. $0.30 + 0.16 + 0.18 = 0.64$ or 64

Exercise 2.4.3

Use the heights of the 100 male semiprofessional soccer players in Table **2.4.4**. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Answer

a. 29%
b. 36%
c. 77%
d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Exercise 2.4.4

From Table **2.4.5**, find the number of towns that have rainfall between 2.95 and 9.01 inches.

Answer

$6 + 7 + 15 = 28$ towns

COLLABORATIVE EXERCISE 2.4.7

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

a. What percentage of the students in your class have no siblings?
b. What percentage of the students have from one to three siblings?
c. What percentage of the students have fewer than three siblings?

Example 2.4.7

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table was produced:

***Table* 2.4.6:** Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?

d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Answer

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 2.4.8

Table represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Answer

$\frac{9}{50}$

Example 2.4.9

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

Table **2.4.7**:

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?

f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Answer

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

Exercise 2.4.10

Table contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Table **2.4.8**:

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Answer

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

1. “State & County QuickFacts,” U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

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

Chapter Review

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

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

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

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

Exercise 2.4.11

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

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

Answer

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

j. interval
k. ratio
l. ordinal

Glossary

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency

the number of times a value of the data occurs

Relative Frequency

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

Contributors

- Template:ContribOpenStax

This page titled 2.4: Frequency, Frequency Tables, and Levels of Measurement is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.5: Experimental Design and Ethics

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

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

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

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

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

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

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

Example 2.5.1

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

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

Answer

- The *population* is men aged 50 to 84.
- The *sample* is the 400 men who participated.

- The *experimental units* are the individual men in the study.
- The *explanatory variable* is oral medication.
- The *treatments* are aspirin and a placebo.
- The *response variable* is whether a subject had a heart attack.

Example 2.5.2

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

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

Answer

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

Example 2.5.3

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

Answer

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

Exercise 2.5.4

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

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

Answer

a. Explanatory: presence of distraction from texting; response: response time measured in seconds
b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?

d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.
f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Example 2.5.5

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

A researcher is collecting data in a community.

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

Answer

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

Exercise 2.5.6

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

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

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

Answer

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.

d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

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

Chapter Review

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

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

Exercise 2.5.7

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

Exercise 2.5.7

Discuss potential violations of the rule requiring informed consent.

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

Answer

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

Footnotes

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

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

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

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

Glossary

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

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

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

Experimental Unit
any individual or object to be measured

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

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

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

Informed Consent

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

Institutional Review Board

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

Placebo

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

Blinding

not telling participants which treatment a subject is receiving

Double-blinding

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

Contributors

- Template:ContribOpenStax

This page titled 2.5: Experimental Design and Ethics is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.6: Data Collection Experiment (Worksheet)

This page titled 2.6: Data Collection Experiment (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.7: Sampling Experiment (Worksheet)

This page titled 2.7: Sampling Experiment (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

2.8: Sampling and Data (Exercises)

This page titled 2.8: Sampling and Data (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

3: 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.

3.1: Prelude to Descriptive Statistics

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

3.3: Histograms, Frequency Polygons, and Time Series Graphs

3.4: Measures of the Location of the Data

Measures of the Location of the Data (Exercises)

3.5: Box Plots

3.6: Measures of the Center of the Data

3.7: Skewness and the Mean, Median, and Mode

3.8: Measures of the Spread of the Data

3.9: Descriptive Statistics (Worksheet)

3.10: Descriptive Statistics (Exercises)

Contributors

- Template:ContribOpenStax

This page titled 3: Descriptive Statistics is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.1: Prelude to Descriptive Statistics

Skills to Develop

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

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- 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.

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

Contributors

- Template:ContribOpenStax

This page titled 3.1: Prelude to Descriptive Statistics is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

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

One simple graph, the *stem-and-leaf graph* or *stemplot*, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 3.2.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Graph

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Exercise 3.2.2

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Answer

Stem	Leaf
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example 3.2.3

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

HINT: The leaves are to the right of the decimal.

Answer

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Exercise 3.2.4

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Answer

Stem	Leaf
0	5 7
1	1 2 2 3 3 5 5 7 7 8 9
2	0 2 5 6 8 8 8
3	5 8
4	4 8 9
5	2 5 7 8
6	
7	
8	0

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Example 3.2.5: side-by-side stem-and-leaf plot

A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Tables **3.2.1** and **3.2.2** show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

***Table* 3.2.1:** *Presidential Ages at Inauguration*

President	Age*at Inauguration*	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51	Trump	70

3.2.2 *Presidential Age at Death*

President	Age	President	Age	President	Age

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Answer

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Exercise 3.2.6

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Answer

3.2.1 Atlanta Hawks Wins and Losses

Number of Wins		Number of Losses
3	1	9
9 8 8 6 5	2	5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3	0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 8 7 6 6 6 3 3 3 2 2 1 1 0	4	0 0 1 1 2 4 5 6 6 7 7 8 9
7 7 6 3 2 0 0 0 0	5	1 1 1 2 3 4 4 6 7
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are

connected using line segments.

Example 3.2.7

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

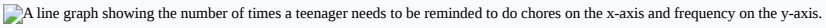


Figure 2.2.1.

Exercise 3.2.8

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Answer

Figure 2.2.2.

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

Example 3.2.9

By the end of 2011, Facebook had over 146 million users in the United States. Table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Answer

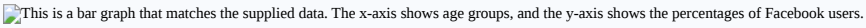


Figure 2.2.3.

Exercise 3.2.10

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

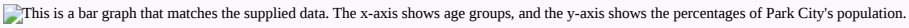


Figure 2.2.4.

Example 3.2.11

The columns in Table contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution



Figure 2.2.5

Exercise 3.2.12

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Answer

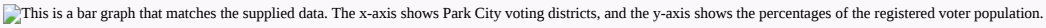

This is a bar graph that matches the supplied data. The x-axis shows Park City voting districts, and the y-axis shows the percentages of the registered voter population.

Figure 2.2.6

Summary

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

References

1. Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://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

- Template:ContribOpenStax

This page titled 3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.3: Histograms, Frequency Polygons, and Time Series Graphs

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

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

The 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:

$$RF = \frac{f}{n} \tag{3.3.1}$$

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 2.3.1

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; 67.5; 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.

$$\frac{74.05 - 59.95}{8} = 1.76 \quad (3.3.2)$$

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.

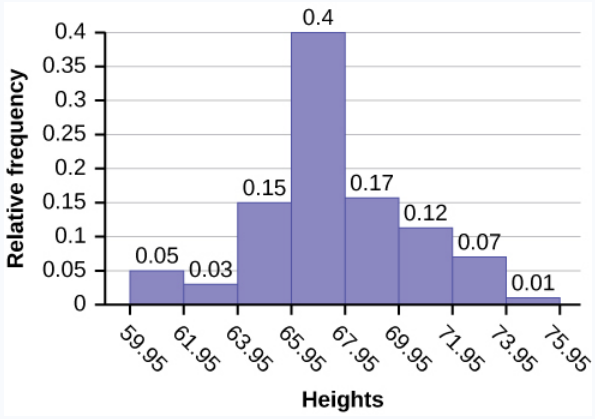


***Figure 2.3.1**: Histogram of something*

Exercise 2.3.1

continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5

12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Answer

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$\frac{14.05-8.95}{6} = 0.85$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example 2.3.2

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
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.

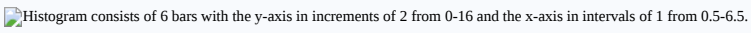


Figure 2.3.2.

Note

Go to [link]. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example.

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.

- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rdpicture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Exercise 2.3.2

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3

20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Answer

1.5

1.5 to 2.5

2.5 to 3.5

Example 2.3.3

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



This is a histogram that matches the supplied data. The x-axis consists of 5 bars in intervals of 5 from 0 to 25. The y-axis is marked in increments of 1 from 0 to 10. The x-axis shows the number of hours spent playing video games on the weekends, and the y-axis shows the number of students.

Figure 2.3.3.

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.

Exercise 2.3.3

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22; 35; 15; 26; 40; 28; 18; 20; 25; 34; 39; 42; 24; 22; 19; 27; 22; 34; 40; 20; 38 and 28

Use 10–19 as the first interval.

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

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 2.3.4

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 2.3.4.

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

Exercise 2.3.4

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

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

Answer

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

.


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

Figure 2.3.5.

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

Example 2.3.5

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

Frequency Distribution for Calculus Final Test Scores

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

Frequency Distribution for Calculus Final Grades

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


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

Figure 2.3.6.

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

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

Constructing a Time Series Graph

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

Example 2.3.6

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

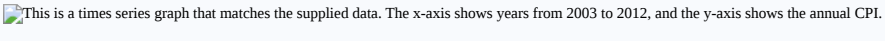


Figure 2.3.7.

Exercise 2.3.5

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

CO2 Emissions			
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

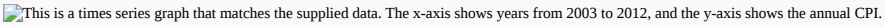


Figure 2.3.8.

Uses of a Time Series Graph

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

Chapter 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

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 http://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 http://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

- Template:ContribOpenStax

This page titled 3.3: Histograms, Frequency Polygons, and Time Series Graphs is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.4: Measures of the Location of the Data

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

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7 \tag{3.4.1}$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 – Q_1 \tag{2.4.1}$$

The *IQR* can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(*IQR*) below the first quartile or more than (1.5)(*IQR*) above the third quartile**. Potential outliers always require further investigation.

Definition: Outliers

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.4.1

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Answer

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Exercise 2.4.1

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

$33,000; $64,500; $28,000; $54,000; $72,000; $68,500; $69,000; $42,000; $54,000; $120,000; $40,500

Answer

Order the data from smallest to largest.

$28,000; $33,000; $40,500; $42,000; $54,000; $54,000; $64,500; $68,500; $69,000; $72,000; $120,000

Median = $54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

$$IQR = \$69,000 - \$40,500 = \$28,500$$

$$(1.5)(IQR) = (1.5)(\$28,500) = \$42,750$$

$$Q_1 - (1.5)(IQR) = \$40,500 - \$42,750 = -\$2,250$$

\[Q_{3} + (1.5)(*IQR*) = $69,000 + $42,750 = $111,750\nonumber\]

No salary is less than –$2,250. However, $120,000 is more than $11,750, so $120,000 is a potential outlier.

Example 2.4.2

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.

b. Any outliers in either set.

Answer

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The *IQR* for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

The *IQR* for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

- $Q_1 - IQR(1.5) = 56–26.5(1.5) = 16.25$
- $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

- $Q_1 - IQR(1.5) = 78–11(1.5) = 61.5$
- $Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Exercise 2.4.2

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class *B*

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Answer

Class *A*

Order the data from smallest to largest.

65; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99

$Median = \frac{80+81}{2}$ = 80.5

$Q_1 = \frac{69+76}{2} = 72.5$

$Q_3 = \frac{90+91}{2} = 90.5$

$IQR = 90.5 - 72.5 = 18$

Class *B*

Order the data from smallest to largest.

68; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 99; 100

$Median = \frac{80+80}{2} = 80$

$Q_1 = \frac{72+73}{2} = 72.5$

$Q_3 = \frac{92+95}{2} = 93.5$

$IQR = 93.5 - 72.5 = 21$

The data for Class *B* has a larger *IQR*, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class *B* are more spread out and not clustered about the median.

Example 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 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Exercise 2.4.3

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Answer

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example 2.4.4

Using Table:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

Exercise 2.4.4

Refer to the Table. Find the third quartile. What is another name for the third quartile?

Answer

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

COLLABORATIVE STATISTICS

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

a. How many students were surveyed?
b. What kind of sampling did you do?
c. Construct two different histograms. For each, starting value = _____ ending value = ____.
d. Find the median, first quartile, and third quartile.
e. Construct a table of the data to find the following:
   i. the 10th percentile
   ii. the 70th percentile
   iii. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the kth percentile. Here is one of them.

- $k =$ the kth percentile. It may or may not be part of the data.
- $i =$ the index (ranking or position of a data value)
- $n =$ the total number of data

Order the data from smallest to largest.

Calculate $i = \frac{k}{100}(n+1)$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) = \dfrac{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.

Exercise 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

Calculate the 20th percentile and the 55th percentile.

Answer

$k = 20$. Index $= i = \frac{k}{100}(n+1) = \frac{20}{100}(29+1) = 6$. The age in the sixth position is 27. The 20th percentile is 27 years.

$k = 55$. Index $= i = \frac{k}{100}(n+1) = \frac{55}{100}(29+1) = 16.5$. Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55. The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

Note 2.4.2

You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.

- $x =$ the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- $y =$ the number of data values equal to the data value for which you want to find the percentile.
- $n =$ the total number of data.
- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.4.6

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

$x = 18$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{18 + 0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

$x = 3$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{3 + 0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12thpercentile.

Exercise 2.4.6

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Answer

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$x = 15$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(1)}{29}(100) = 53.45$. 47 is the 53rd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$x = 15$ and $y = 2$. $\frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(2)}{29}(100) = 31.03$. 31 is the 31st percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example 2.4.7

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Answer

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Exercise 2.4.7

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Answer

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

Example 2.4.8

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.

Exercise 2.4.8

On a 60 point written assignment, the 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.

Answer

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example 2.4.9

At a community college, it was found that the 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.

Exercise 2.4.9

During a season, the 40^{th} percentile for points scored per player in a game is eight. Interpret the 40^{th} percentile in the context of this situation.

Answer

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Example 2.4.10

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \quad (3.4.2)$$

.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

1. Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at http://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.

Chapter Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$$i = \frac{k}{100}(n+1) \tag{3.4.3}$$

where i = the ranking or position of a data value,

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value: Missing or unrecognized delimiter for \right(100)\)

where $x =$ the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

$y =$ the number of data values equal to the data value for which you want to find the percentile,

$n =$ total number of data

Glossary

Interquartile Range

or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier

an observation that does not fit the rest of the data

Percentile

a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Contributors

- Template:ContribOpenStax

This page titled 3.4: Measures of the Location of the Data is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Measures of the Location of the Data (Exercises)

Exercise 2.4.10

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40th percentile.
b. Find the 78th percentile.

Answer

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

Exercise 2.4.11

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile of 37.
b. Find the percentile of 72.

Exercise 2.4.12

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Answer

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and \(y = 1\). $\frac{x + 0.5y}{n}(100) = \frac{143 + 0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercise 2.4.13

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 2.4.14

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Answer

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise 2.4.15

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise 2.4.16

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Answer

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise 2.4.17

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise 2.4.18

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Answer

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise 2.4.19

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise 2.4.20

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Answer

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Use Exercise to calculate the following values:

Exercise 2.4.21

First quartile = _____

Exercise 2.4.22

Second quartile = median = 50th percentile = _____

Answer

4

Exercise 2.4.23

Third quartile = _____

Exercise 2.4.24

Interquartile range (*IQR*) = _____ – _____ = _____

Answer

$6 - 4 = 2$

Exercise 2.4.25

10^{th} percentile = _____

Exercise 2.4.26

70^{th} percentile = _____

Answer

6

Measures of the Location of the Data (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

3.5: 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.

See the calculator instructions on the TI web site or in the appendix.

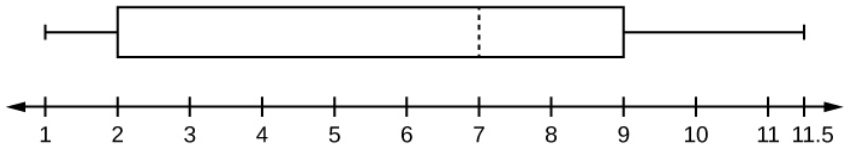


***Figure* 3.5.1**

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example 3.5.1

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 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
- *Q*1: First quartile = 64.5
- *Q*2: Second quartile or median= 66
- *Q*3: Third quartile = 70

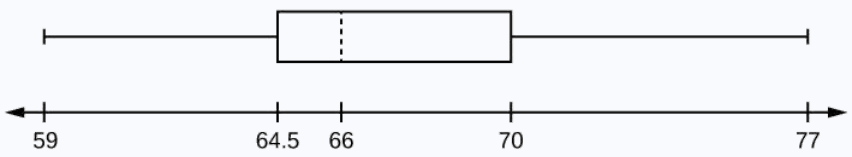


Figure 3.5.2

a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. $\text{Range} = \text{maximum value} - \text{the minimum value} = 77 - 59 = 18$
d. Interquartile Range: $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$.
e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Calculator

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4: Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE, and use the arrow keys to examine the box plot.

Exercise 3.5.1

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

Answer

Figure 3.5.3

$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



***Figure* 3.5.4**

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example 3.5.2

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5
- Q_3 = 82.5
- Max = 99

b.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

d.

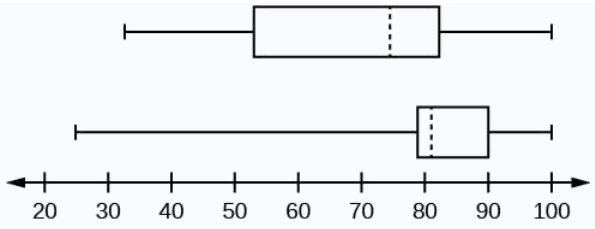


***Figure* 3.5.5**

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.

Exercise 3.5.2

The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Answer

***Figure* 3.5.6**

IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

Example 3.5.3

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.

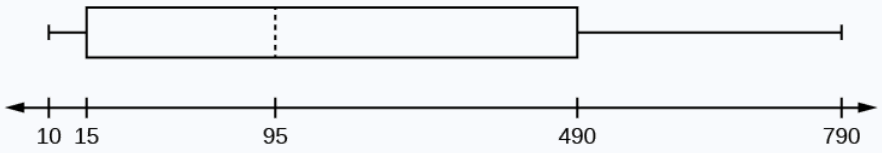


***Figure* 3.5.7**

Exercise 3.5.3

Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

Answer

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330

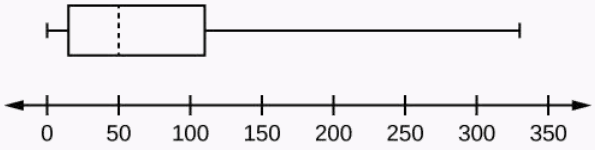


***Figure* 3.5.8**

References

1. Data from *West Magazine*.

Chapter Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise 2.5.4

Construct a box plot below. Use a ruler to measure and scale accurately.

Exercise 2.5.5

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Answer

More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Bringing It Together

Exercise 2.5.6

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?

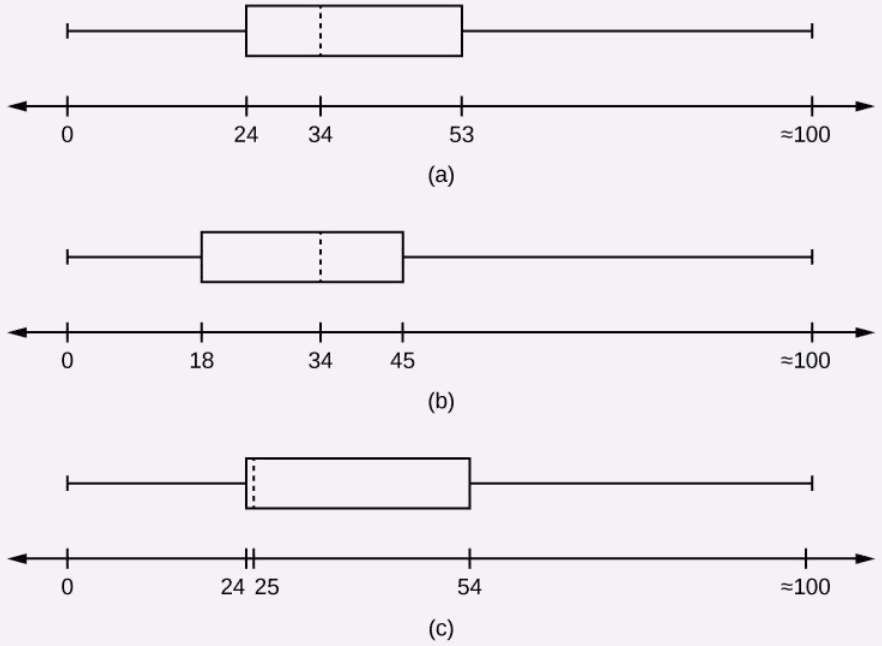


Figure 2.5.9.

Answer

a. For graph, check student's solution.
b. 49.7% of the community is under the age of 35.
c. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Contributors

- Template:ContribOpenStax

This page titled 3.5: Box Plots is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.6: Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the mean (average) and the median. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

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

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{3.6.1}$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \tag{3.6.2}$$

In the second example, the frequencies are

$$3(1)+2(2)+1(3)+5(4). \tag{3.6.3}$$

You can quickly find the location of the median by using the expression

$$\frac{n+1}{2} \tag{3.6.4}$$

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \tag{3.6.5}$$

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

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \tag{3.6.6}$$

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 3.6.1

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:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6 \quad (3.6.7)$$

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

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \quad (3.6.8)$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

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

$$M = \frac{24+24}{2} = 24 \quad (3.6.9)$$

Calculator

To find the mean and the median:

Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$\bar{x}$ = 23.6, M = 24

Exercise 3.6.1

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Answer

Mean: $3+4+5+7+7+7+7+8+8+9+9+10+10+10+10+10+11+12+12+13+14+14+15+15+17+17+18+19+19+19+21+21+22+22+23+24+24+24=544$

$$\frac{544}{39} = 13.95 \quad (3.6.10)$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example 3.6.2

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

$$\bar{x} = \frac{5,000,000+49(30,000)}{50} = 129,400 \quad (3.6.11)$$

$M = 30,000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Exercise 3.6.2

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Answer

The median is the better measure of the “center” than the mean because 59 of the values are $280,000 and one is $2,500,000. The $2,500,000 is an outlier. Either $280,000 or $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 3.6.3

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Answer

The most frequent score is 72, which occurs five times. Mode = 72.

Exercise 3.6.3

The number of books checked out from the library from 25 students are as follows:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 7; 8; 8; 8; 9; 10; 10; 11; 11; 12; 12

Find the mode.

Answer

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example 3.6.4

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Exercise 3.6.4

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Answer

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don't represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{3.6.12}$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary+upper boundary}}{2}. \tag{3.6.13}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{3.6.14}$$

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

Example 3.6.5

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm 53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Exercise 3.6.5

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Answer

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.

$$Mean = (1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9) = 409.75 \quad (3.6.15)$$

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).

Chapter 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

$$\mu = \frac{\sum fm}{\sum f} \quad (3.6.16)$$

where f = interval frequencies and m = interval midpoints.

Exercise 2.6.6

Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

Exercise 2.6.7

Calculate the mean.

Answer

Mean: $16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738$;

$\frac{738}{27} = 27.33$

Exercise 2.6.8

Identify the median.

Exercise 2.6.9

Identify the mode.

Answer

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise 2.6.10

sample mean = $\bar{x}$ = _____

Exercise 2.6.11

median = _____

Answer

4

Bringing It Together

Exercise 2.6.12

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 miles	6.0 miles
ss	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know? <figure id="eip-idm218452576">

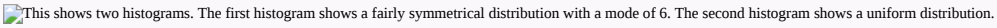


Figure 2.6.1

</figure>
d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know? <figure id="eip-idm179364784">

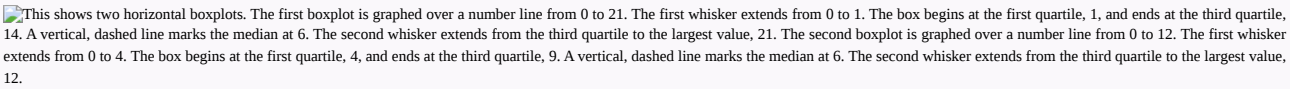


Figure 2.6.2

</figure>

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise 2.6.13

What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Answer

a

Exercise 2.6.14

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise 2.6.15

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Answer

b

Glossary

Frequency Table
a data representation in which grouped data is displayed along with the corresponding frequencies

Mean
a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median
a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint
the mean of an interval in a frequency table

Mode
the value that appears most frequently in a set of data

Contributors

- Template:ContribOpenStax

This page titled 3.6: Measures of the Center of the Data is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.7: Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

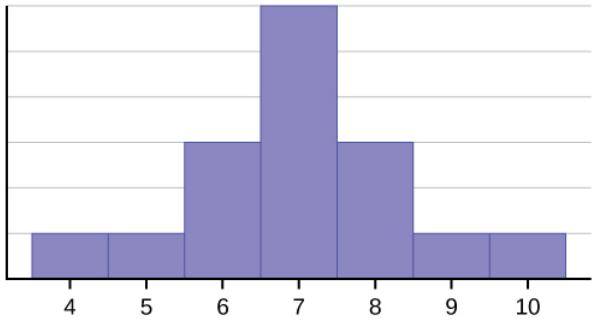


***Figure* 3.7.1**

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

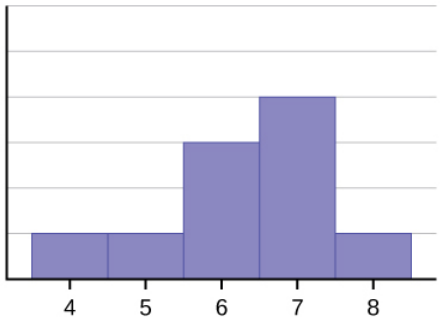


***Figure* 3.7.2**

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.

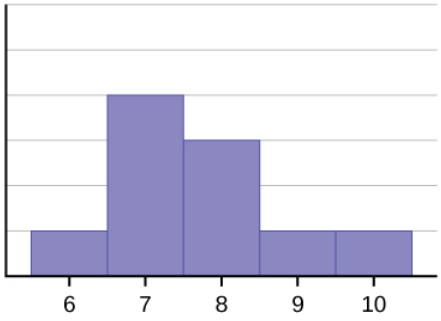


Figure 3.7.3

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

> *Generally, if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.*

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example 3.7.1

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

- Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2
- Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1
- Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

Solution

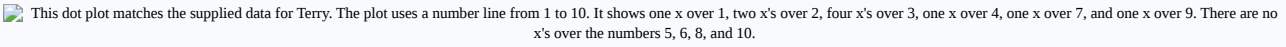
a. 

Figure 3.7.4: Terry's distribution has a right (positive) skew.

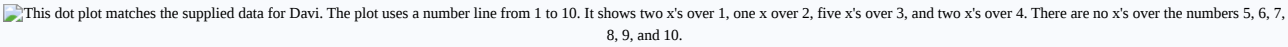


Figure 3.7.5: Davis' distribution has a left (negative) skew

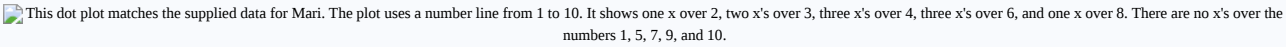


Figure 3.7.6: Maris' distribution is symmetrically shaped.

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.

Exercise 3.7.1

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

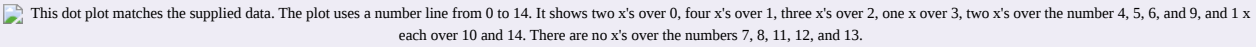


Figure 3.7.7.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.

***Figure* 3.7.1**

Chapter 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 **right (or positive) skewed** distribution has a shape like Figure **3.7.3**.
- A **left (or negative) skewed** distribution has a shape like Figure **3.7.2**.
- A **symmetrical** distribution looks like Figure **3.7.1**.

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise 2.7.2

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

Answer

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise 2.7.3

16; 17; 19; 22; 22; 22; 22; 22; 23

Exercise 2.7.4

87; 87; 87; 87; 87; 88; 89; 89; 90; 91

Answer

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise 2.7.5

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise 2.7.6

When the data are symmetrical, what is the typical relationship between the mean and median?

Answer

When the data are symmetrical, the mean and median are close or the same.

Exercise 2.7.7

What word describes a distribution that has two modes?

Exercise 2.7.8

Describe the shape of this distribution.


This is a historgram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right.

Figure 2.7.9

Answer

The distribution is skewed right because it looks pulled out to the right.

Exercise 2.7.9

Describe the relationship between the mode and the median of this distribution.

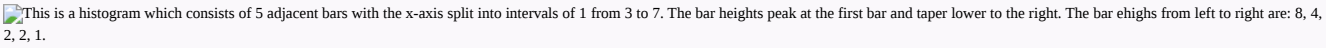

This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar ehighs from left to right are: 8, 4, 2, 2, 1.

Figure 2.7.10

Exercise 2.7.10

Describe the relationship between the mean and the median of this distribution.

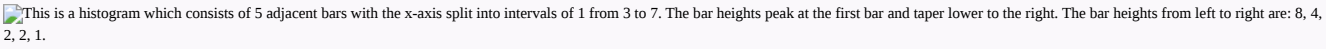

This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Figure 2.7.11

Answer

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise 2.7.11

Describe the shape of this distribution.

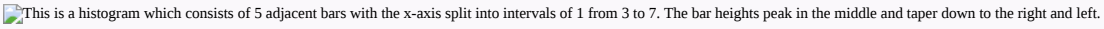

This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak in the middle and taper down to the right and left.

Figure 2.7.12

Exercise 2.7.12

Describe the relationship between the mode and the median of this distribution.

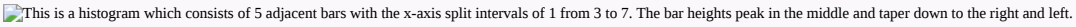

This is a histogram which consists of 5 adjacent bars with the x-axis split intervals of 1 from 3 to 7. The bar heights peak in the middle and taper down to the right and left.

Figure 2.7.13

Answer

The mode and the median are the same. In this case, they are both five.

Exercise 2.7.13

Are the mean and the median the exact same in this distribution? Why or why not?

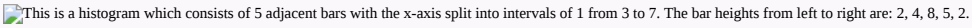

This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights from left to right are: 2, 4, 8, 5, 2.

Figure 2.7.14

Exercise 2.7.14

Describe the shape of this distribution.

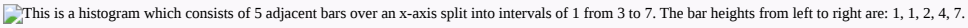


Figure 2.7.15

Answer

The distribution is skewed left because it looks pulled out to the left.

Exercise 2.7.15

Describe the relationship between the mode and the median of this distribution.

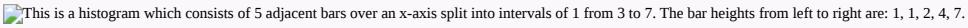


Figure 2.7.16

Exercise 2.7.16

Describe the relationship between the mean and the median of this distribution.

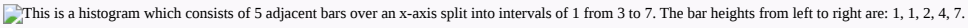


Figure 2.7.17

Answer

The mean and the median are both six.

Exercise 2.7.17

The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

Exercise 2.7.18

Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

Answer

The mode is 12, the median is 13.5, and the mean is 15.1. The mean is the largest.

Exercise 2.7.19

Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

Exercise 2.7.20

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Answer

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise 2.7.21

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Contributors

- Template:ContribOpenStax

This page titled 3.7: Skewness and the Mean, Median, and Mode is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.8: Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.



***Figure* 3.8.1**

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**

- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation value = mean + (#ofSTDEVs)(standard deviation) can be expressed for a sample and for a population.

- sample:

Misplaced '#'

- Population:

Misplaced '#'

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

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

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{3.8.1}$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \tag{3.8.2}$$

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

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{3.8.3}$$

or

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}} \quad (3.8.4)$$

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

In Equations 3.8.2 and 3.8.4, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Section 2.6. How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the **sampling variability of a statistic** by its standard error.

The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in Chapter 7. The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. 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 3.8.1

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;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525 \quad (3.8.5)$$

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):

$$s^2 = \frac{9.7375}{20-1} = 0.5125 \quad (3.8.6)$$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891 \quad (3.8.7)$$

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

Exercise 2.8.1

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Answer

$\mu = 30.68$

$s = 6.09$

$(\bar{x} + 2s = 30.68 + (2)(6.09) = 42.86.$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For Example 3.8.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population variance.

> Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example 3.8.2

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.

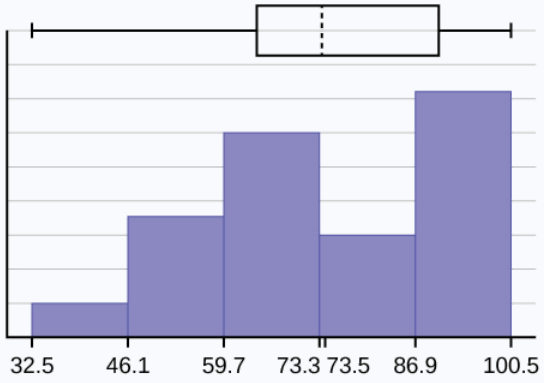


Figure 2.9.2.

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.

https://socialsci.libretexts.org/@go/page/19032

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Exercise 3.8.2

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Answer

$\mu = 9.3$ and $s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{3.8.8}$$

where f interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example 3.8.3

Find the standard deviation for the data in Table **3.8.3**.

Table **3.8.3**

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.

Find the standard deviation for the data from the previous example

Class	0-2	3-5	6-8	9–11	12–14	15–17
Frequency, f	1	6	10	7	0	2

First, press the **STAT** key and select **1:Edit**

Figure 2.8.3

Input the midpoint values into **L1** and the frequencies into **L2**

Figure 2.8.4

Select **STAT**, **CALC**, and **1: 1-Var Stats**

Figure 2.8.5

Select **2nd** then **1** then , **2nd** then **2 Enter**

Figure 2.8.6

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- Misplaced '#'
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol \(z\). In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example 3.8.4

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

Misplaced '#'

For John,

z = Misplaced '#'

For Ali,

Misplaced '#'

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Exercise 3.8.4

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Answer

For Angie:

$$z = \left(\frac{26.2 - 27.2}{0.8}\right) = -1.25 \tag{3.8.9}$$

For Beth:

$$z = \left(\frac{27.3 - 30.1}{1.4}\right) = -2 \tag{3.8.10}$$

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 http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Chapter Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$. .

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{3.8.11}$$

where s_x **sample standard deviation** and $\bar{x}$ **= sample mean**

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Exercise 2.8.4

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Answer

s = 34.5

Exercise 2.8.5

Find the value that is one standard deviation below the mean.

Exercise 2.8.6

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Answer

For Fredo:

$$z = \frac{0.158 - 0.166}{0.012} = -0.67$$

For Karl:

$$z = \frac{0.177 - 0.189}{0.015} = -0.8$$

Fredo's z-score of –0.67 is higher than Karl's z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise 2.8.7

Use Table to find the value that is three standard deviations:

- above the mean
- below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise 2.8.5

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Answer

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Bringing It Together

Exercise 2.8.7

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

a. Find the sample mean $\bar{x}$.

b. Find the approximate sample standard deviation, s.

Answer

a. 1.48

b. 1.12

Exercise 2.8.8

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40^{th} percentile.
k. Find the 90^{th} percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise 2.8.9

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _____ to _____.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, σ.
   iii. the weight that is two standard deviations below the mean.

iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

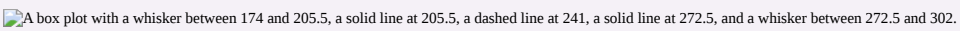
Answer

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302

b. 241

c. 205.5

d. 272.5

e. 

f. 205.5, 272.5

g. sample

h. population

i. i. 236.34
ii. 37.50
iii. 161.34
iv. 0.84 std. dev. below the mean

j. Young

Exercise 2.8.10

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

a. What is the mean change score?

b. What is the standard deviation for this population?

c. What is the median change score?

d. Find the change score that is 2.2 standard deviations below the mean.

Exercise 2.8.11

Refer to Figure determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

<figure id="fs-idm70725344">

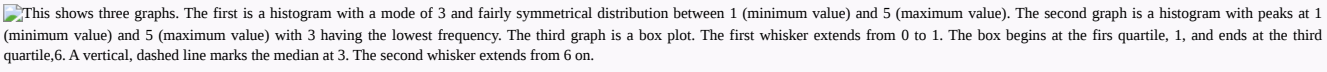


Figure 2.8.6.

</figure>

a. The medians for all three graphs are the same.

b. We cannot determine if any of the means for the three graphs is different.

c. The standard deviation for graph b is larger than the standard deviation for graph a.

d. We cannot determine if any of the third quartiles for the three graphs is different.

Answer

a. True

b. True
c. True
d. False

Exercise 2.8.12

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _____ days to _____ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise 2.8.13

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Answer

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. $X =$ the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise 2.8.14

The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise 2.8.15

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Answer

a

Exercise 2.8.16

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the Table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR*to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Contributors

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.

Template:ContribOpenStax

This page titled 3.8: Measures of the Spread of the Data is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.9: Descriptive Statistics (Worksheet)

This page titled 3.9: Descriptive Statistics (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

3.10: Descriptive Statistics (Exercises)

This page titled 3.10: Descriptive Statistics (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

4: Graphing Distributions

Template:TopHeblLane

Graphing data is the first and often most important step in data analysis. In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves. This is all the more unfortunate because computers can create many types of graphs quickly and easily. This chapter covers some classic types of graphs such as bar charts that were invented by William Playfair in the 18th century as well as graphs such as box plots invented by John Tukey in the 20th century.

Topic hierarchy

4.1: Graphing Qualitative Variables
4.2: Quantitative Variables
4.3: Stem and Leaf Displays
4.4: Histograms
4.5: Frequency Polygons
4.6: Box Plots
4.7: Box Plot Demo
4.8: Bar Charts
4.9: Line Graphs
4.10: Dot Plots
4.11: Statistical Literacy
4.12: Graphing Distributions (Exercises)

Contributors

- Template:ContribHeblLane

This page titled 4: Graphing Distributions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.1: Graphing Qualitative Variables

Skills to Develop

- Create a frequency table
- Determine when pie charts are valuable and when they are not
- Create and interpret bar charts
- Identify common graphical mistakes

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 4.1.1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" is $85/500 = 0.17$.

***Table* 4.1.1:** *Frequency Table for the iMac Data*

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Pie Charts

The pie chart in Figure 4.1.1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

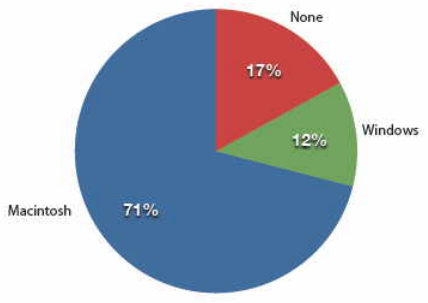


Figure* 4.1.1*: *Pie chart of iMac purchases illustrating frequencies of previous computer ownership*

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted, "The only worse design than a pie chart is several of them."

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure $4.1.2$. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations as is typical in pie charts.

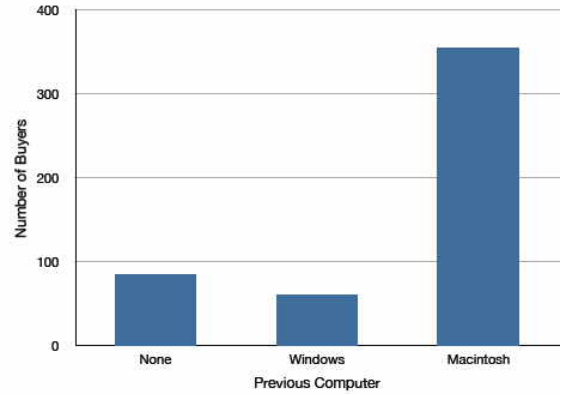


Figure* 4.1.2*: *Bar chart of iMac purchases as a function of previous computer ownership*

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure $4.1.3$ shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

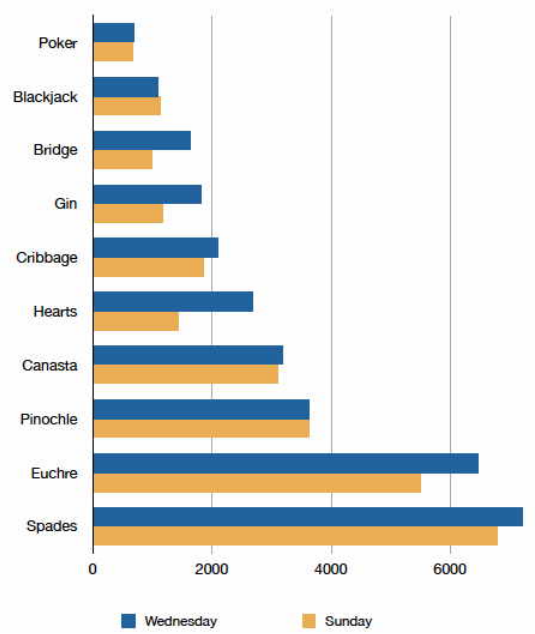


Figure* 4.1.3*: *A bar chart of the number of people playing different card games on Sunday and Wednesday*

The bars in Figure **4.1.3** are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in the section Bar Charts.

Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, **3**-dimensional bar charts such as the one shown in Figure **4.1.4** are usually not as effective as their two-dimensional counterparts.

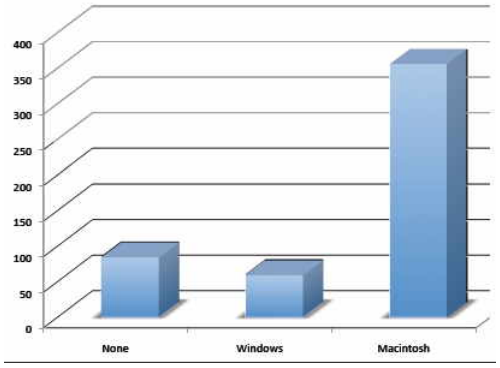


Figure* 4.1.4*: *A three-dimensional version of Figure* **4.1.2**

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure **4.1.5** presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure **4.1.5** is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about **6** to **1**. But the ratio of the two areas in Figure **4.1.5** is about **35** to **1**. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure **4.1.5** instead of Figure **4.1.2**! Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than **1.05** or less than **0.95** produce unacceptable distortion.

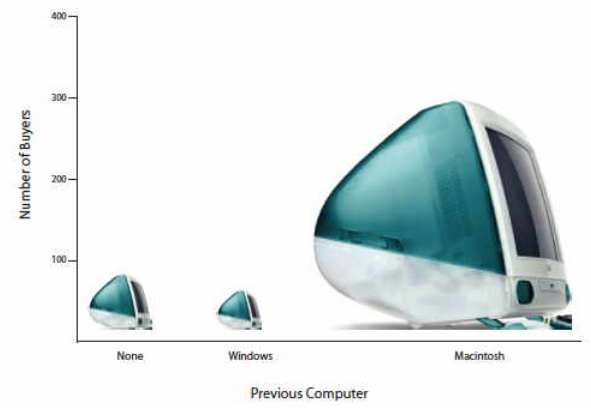


Figure* 4.1.5*: *A redrawing of Figure* **4.1.2** *with a lie factor greater than* **8**

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure **4.1.6** shows the iMac data with a baseline of **50**. Once again, the differences in areas suggest a different story than the true differences in percentages. The percentage of Windows-switchers seems minuscule compared to its true value of **12%**.

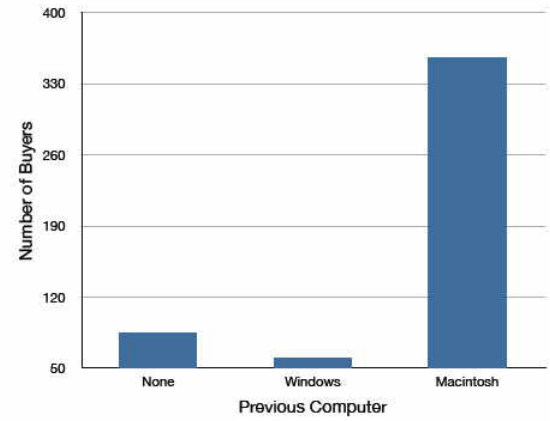


Figure* 4.1.6*: *A redrawing of Figure* **4.1.2** *with a baseline of* **50**

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure **4.1.7** inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure **4.1.7** is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

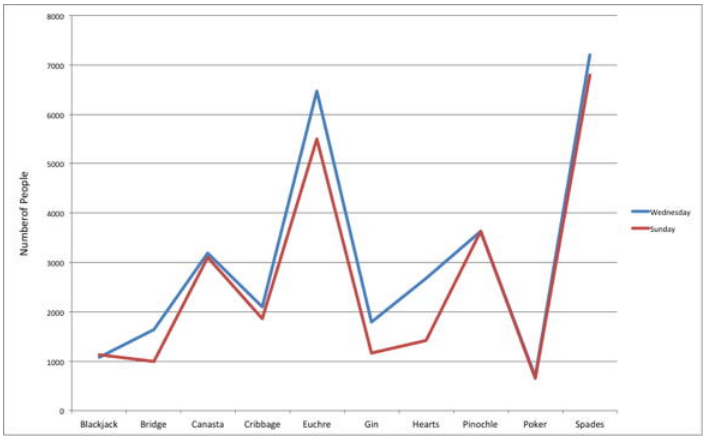


Figure* 4.1.7*: *A line graph used inappropriately to depict the number of people playing different card games on Sunday and*

Wednesday.

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

Contributors

- Template:ContribHeblLane

This page titled 4.1: Graphing Qualitative Variables is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.2: Quantitative Variables

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, and favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs:

1. stem and leaf displays
2. histograms
3. frequency polygons
4. box plots
5. bar charts
6. line graphs
7. scatter plots (discussed in a different chapter)
8. dot plots

Some graph types such as stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best-suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

Contributors

- Template:ContribHeblLane

This page titled 4.2: Quantitative Variables is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.3: Stem and Leaf Displays

Skills to Develop

- Create and interpret basic stem and leaf displays
- Create and interpret back-to-back stem and leaf displays
- Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, an example will get us started. Consider Table 4.3.1 that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

Table* 4.3.1*: *Number of touchdown passes*

37	33	33	32	29	28	28	23	22
22	22	21	21	21	20	20	19	19
18	18	18	18	16	15	14	14	14
12	12	9	6					

(4.3.1)

A stem and leaf display of the data is shown in Figure 4.3.1. The left portion of Figure 4.3.1 contains the stems. They are the numbers $3, 2, 1,$ *and* 0, arranged as a column to the left of the bars. Think of these numbers as $10's$ digits. A stem of 3, for example, can be used to represent the $10's$ digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the $1's$ digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

3	2	3	3	7								
2	0	0	1	1	1	2	2	2	3	8	8	9
1	2	2	4	4	4	5	6	8	8	8	8	9	9
0	6	9										

(4.3.2)

Figure* 4.3.1*: *Stem and leaf display of the number of touchdown passes*

To make this clear, let us examine Figure 4.3.1 more closely. In the top row, the four leaves to the right of stem 3 are $2, 3, 3,$ *and* 7. Combined with the stem, these leaves represent the numbers $32, 33, 33,$ *and* 37, which are the numbers of TD passes for the first four teams in Table 4.3.1. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table 4.3.1, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 4.3.1 than in Table 4.3.1. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure 4.3.2 shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table 4.3.1. The second row is reserved for the numbers from 30 to 34 and holds the $32, 33,$ *and* 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

3	7									
3	2	3	3							
2	8	8	9							
2	0	0	1	1	1	2	2	2	3	
1	5	6	8	8	8	8	9	9		
1	2	2	4	4	4					
0	6	9								

(4.3.3)

Figure* 4.3.2*: *Stem and leaf display with the stems split in two*

Figure 4.3.2 is more revealing than Figure $1 PageIndex 2$ because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a "back-to-back stem and leaf graph." Figure 4.3.3 shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with $11, 12,$ *and* 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

1 1	4	
	3	7
3 3 2	3	2 3 3
8 8 6 5	2	8 8 9
4 4 3 3 1 1 1 0	2	0 0 1 1 1 2 2 2 3
9 8 7 7 7 6 6 6 5	1	5 6 8 8 8 8 9 9
3 2 1	1	2 2 4 4 4 4
7	0	6 9

(4.3.4)

Figure* 4.3.3*: *Back-to-back stem and leaf display.*

The left side shows the 1998 TD data and the right side shows the 2000 TD data. Figure 4.3.3 helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table 4.3.2 shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like “punch”) under two conditions. In one condition, the words were preceded by a non-weapon word such as "bug." In the second condition, the same words were preceded by a weapon word such as "gun" or "knife." The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

Table* 4.3.2*: *The effects of priming (thousandths of a second)*

43.2	42.9	35.6	25.6	25.4	23.6		
20.5	19.9	14.4	12.7	11.3	10.2		
10.0	9.1	7.5	5.4	4.7	3.8	2.1	1.2
−0.2	−6.3	−6.7	−8.8	−10.4	−10.5		
−14.9	−14.9	−15.0	−18.5	−27.4			

(4.3.5)

You see that the numbers range from 43.2 to −27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value −27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure 4.3.4. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf), the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number −27. The second-to-last row represents the numbers −10, −10, −15, etc. Once again, we have rounded the original values from Table 4.3.2.

4	3 3
3	6
2	0 0 4 5 6
1	0 0 1 3 4
0	1 2 4 5 5 8 9
−0	0 6 7 9
−1	0 0 5 5 5 9
−2	7

(4.3.6)

Figure* 4.3.4*: *Stem and leaf display with negative numbers and rounding*

Observe that the figure contains a row headed by "0" and another headed by "−0". The stem of 0 is for numbers between 0 and 9, whereas the stem of −0 is for numbers between 0 and −9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, −6, −7, *and* −9. Values that are exactly 0 before rounding should be split as evenly as possible between the "0" and "−0" rows. In Table 4.3.2, none of the values are 0 before rounding. The "0" that appears in the "−0" row comes from the original value of −0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure 4.3.5 portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.

```
4 | 8 9 9
4 | 6
4 | 4 4 5 5
4 | 3 3 3
4 | 0 1
3 | 9 9
3 | 6 7 7 7 7 7
3 | 5 5
3 | 2 2 3
3 | 1 1 1
2 | 8 8 9 9
2 | 6 6 6 6 6 7
2 | 4 4 4 4 5 5
2 | 2 2 3 3 3
2 | 0 0 0 0 0 0
1 | 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9
1 | 6 6 6 6 6 6 7 7 7 7 7 7
1 | 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5
1 | 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3
1 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

(4.3.7)

Figure* 4.3.5*: *Stem and leaf display of populations of* 185 *US cities with populations between* 100,000 *and* 500,000 *in* 1998.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example, the largest number (493,559) was rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: $0-1, 2-3, 4-5, 6-7$, and $8-9$.

Whether your data can be suitably represented by a stem and leaf graph depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure 4.3.5 fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

Contributors

- Template:ContribHeblLane

This page titled 4.3: Stem and Leaf Displays is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.4: Histograms

Skills to Develop

- Create a grouped frequency distribution
- Create a histogram based on a grouped frequency distribution
- Determine an appropriate bin width

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psycho\logy test. The test consists of 197 items, each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 4.4.1.

***Table* 4.4.1:** *Grouped Frequency Distribution of Psycho\logy Test Scores*

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 4.4.1.

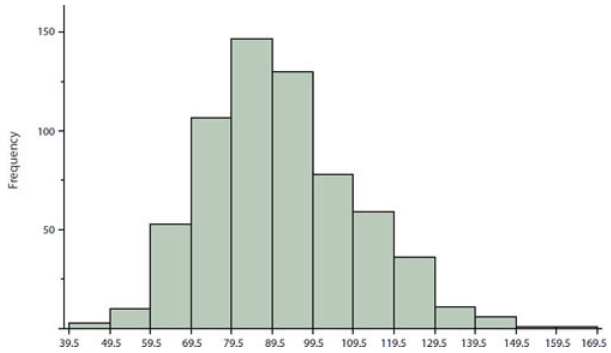


***Figure* 4.4.1:** *Histogram of scores on a psycho\logy test.*

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in the chapter " Summarizing Distributions.")

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence-sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

Sturges' rule

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some "rules of thumb" that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) Sturges' rule is to set the number of intervals as close as possible to $1 + \log_2(N)$, where $\log_2(N)$ is the base 2 log of the number of observations. The formula can also be written as $1 + 3.3 \log_2(N)$, where $\log_{10}(N)$ is the log base 10 of the number of observations. According to Sturges' rule, 1000 observations would be graphed with 11 class intervals since 10 is the closest integer to $\log_2(1000)$. We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by Sturges' rule. For the psychology test example used above, Sturges' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure $4.4.1$ to create a histogram that seemed clearest. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

To provide experience in constructing histograms, we have developed an interactive demonstration. The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Contributors

- Template:ContribHeblLane

Interactive histogram

This page titled 4.4: Histograms is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.5: Frequency Polygons

Skills to Develop

- Create and interpret frequency polygons
- Create and interpret cumulative frequency polygons
- Create and interpret overlaid frequency polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

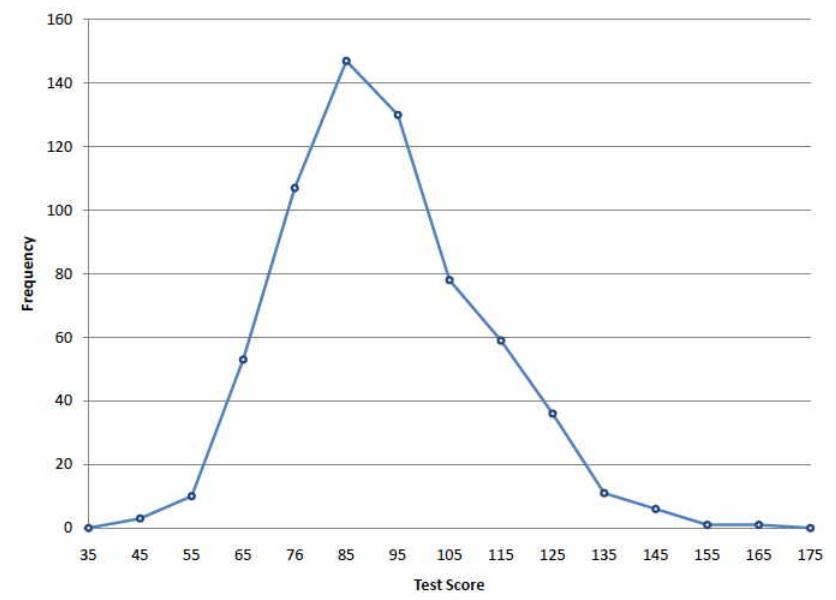


Figure* 4.5.1*: *Frequency polygon for the psychology test scores*

A frequency polygon for 642 psychology test scores shown in Figure 4.5.1 was constructed from the frequency table shown in Table 4.5.1.

Table* 4.5.1*: *Frequency Distribution of Psychology Test Scores.*

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	179.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 4.5.1. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

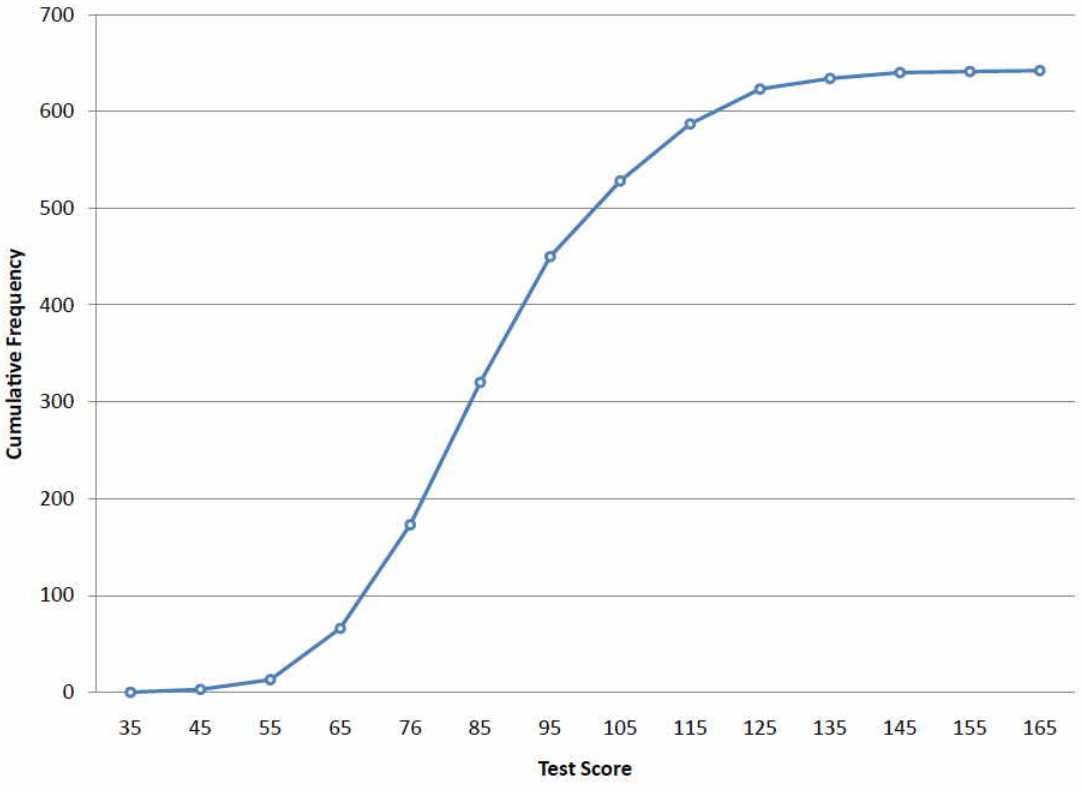


***Figure* 4.5.2:** *Cumulative frequency polygon for the psychology test scores*

A cumulative frequency polygon for the same test scores is shown in Figure 4.5.2. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled 35, three in the interval 45, and 10 in the interval 55. Therefore, the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

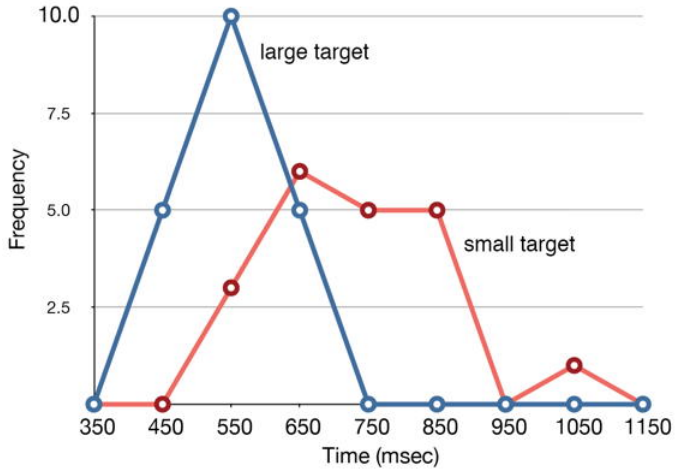


***Figure* 4.5.3:** *Overlaid frequency polygons*

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 4.5.3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in

Figure 4.5.3. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

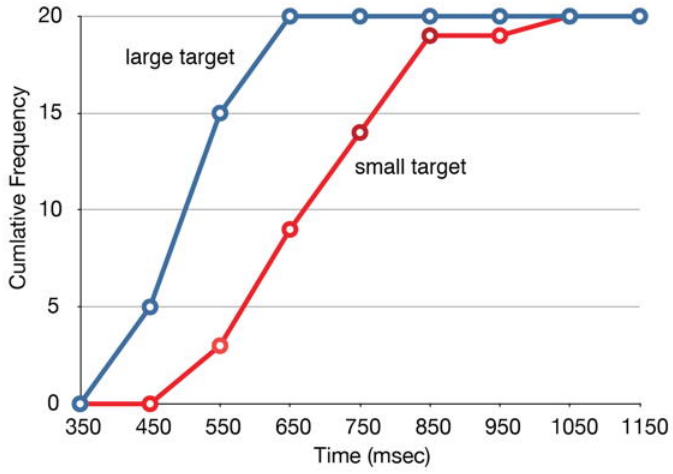


Figure* 4.5.1*: *Overlaid cumulative frequency polygons*

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4.5.4 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

Contributor

- Template:ContribHeblLane

This page titled 4.5: Frequency Polygons is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.6: Box Plots

Skills to Develop

- Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
- Create a box plot
- Create parallel box plots
- Determine whether a box plot is appropriate for a given data set

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section, we present another important graph called a box plot. Box plots are useful for identifying outliers and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. As part of the "Stroop Interference Case Study," students in introductory statistics were presented with a page containing **30** colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the **16** men and **31** women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25^{th}, 50^{th}, *and* 75^{th} percentiles in the distribution of scores. Figure **4.6.1**shows how these three statistics are used. For each gender, we draw a box extending from the 25^{th} percentile to the 75^{th} percentile. The 50^{th} percentile is drawn inside the box. Therefore,

- the bottom of each box is the 25^{th} percentile,
- the top is the 75^{th} percentile,
- and the line in the middle is the 50^{th} percentile.

The data for the women in our sample are shown in Table **4.6.1**.

***Table* 4.6.1:** *Women's times*

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25^{th} percentile is **17**, the 50^{th} percentile is **19**, and the 75^{th} percentile is **20**. For the men (whose data are not shown), the 25^{th} percentile is **19**, the 50^{th} percentile is **22.5**, and the 75^{th} percentile is **25.5**.

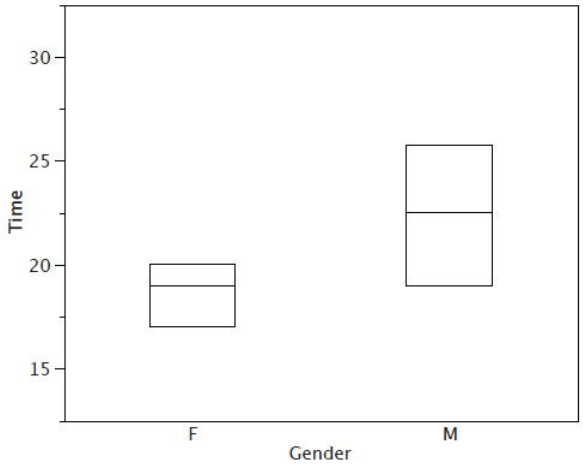


***Figure* 4.6.1:** *The first step in creating box plots*

Before proceeding, the terminology in Table **4.6.2** is helpful.

Table 4.6.2*:* *Box plot terms and values for women's times*

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (**24** and **14** for the women's data).

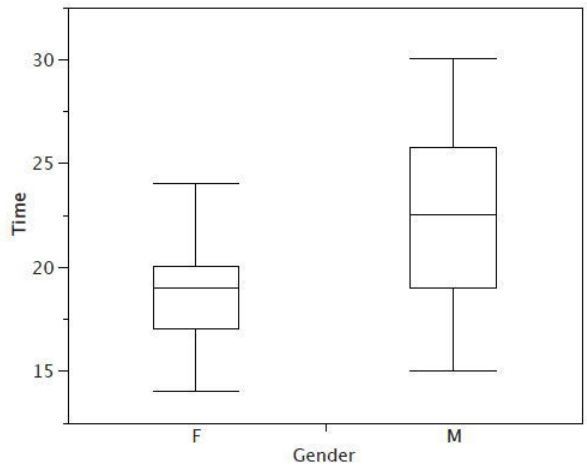


Figure **4.6.2***:* *The box plots with the whiskers drawn*

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "$o's$" and far out values are indicated by asterisks ($*$). In our data, there are no far out values and just one outside value. This outside value of **29** is for the women and is shown in Figure **4.6.3**.

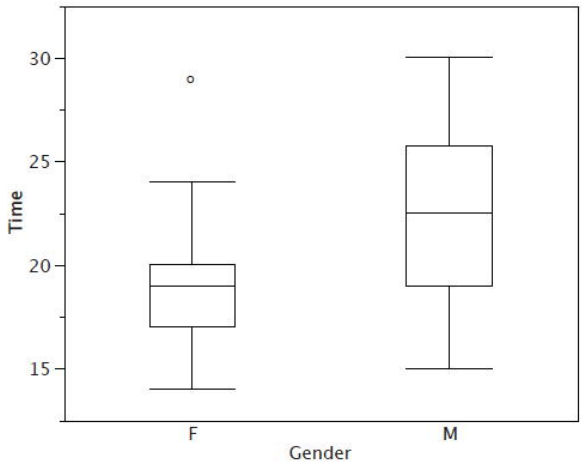


Figure* 4.6.3*: *The box plots with the outside value shown*

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure **4.6.4** shows the result of adding means to our box plots.

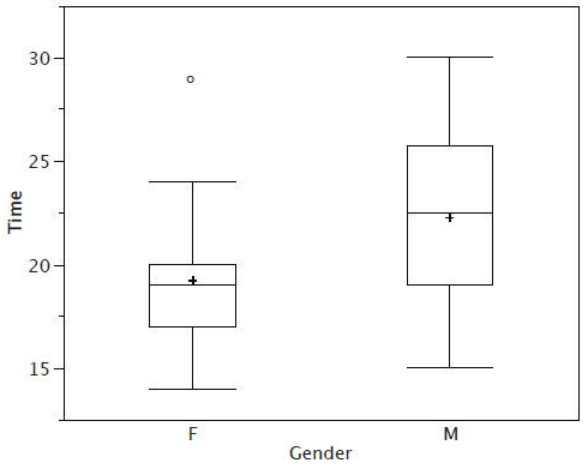


Figure* 4.6.4*: *The completed box plots*

Figure **4.6.4** provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25^{th} and 75^{th} percentiles), we see that half the women's times are between **17** and **20** seconds, whereas half the men's times are between **19** and **25.5**. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure **4.6.5** shows the box plot for the women's data with detailed labels.

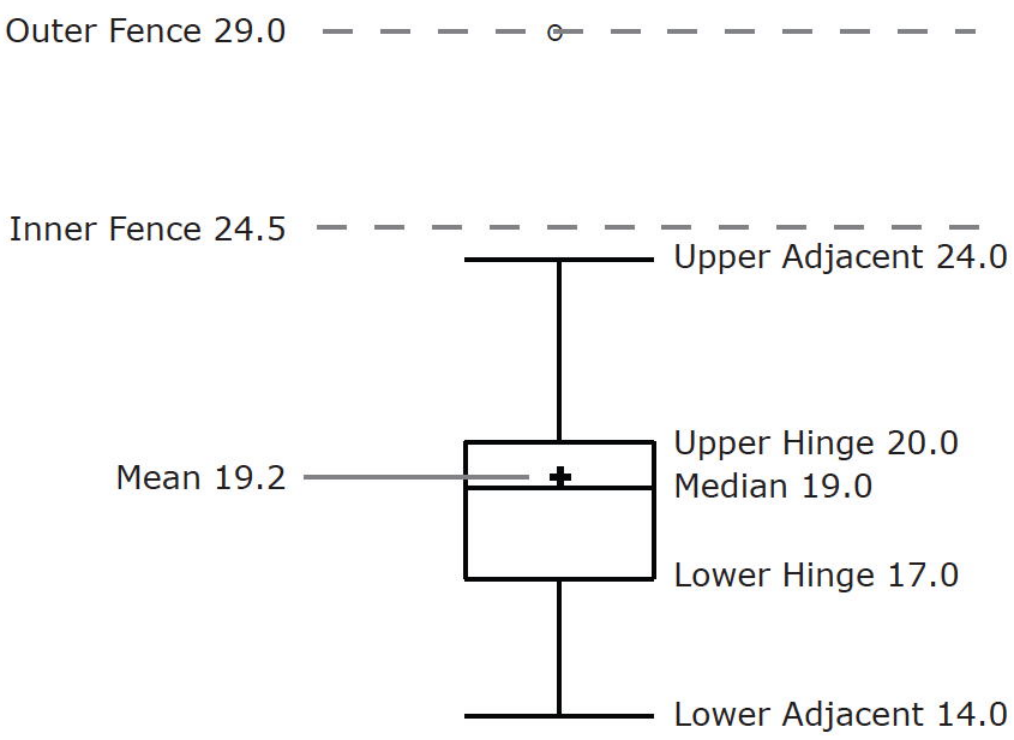


Figure* 4.6.5*: *The box plot for the women's data with detailed labels*

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram and/or a stem and leaf display.

Here are some other examples of box plots:

Example 4.6.1: Time to move the mouse over a target

The data come from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On **20** of the trials, the target was a small rectangle; on the other **20**, the target was a large rectangle. Time to reach the target was recorded on each trial. The box plots of the two distributions are shown below. You can see that although there is some overlap in times, it generally took longer to move the mouse to the small target than to the large one.

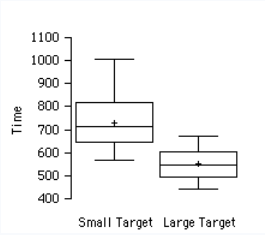


Figure* 4.6.6*: *Target Box Plots*

Example 4.6.2: Draft lottery

In **1969** the war in Vietnam was at its height. An agency called the *Selective Service* was charged with finding a fair procedure to determine which young men would be conscripted ("drafted") into the U.S. military. The procedure was supposed to be fair in the sense of not favoring any culturally or economically defined subgroup of American men. It was decided that choosing "draftees" solely on the basis of a person's birth date would be fair. A birthday lottery was thus devised. Pieces of paper representing the **366** days of the year (including **February 29**) were placed in plastic capsules, poured into a rotating drum, and

then selected one at a time. The lower the draft number, the sooner the person would be drafted. Men with high enough numbers were not drafted at all.

The first number selected was **258**, which meant that someone born on the 258^{th} day of the year (**September 14**) would be among the first to be drafted. The second number was **115**, so someone born on the 115^{th} day (**April 24**) was among the second group to be drafted. All **366** birth dates were assigned draft numbers in this way.

To crate box plots, we divided the **366** days of the year into thirds. The first third goes from **January 1 to May 1**, the second from **May 2 to August 31**, and the last from **September 1 to December 31**. The three groups of birth dates yield three groups of draft numbers. The draft number for each birthday is the order it was picked in the drawing. The figure below contains box plots of the three sets of draft numbers. As you can see, people born later in the year tended to have lower draft numbers.

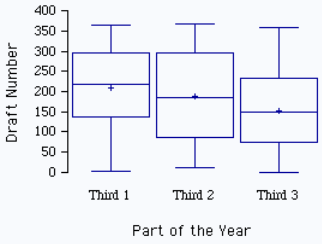


Figure* 4.6.7*: *Draft Lottery Box Plots*

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure **4.6.8** are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.
5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (**31** women and **16** men).

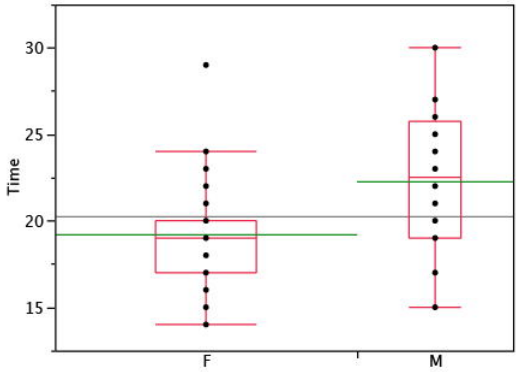


Figure* 4.6.8*: *Box plots showing the individual scores and the means*

Each dot in Figure **4.6.8** represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter thepoints. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure **4.6.9** shows what jittering looks like.

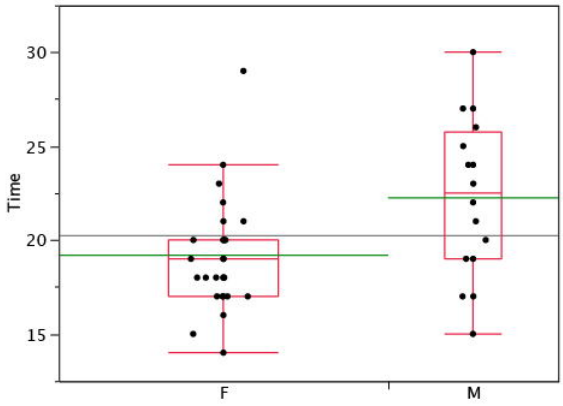


Figure* 4.6.9*: *Box plots with the individual scores jittered*

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

Contributor

- Template:ContribHeblLane

This page titled 4.6: Box Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.7: Box Plot Demo

Skills to Develop

- Understand what it means for a distribution to balance on a fulcrum
- Learn which measure of central tendency will balance a distribution

Instructions

- The box plot in the simulation is based on the data shown to the left of the box plot.
- Notice that various aspects of the box plot such as the mean and median are labeled.
  - These labels can be hidden by unchecking the "show labels on box plot" button.
- Beneath the data is a pair of buttons that let you specify whether you want to be able to enter data or to see statistics based on the data.
- When you use the simulation, try to modify the data in various ways and see how it affects the box plot.
  - Try putting in some extreme values and see if they get labeled as outliers.
  - Outside values are shown as "$o's$" beyond the inner fence.
  - Far out values are shown as $*'s$ outside the outer fence.
- You can delete all the data (by pressing the "Clear All button") and enter your own data.
  - Your data can be typed in or pasted in from another application.
  - When pasting, use keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).
  - When you change the data, the box plot will disappear.

After you have entered new data, click the "Draw box plot" button to redraw the box plot.

Illustrated Instructions

The screenshot below shows the box plot simulation with its default data on the left.

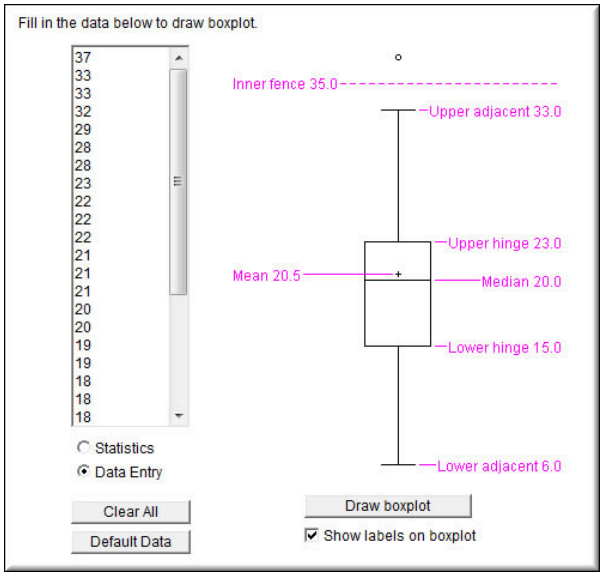


Figure* 4.7.1*: *Box Plot simulation*

You can change some or all of the data on the left by editing individual numbers or pasting in your own data. When pasting, you must use the keyboard shortcut for pasting (Command-V for Mac, CTRL-V for Windows).

Contributor

- Template:ContribHeblLane

This page titled 4.7: Box Plot Demo is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.8: Bar Charts

Skills to Develop

- Create and interpret bar charts
- Judge whether a bar chart or another graph such as a box plot would be more appropriate

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 4.8.1 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

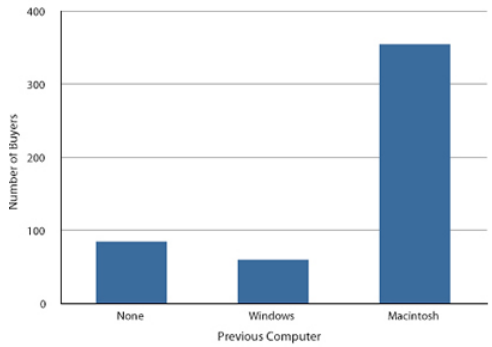


***Figure** 4.8.1**:** iMac buyers as a function of previous computer ownership*

In this section, we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 4.8.2 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24 2000 to May 24 2001. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity *percentage increase.*

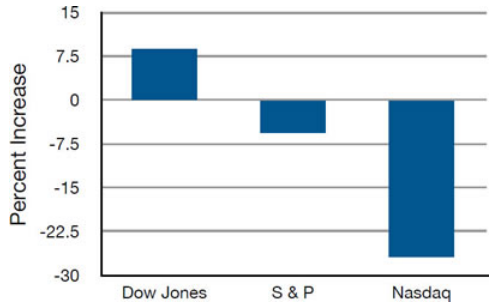


***Figure** 4.8.2**:** Percent increase in three stock indexes from May 24th 2000 to May 24th 2001*

Bar charts are particularly effective for showing change over time. Figure 4.8.3, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

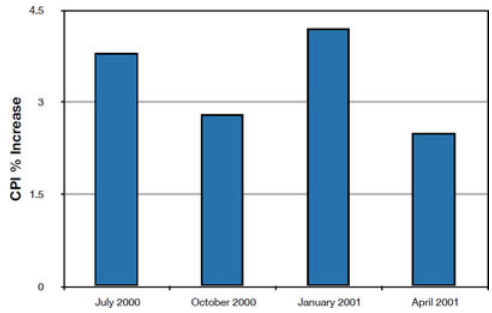


Figure* 4.8.3*: *Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated*

Bar charts are often used to compare the means of different experimental conditions. Figure **4.8.4** shows the mean time it took one of us (DL) to move the mouse to either a small target or a large target. On average, more time was required for small targets than for large ones.

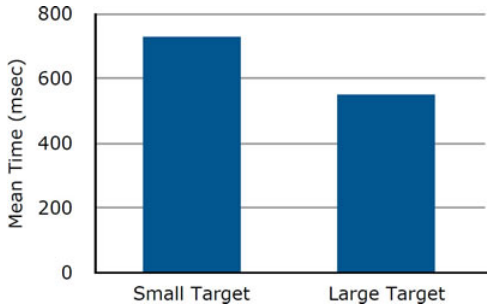


Figure* 4.8.4*: *Bar chart showing the means for the two conditions*

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the mouse-movement data is shown in Figure **4.8.5**. You can see that Figure **4.8.5** reveals more about the distribution of movement times than does Figure **4.8.4**.

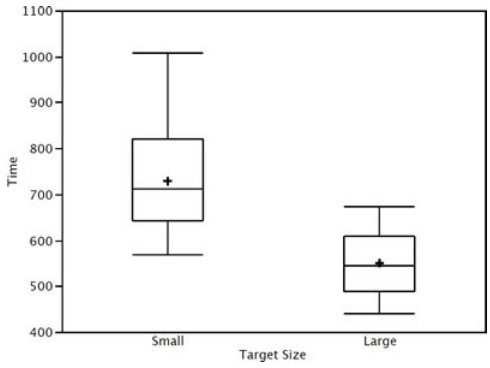


Figure* 4.8.5*: *Box plots of times to move the mouse to the small and large targets*

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

Contributors

- Template:ContribHeblLane

This page titled 4.8: Bar Charts is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.9: Line Graphs

Skills to Develop

- Create and interpret line graphs
- Judge whether a line graph would be appropriate for a given data set

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure 4.9.1 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

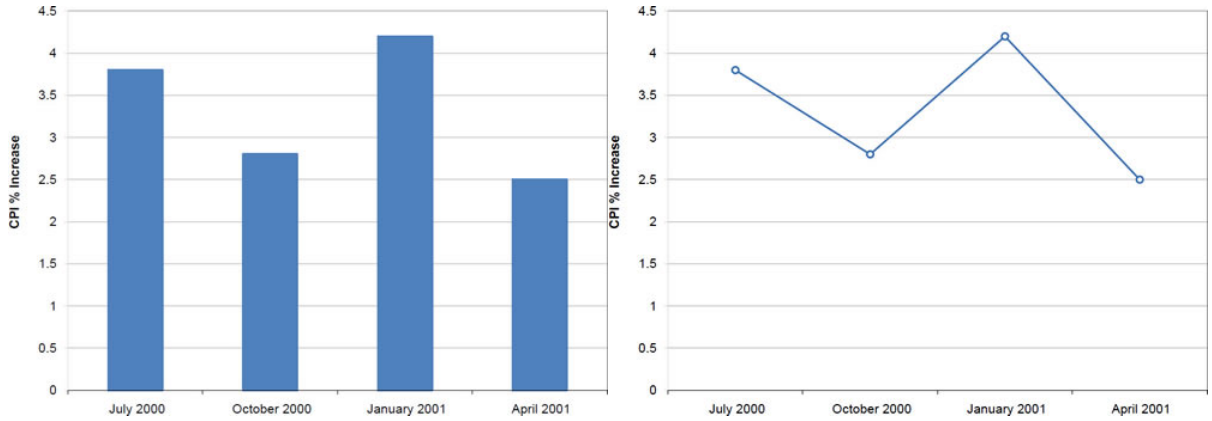


Figure* 4.9.1*: (left) *A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated. (right)* ***Figure* 4.9.2*:*** *A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.*

A line graph of these same data is shown in Figure 4.9.2. Although the figures are similar, the line graph emphasizes the change from period to period.

Line graphs are appropriate only when both the ***X***- and ***Y***-axes display ordered (rather than qualitative) variables. Although bar graphs can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 4.9.3, for example, shows percent increases and decreases in five components of the Consumer Price Index (CPI). The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

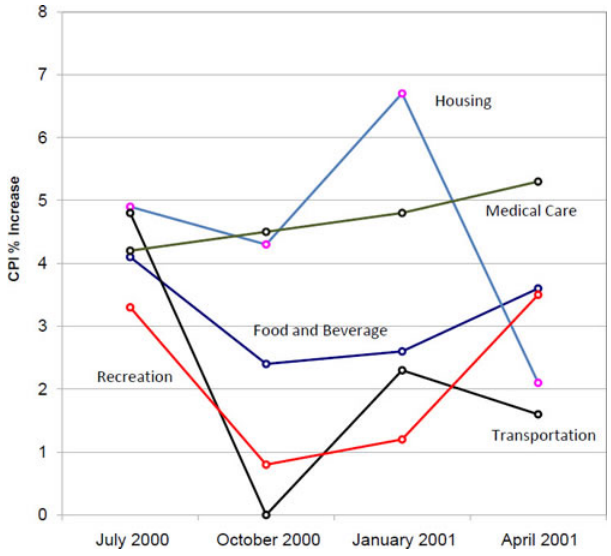


Figure* 4.9.3*: *A line graph of the percent change in five components of the CPI over time.*

Let us stress that it is misleading to use a line graph when the ***X***-axis contains merely qualitative variables. Figure **4.9.4** inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure **4.9.4** is that it gives the false impression that the games are naturally ordered in a numerical way.

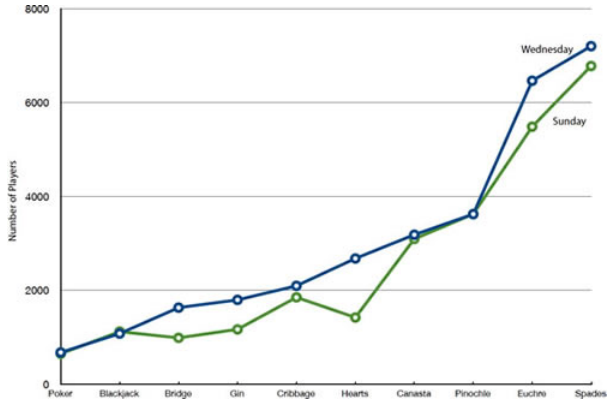


Figure* 4.9.4*: *A line graph, inappropriately used, depicting the number of people playing different card games on Sunday and Wednesday.*

Contributors

- Template:ContribHeblLane

This page titled 4.9: Line Graphs is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.10: Dot Plots

Skills to Develop

- Create and interpret dot plots
- Judge whether a dot plot would be appropriate for a given data set

Dot plots can be used to display various types of information. Figure 4.10.1 uses a dot plot to display the number of M & M's of each color found in a bag of M & M's. Each dot represents a single M & M. From the figure, you can see that there were 3 blue M & M's, 19 brown M & M's, etc.

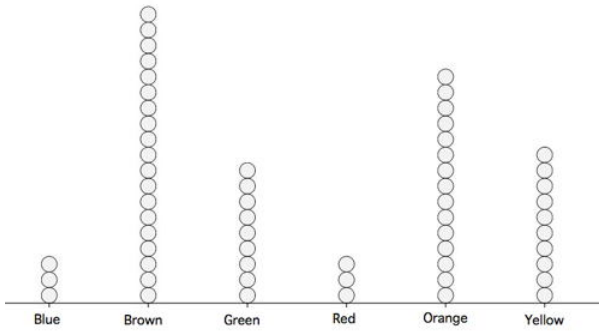


Figure* 4.10.1*: *A dot plot showing the number of M & M's of various colors in a bag of M & M's*

The dot plot in Figure 4.10.2 shows the number of people playing various card games on the Yahoo website on a Wednesday. Unlike Figure 4.10.1, the location rather than the number of dots represents the frequency.

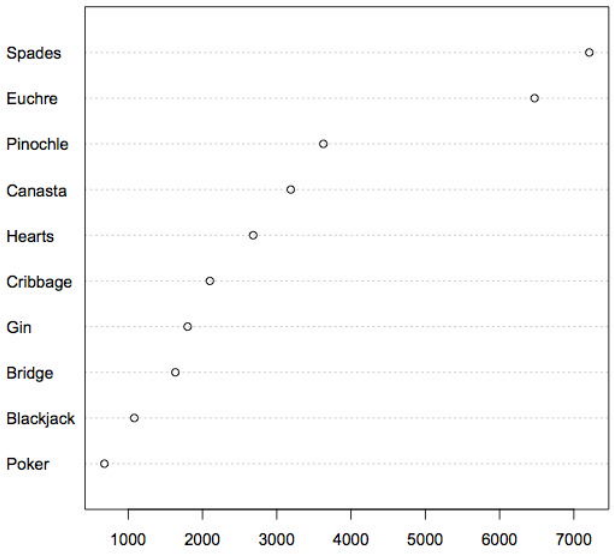


Figure* 4.10.2*: *A dot plot showing the number of people playing various card games on a Wednesday*

The dot plot in Figure 4.10.3 shows the number of people playing on a Sunday and on a Wednesday. This graph makes it easy to compare the popularity of the games separately for the two days, but does not make it easy to compare the popularity of a given game on the two days.

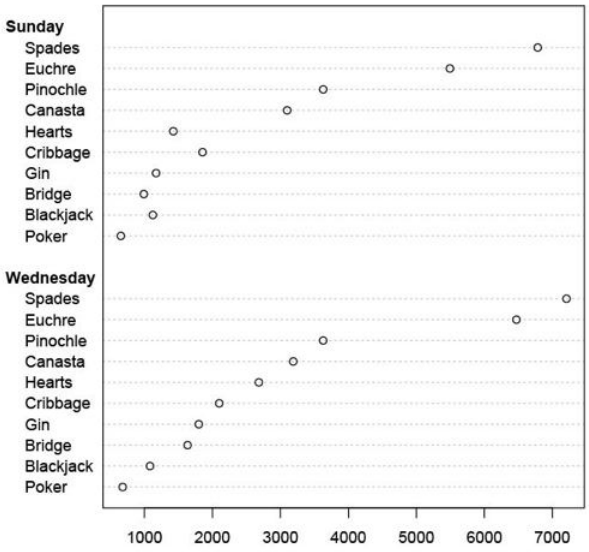


Figure* 4.10.3*: *A dot plot showing the number of people playing various card games on a Sunday and on a Wednesday*

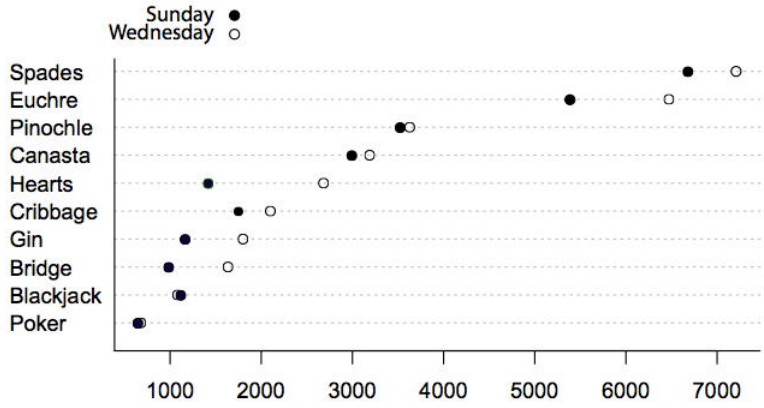


Figure* 4.10.4*: *An alternate way of showing the number of people playing various card games on a Sunday and on a Wednesday*

The dot plot in Figure **4.10.4** makes it easy to compare the days of the week for specific games while still portraying differences among games.

Contributors

- Template:ContribHeblLane

This page titled 4.10: Dot Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.11: Statistical Literacy

Are Commercial Vehicles in Texas Unsafe?

Prerequisites

Graphing Distributions

A news report on the safety of commercial vehicles in Texas stated that one out of five commercial vehicles have been pulled off the road in 2012 because they were unsafe. In addition, 12,301 commercial drivers have been banned from the road for safety violations.

The author presents the bar chart below to provide information about the percentage of fatal crashes involving commercial vehicles in Texas since 2006. The author also quotes DPS director Steven McCraw:

> Commercial vehicles are responsible for approximately 15 percent of the fatalities in Texas crashes. Those who choose to drive unsafe commercial vehicles or drive a commercial vehicle unsafely pose a serious threat to the motoring public.

Example 4.11.1

Based on what you have learned in this chapter, does this bar chart below provide enough information to conclude that unsafe or unsafely driven commercial vehicles pose a serious threat to the motoring public? What might you conclude if 30 percent of all the vehicles on the roads of Texas in 2010 were commercial and accounted for 16 percent of fatal crashes?

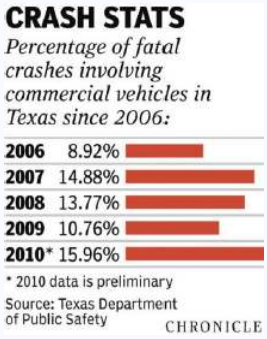


***Figure* 4.11.1:** *Crash Statistics for commercial vehicles in Texas*

Solution:

This bar chart does not provide enough information to draw such a conclusion because we don't know, on the average, in a given year what percentage of all vehicles on the road are commercial vehicles. For example, if 30 percent of all the vehicles on the roads of Texas in 2010 are commercial ones and only 16 percent of fatal crashes involved commercial vehicles, then commercial vehicles are safer than non-commercial ones. Note that in this case 70 percent of vehicles are non-commercial and they are responsible for 84 percent of the fatal crashes.

Linear By Design

Example 4.11.2

Fox News aired the line graph below showing the number unemployed during four quarters between 2007 and 2010.



Figure* 4.11.2*: *Fox news graph showing job loss by quarter*

Does Fox News' line graph provide misleading information? Why or Why not?

Solution:

There are major flaws with the Fox News graph. First, the title of the graph is misleading. Although the data show the number unemployed, Fox News' graph is titled "Job **Loss** by Quarter." Second, the intervals on the x-axis are misleading. Although there are **6** months between September **2008** and March **2009** and **15** months between March **2009** and June **2010**, the intervals are represented in the graph by very similar lengths. This gives the false impression that unemployment increased steadily.

The graph presented below is corrected so that distances on the x-axis are proportional to the number of days between the dates. This graph shows clearly that the rate of increase in the number unemployed is greater between September **2008** and March **2009** than it is between March **2009** and June **2010**.

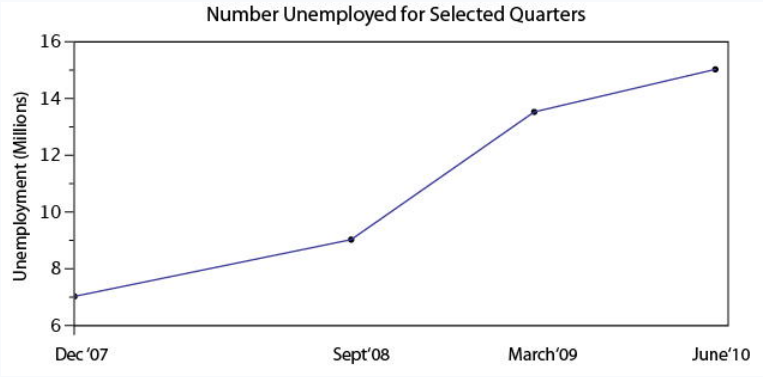


Figure* 4.11.3*: *Corrected Fox News graph*

Contributors

- Template:ContribHeblLane
- Seyd Ercan and David Lane

This page titled 4.11: Statistical Literacy is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

4.12: Graphing Distributions (Exercises)

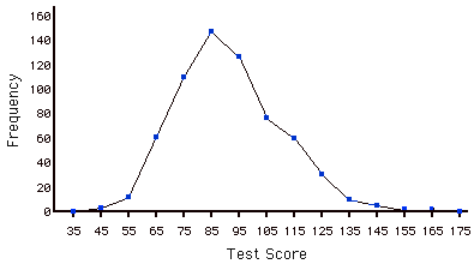
General Questions

Q1

Name some ways to graph quantitative variables and some ways to graph qualitative variables. (relevant section & relevant section)

Q2

Based on the frequency polygon displayed below, the most common test grade was around what score? Explain. (relevant section)



Q3

An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from **three** chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots? (relevant section)

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

Q4

You have to decide between displaying your data with a histogram or with a stem and leaf display. What factor(s) would affect your choice? (relevant section & relevant section)

Q5

In a box plot, what percent of the scores are between the lower and upper hinges? (relevant section)

Q6

A student has decided to display the results of his project on the number of hours people in various countries slept per night. He compared the sleeping patterns of people from the US, Brazil, France, Turkey, China, Egypt, Canada, Norway, and Spain. He was planning on using a line graph to display this data. Is a line graph appropriate? What might be a better choice for a graph? (relevant section & relevant section)

Q7

For the data from the **1977 Stat. and Biom. 200** class for eye color, construct: (relevant section)

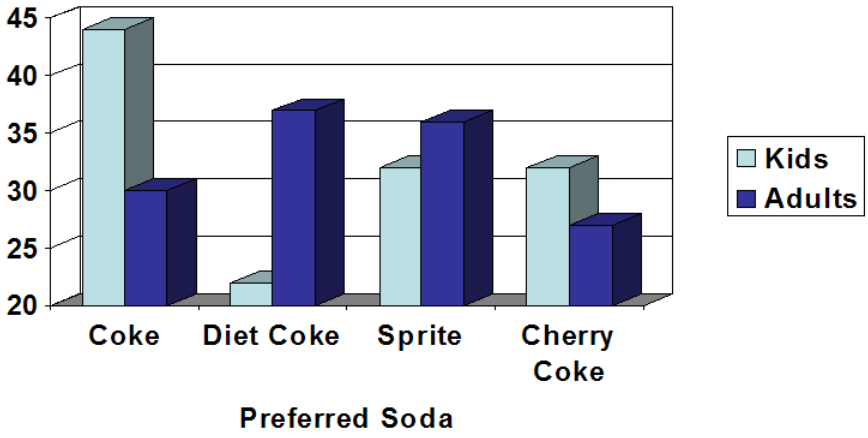
a. pie graph
b. horizontal bar graph
c. vertical bar graph
d. a frequency table with the relative frequency of each eye color

Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

(Question submitted by J. Warren, UNH)

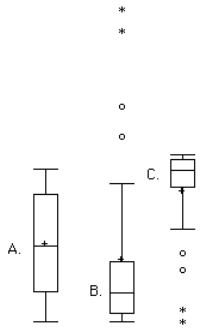
Q8

A graph appears below showing the number of adults and children who prefer each type of soda. There were **130** adults and kids surveyed. Discuss some ways in which the graph below could be improved. (relevant section)



Q9

Which of the box plots below has a large positive skew? Which has a large negative skew? (relevant section & relevant section)



Questions from Case Studies

The following questions are from the Angry Moods (AM) case study.

Q10

(AM#6) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the "Anger-Out" scores:

a. Create parallel box plots. (relevant section)
b. Create a back to back stem and leaf displays (You may have trouble finding a computer to do this so you may have to do it by hand.) (relevant section)

Q11

(AM#9) Create parallel box plots for the Anger-In scores by sports participation. (relevant section)

Q12

(AM#11) Plot a histogram of the distribution of the Control-Out scores. (relevant section)

Q13

(AM#14) Create a bar graph comparing the mean Control-In score for the athletes and the non-athletes. What would be a better way to display this data? (relevant section)

Q14

(AM#18) Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger? (relevant section)

The following questions are from the Flatulence (F) case study.

Q15

(F#1) Plot a histogram of the variable "per day." (relevant section)

Q16

(F#7) Create parallel box plots of "how long" as a function gender. Why is the 25^{th} percentile not showing? What can you say about the results? (relevant section)

Q17

(F#9) Create a stem and leaf plot of the variable "how long" What can you say about the shape of the distribution? (relevant section.1)

The following questions are from the Physicians' Reactions (PR) case study.

Q18

(PR#1) Create box plots comparing the time expected to be spent with the average-weight and overweight patients. (relevant section)

Q19

(PR#4) Plot histograms of the time spent with the average-weight and overweight patients. (relevant section)

Q20

(PR#5) To which group does the patient with the highest expected time belong?

The following questions are from the Smiles and Leniency (SL) case study

Q21

(SL#1) Create parallel box plots for the four conditions. (relevant section)

Q22

(SL#3) Create back to back stem and leaf displays for the false smile and neutral conditions. (It may be hard to find a computer program to do this for you, so be prepared to do it by hand). (relevant section)

The following questions are from the ADHD Treatment (AT) case study.

Q23

(AT#3) Create a line graph of the data. Do certain dosages appear to be more effective than others? (relevant section)

Q24

(AT#5) Create a stem and leaf plot of the number of correct responses of the participants after taking the placebo ($d0$ variable). What can you say about the shape of the distribution? (relevant section)

Q25

Create box plots for the four conditions. You may have to rearrange the data to get a computer program to create the box plots.

The following question is from the SAT and College GPA case study.

Q26

Create histograms and stem and leaf displays of both high-school grade point average and university grade point average. In what way(s) do the distributions differ?

Q27

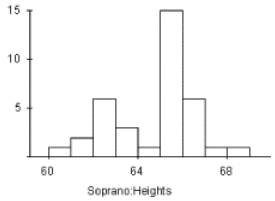
The **April 10th** issue of the Journal of the American Medical Association reports a study on the effects of anti-depressants. The study involved **340** subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (Pfizer's cousin of Lilly's Prozac) or placebo for an **8**-week period. The following are the mean scores (approximately) for the three groups of subjects over the eight-week experiment. The first column is the baseline. Lower scores mean less depression. Create a graph to display these means.

Placebo	22.5	19.1	17.9	17.1	16.2	15.1	12.1	12.3
Wort	23.0	20.2	18.2	18.0	16.5	16.1	14.2	13.0
Zoloft	22.4	19.2	16.6	15.5	14.2	13.1	11.8	10.5

The following questions are from . Visit the site

Q28

For the graph below, of heights of singers in a large chorus, please write a complete description of the histogram. Be sure to comment on all the important features.



Q29

Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired.

a. What is on the Y-axis? Explain.
b. What is on the X-axis?
c. What would be the probable shape of the salary distribution? Explain why.

Select Answers

S2

85

Contributor

- Template:ContribHeblLane

This page titled 4.12: Graphing Distributions (Exercises) is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

CHAPTER OVERVIEW

5: Probability Topics

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.

5.1: Introduction
5.2: Terminology
5.3: Independent and Mutually Exclusive Events
5.4: Two Basic Rules of Probability
5.5: Contingency Tables
5.6: Tree and Venn Diagrams
5.7: Probability Topics (Worksheet)
5.8: Probability Topics (Exericses)

Contributors

- Template:ContribOpenStax

This page titled 5: Probability Topics is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.1: Introduction

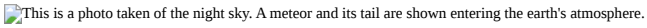

This is a photo taken of the night sky. A meteor and its tail are shown entering the earth's atmosphere.

Figure 3.1.1. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

CHAPTER OBJECTIVES

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

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

COLLABORATIVE EXERCISE

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

This page titled 5.1: Introduction is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.2: Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is 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).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event E = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

Definition: law of large numbers

This important characteristic of probability experiments is known as the law of large numbers which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are

made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

The "OR" Event

An outcome is in the event **A OR B** if the outcome is in **A** or is in **B** or is in both **A** and **B**. For example, let $\mathbf{A} = \{1, 2, 3, 4, 5\}$ and $\mathbf{B} = \{4, 5, 6, 7, 8\}$. **A OR B** $= \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event **A AND B** if the outcome is in both **A** and **B** at the same time. For example, let **A** and **B** be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then **A AND B** $= 4, 5$.

The **complement** of event **A** is denoted **A'** (read "A prime"). **A'** consists of all outcomes that are **NOT** in **A**. Notice that $P(\mathbf{A}) + P(\mathbf{A'}) = 1$. For example, let $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$ and let $\mathbf{A} = 1, 2, 3, 4$. Then, $\mathbf{A'} = 5, 6$. $P(A) = \frac{2}{6}$, $P(\mathbf{A'}) = \frac{2}{6}$, and $P(\mathbf{A}) + P(\mathbf{A'}) = \frac{4}{6} + \frac{2}{6} = 1$

The conditional probability of **A** given **B** is written $P(\mathbf{A}|\mathbf{B})$. $P(\mathbf{A}|\mathbf{B})$ 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 $P(\mathbf{A}|\mathbf{B})$ is $P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{A} \text{ AND } \mathbf{B})}{P(\mathbf{B})}$ where $P(\mathbf{B})$ is greater than zero.

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

$$P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{A} \text{ AND } \mathbf{B})}{P(\mathbf{B})} \tag{5.2.1}$$

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

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 5.2.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____

Let event $A =$ the even numbers and event $B =$ numbers greater than 13.

b. $A =$ _____, $B =$ _____
c. $P(\mathbf{A}) =$ _____, $P(\mathbf{B}) =$ _____
d. **A AND B** = _____, **A OR B** = _____
e. $P(\mathbf{A} \text{ AND } \mathbf{B}) =$ _____, $P(\mathbf{A} \text{ OR } \mathbf{B}) =$ _____
f. $\mathbf{A'} =$ _____, $P(\mathbf{A'}) =$ _____
g. $P(\mathbf{A}) + P(\mathbf{A'}) =$ _____
h. $P(\mathbf{A}|\mathbf{B}) =$ _____, $P(\mathbf{B}|\mathbf{A}) =$ _____; are the probabilities equal?

Answer

a. $\mathbf{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$

b. $\mathrm{A} = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, \mathrm{B} = \{14, 15, 16, 17, 18, 19\}$
c. $P(\mathrm{A}) = \frac{9}{19}$, $P(\mathrm{B}) = \frac{6}{19}$
d. $\mathrm{A\ AND\ B} = \{14, 16, 18\}$, $\mathrm{A\ OR\ B} = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(\mathrm{A\ AND\ B}) = \frac{3}{19}$, $P(\mathrm{A\ OR\ B}) = \frac{12}{19}$
f. $\mathrm{A}' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(\mathrm{A}') = \frac{10}{19}$
g. $P(\mathrm{A}) + P(\mathrm{A}') = 1 \left(\left(\frac{9}{19} + \frac{10}{19} = 1\right)\right.$
h. $P(\mathrm{A|B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}} = \frac{3}{6}$, $P(\mathrm{B|A}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(A)}} = \frac{3}{9}$, No

Exercise 5.2.1

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ _____
Let event $A =$ the sum is even and event $B =$ the first number is prime.
b. $A =$ _____, $B =$ _____
c. $P(\mathrm{A}) =$ _____, $P(\mathrm{B}) =$ _____
d. $\mathrm{A\ AND\ B} =$ _____, $\mathrm{A\ OR\ B} =$ _____
e. $P(\mathrm{A\ AND\ B}) =$ _____, $P(\mathrm{A\ OR\ B}) =$ _____
f. $\mathrm{B}' =$ _____, $P(\mathrm{B}') =$ _____
g. $P(\mathrm{A}) + P(\mathrm{A}') =$ _____
h. $P(\mathrm{A|B}) =$ _____, $P(\mathrm{B|A}) =$ _____; are the probabilities equal?

Answer

a. $\mathrm{S} = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
b. $\mathrm{A} = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$
$\mathrm{B} = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
c. $P(\mathrm{A}) = \frac{1}{2}$, $P(\mathrm{B}) = \frac{2}{3}$
d. $\mathrm{A\ AND\ B} = \{(2,2), (2,4), (3,1), (3,3)\}$
$\mathrm{A\ OR\ B} = \{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
e. $P(\mathrm{A\ AND\ B}) = \frac{1}{3}, P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
f. $\mathrm{B}' = \{(1,1), (1,2), (1,3), (1,4)\}, P(\mathrm{B}') = \frac{1}{3}$
g. $P(\mathrm{B}) + P(\mathrm{B}') = 1$
h. $P(\mathrm{A|B}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{1}{2}, P(\mathrm{B|A}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{2}{3}$, No.

Example 5.2.2A

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event $\mathrm{T} =$ the outcome is two.
b. Event $\mathrm{A} =$ the outcome is an even number.
c. Event $\mathrm{B} =$ the outcome is less than four.
d. The complement of A.
e. $\mathrm{A\ GIVEN\ B}$
f. $\mathrm{B\ GIVEN\ A}$
g. $\mathrm{A\ AND\ B}$
h. $\mathrm{A\ OR\ B}$
i. $\mathrm{A\ OR\ B}'$
j. Event $\mathrm{N} =$ the outcome is a prime number.
k. Event $\mathrm{I} =$ the outcome is seven.

Solution

a. $\mathrm{T} = \{2\}$, $P(\mathrm{T}) = \frac{1}{6}$

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

Example 5.2.2B

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events $M =$ the subject is male, $F =$ the subject is female, $R =$ the subject is right-handed, $L =$ the subject is left-handed. Compute the following probabilities:

a. $P(\text{M})$
b. $P(\text{F})$
c. $P(\text{R})$
d. $P(\text{L})$
e. $P(\text{M AND R})$
f. $P(\text{F AND L})$
g. $P(\text{M OR F})$
h. $P(\text{M OR R})$
i. $P(\text{F OR L})$
j. $P(\text{M'})$
k. $P(\text{R|M})$
l. $P(\text{F|L})$
m. $P(\text{L|F})$

Answer

a. $P(\text{M}) = 0.52$
b. $P(\text{F}) = 0.48$
c. $P(\text{R}) = 0.87$
d. $P(\text{L}) = 0.13$
e. $P(\text{M AND R}) = 0.43$
f. $P(\text{F AND L}) = 0.04$
g. $P(\text{M OR F}) = 1$
h. $P(\text{M OR R}) = 0.96$
i. $P(\text{F OR L}) = 0.57$
j. $P(\text{M'}) = 0.48$
k. $P(\text{R|M}) = 0.8269$ (rounded to four decimal places)
l. $P(\text{F|L}) = 0.3077$ (rounded to four decimal places)
m. $P(\text{L|F}) = 0.0833$

References

1. “Countries List by Continent.” Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

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

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

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

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:

- $0 \leq P(\mathrm{A}) \leq 1$
- If A and B are any two mutually exclusive events, then $\mathrm{P(A\ OR\ B)} = P(\mathrm{A}) + P(\mathrm{B})$.
- $P(\mathrm{S}) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event $\mathrm{A\ AND\ B}$ if the outcome is in both $\mathrm{A\ AND\ B}$ 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

- Template:ContribOpenStax

Exercise 3.2.2

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Answer

a. $P(L') = P(S)$
b. $P(M \text{ OR } S)$
c. $P(F \text{ AND } L)$
d. $P(M|L)$
e. $P(L|M)$
f. $P(S|F)$
g. $P(F|L)$
h. $P(F \text{ OR } L)$
i. $P(M \text{ AND } S)$
j. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let $H =$ the event of getting a hat.

Let $N =$ the event of getting a noisemaker.

Let $F =$ the event of getting a finger trap.

Let $C =$ the event of getting a bag of confetti.

Exercise 3.2.3

Find $P(H)$.

Exercise 3.2.4

Find $P(\mathrm{N})$.

Answer

$P(\mathrm{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise 3.2.5

Find $P(\mathrm{F})$.

Exercise 3.2.6

Find $P(\mathrm{C})$.

Answer

$P(\mathrm{C}) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let $B =$ the event of getting a blue jelly bean

Let $G =$ the event of getting a green jelly bean.

Let $O =$ the event of getting an orange jelly bean.

Let $P =$ the event of getting a purple jelly bean.

Let $R =$ the event of getting a red jelly bean.

Let $Y =$ the event of getting a yellow jelly bean.

Exercise 3.2.7

Find $P(\mathrm{B})$.

Exercise 3.2.8

Find $P(\mathrm{G})$.

Answer

$P(\mathrm{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise 3.2.9

Find $P(\mathrm{P})$.

Exercise 3.2.10

Find $P(\mathrm{R})$.

Answer

$P(\mathrm{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise 3.2.11

Find $P(\mathrm{Y})$.

Exercise 3.2.12

Find $P(\mathrm{O})$.

Answer

$P(textO) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let $\mathbf{A}$ = the event that a country is in Asia.

Let $\mathbf{E}$ = the event that a country is in Europe.

Let $\mathbf{F}$ = the event that a country is in Africa.

Let $\mathbf{N}$ = the event that a country is in North America.

Let $\mathbf{O}$ = the event that a country is in Oceania.

Let $\mathbf{S}$ = the event that a country is in South America.

Exercise 3.2.13

Find $P(\mathbf{A})$.

Exercise 3.2.14

Find $P(\mathbf{E})$.

Answer

$P(\mathbf{E}) = \frac{47}{194} = 0.24$

Exercise 3.2.15

Find $P(\mathbf{F})$.

Exercise 3.2.16

Find $P(\mathbf{N})$.

Answer

$P(\mathbf{N}) = \frac{23}{194} = 0.12$

Exercise 3.2.17

Find $P(\mathbf{O})$.

Exercise 3.2.18

Find $P(\mathbf{S})$.

Answer

$P(\mathbf{S}) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise 3.2.19

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise 3.2.20

What is the probability of drawing a club in a standard deck of 52 cards?

Answer

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise 3.2.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise 3.2.22

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Answer

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

Figure 3.2.1.

Let $\mathbf{B} =$ the event of landing on blue.

Let $\mathbf{R} =$ the event of landing on red.

Let $\mathbf{G} =$ the event of landing on green.

Let $\mathbf{Y} =$ the event of landing on yellow.

Exercise 3.2.23

If you land on $\mathbf{Y}$, you get the biggest prize. Find $P(\mathbf{Y})$.

Exercise 3.2.24

If you land on red, you don't get a prize. What is $P(\mathbf{R})$?

Answer

$\mathbf{P}(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let $\mathbf{I} =$ the event that a player in an infielder.

Let $\mathbf{O} =$ the event that a player is an outfielder.

Let $\mathbf{H} =$ the event that a player is a great hitter.

Let $\mathbf{N} =$ the event that a player is not a great hitter.

Exercise 3.2.25

Write the symbols for the probability that a player is not an outfielder.

Exercise 3.2.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(\mathbf{O}\ \mathbf{OR}\ \mathbf{H})$

Exercise 3.2.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise 3.2.28

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Answer

$P(\mathbf{H}|\mathbf{I})$

Exercise 3.2.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise 3.2.30

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Answer

$P(\text{N}|\text{O})$

Exercise 3.2.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.2.32

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Answer

$P(\text{I OR N})$

Exercise 3.2.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.2.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\text{I})$

Exercise 3.2.35

What is the word for the set of all possible outcomes?

Exercise 3.2.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.2.37

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let $\text{F} =$ event that book is fiction

Let $\text{N} =$ event that book is nonfiction

What is the sample space?

Exercise 3.2.38

What is the sum of the probabilities of an event and its complement?

Answer

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let $\text{E} =$ the event that it lands on an even number. Let $\text{M} =$ the event that it lands on a multiple of three.

Exercise 3.2.39

What does $P(\text{E}|\text{M})$ mean in words?

Exercise 3.2.40

What does $P(\text{E OR M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

This page titled 5.2: Terminology is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.3: Independent and Mutually Exclusive Events

Independent and *mutually exclusive* do not mean the same thing.

Independent Events

Two events are independent if the following are true:

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

Two events **A** and **B** are independent if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show only one of the above conditions. If two events are NOT independent, then we say that they are dependent.

Sampling a population

Sampling may be done with replacement or without replacement (Figure **5.3.1**):

- **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*.

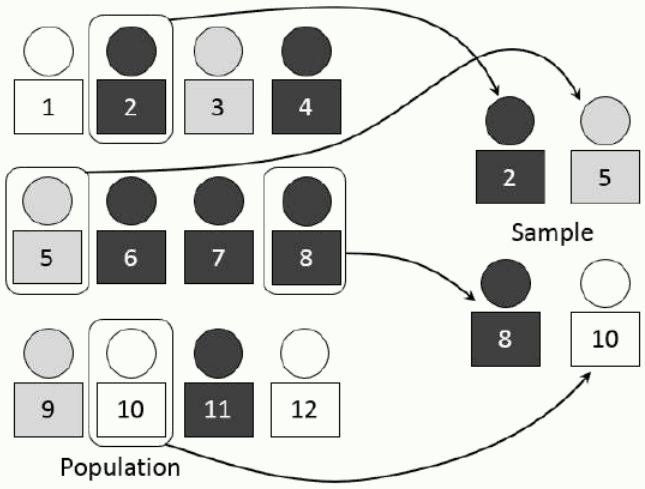


Figure* 5.3.1*: *A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". Image used with permission (CC BY-SA 4.0; Dan Kernler).*

If it is not known whether **A** and **B** are independent or dependent, assume they are dependent until you can show otherwise.

Example 5.3.1: 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.

Exercise 5.3.1

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Answer

a. With replacement
b. No

Example 5.3.2

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?

Solution

a. Without replacement
b. With replacement

Exercise 5.3.2

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. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. QS, 1D, 1C, QD
b. KH, 7D, 6D, KH
c. QS, 7D, 6D, KS

Answer

without replacement: a. Possible; b. Impossible, c. Possible

with replacement: a. Possible; c. Possible, c. Possible

Mutually Exclusive Events

A and B are mutually exclusive events if they **cannot** occur at the same time. This means that A and B do not share any outcomes and $P(\text{A AND B}) = 0$.

For example, suppose the sample space

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}. \tag{5.3.1}$$

Let $\text{A} = \{1, 2, 3, 4, 5\}, \text{B} = \{4, 5, 6, 7, 8\}$, and $\text{C} = \{7, 9\}$. A AND B $= \{4, 5\}$.

$$P(\text{A AND B}) = \frac{2}{10} \tag{5.3.2}$$

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 5.3.3

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 $\text{B} = \text{A}'$. Also, $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$.
- The probabilities for A and for B are $P(\text{A}) = \frac{3}{4}$ and $P(\text{B}) = \frac{1}{4}$.
- Let C = the event of getting all heads. $\text{C} = \{HH\}$. Since $\text{B} = \{TT\}$, $P(\text{B AND C}) = 0$. B and Care mutually exclusive. B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $\text{D} = \{TT\}$. $P(\text{D}) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $\text{E} = \{HT, HH\}$. $P(\text{E}) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. $\text{F} = \{HT, TH, TT\}$. $P(\text{F}) = \frac{3}{4}$

Exercise 5.3.3

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event $A =$ Getting at least one black card $= \{BB, BR, RB\}$

$P(\text{A}) = \frac{3}{4} = 0.75$

Example 5.3.4

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 5.3.3.

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(\text{F}) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(\text{G}) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(\text{H}) = \frac{2}{4}$
4. F and G share HH so $P(\text{F AND G})$ 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 $P(\text{J AND H}) = 0$. J and H are mutually exclusive.

Exercise 5.3.4

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Answer

a. $P(\text{F}) = \frac{1}{4}$
b. $P(\text{G}) = \frac{1}{2}$
c. $P(\text{H}) = \frac{1}{2}$
d. Yes
e. No

Example 5.3.5

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

- Find the complement of A, A′. The complement of A, A′, is B because A and B together make up the sample space. $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$. Also, $P(\text{A}) = \frac{3}{6}$ and $P(\text{B}) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then $\text{C} = \{3, 5\}$. Let event D = all even faces smaller than five. Then $\text{D} = \{2, 4\}$. $P(\text{C AND D}) = 0$ 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. $\text{E} = \{1, 2, 3, 4\}$.

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Answer

No. $\text{C} = \{3, 5\}$ and $\text{E} = \{1, 2, 3, 4\}$. $P(\text{C AND E}) = \frac{1}{6}$. To be mutually exclusive, $P(\text{C AND E})$ must be zero.

- Find $P(\text{C|A})$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(\text{C|A})$, 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, $P(\text{C|A}) = \frac{2}{3}$.

Exercise 5.3.5

Let event A = learning Spanish. Let event B = learning German. Then $\mathrm{A\ AND\ B}$ = learning Spanish and German. Suppose $P(\mathrm{A}) = 0.4$ and $P(\mathrm{B}) = 0.2$. $P(\mathrm{A\ AND\ B}) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$
- $P(\mathrm{B}|\mathrm{A})$
- $P(\mathrm{A\ AND\ B}) = P(\mathrm{A})P(\mathrm{B})$

Answer

$$P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{P(\mathrm{B})} = \frac{0.08}{0.2} = 0.4 = P(\mathrm{A}) \tag{5.3.3}$$

The events are independent because $P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A})$.

Example 5.3.6

Let event G = taking a math class. Let event H = taking a science class. Then, $\mathrm{G\ AND\ H}$ = taking a math class and a science class. Suppose $P(\mathrm{G}) = 0.6$, $P(\mathrm{H}) = 0.5$, and $P(\mathrm{G\ AND\ H}) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(\mathrm{G}|\mathrm{H}) = P(\mathrm{G})$
- $P(\mathrm{H}|\mathrm{G}) = P(\mathrm{H})$
- $P(\mathrm{G\ AND\ H}) = P(\mathrm{G})P(\mathrm{H})$

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 $P(\mathrm{G}|\mathrm{H}) = P(\mathrm{G})$.
b. b. Show $P(\mathrm{G\ AND\ H}) = P(\mathrm{G})P(\mathrm{H})$.

Solution

a. \(P(\text{G|H}) = \dfrac{P(\text{G AND H})}{P(\text{H})} = \dfrac{0.3}{0.5} = 0.6 = P(\text{G})\)
b. $P(\mathrm{G})P(\mathrm{H}) = (0.6)(0.5) = 0.3 = P(\mathrm{G\ AND\ H})$

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 $P(\mathrm{H}|\mathrm{G}) = P(\mathrm{H})$ to show that G and H are independent events.

Exercise 5.3.6

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $\mathrm{S} = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(\mathrm{G\ AND\ O})$?

Answer

Event G and $\mathrm{O} = \{G1, G3\}$

$P(\mathrm{G\ and\ O}) = \frac{2}{10} = 0.2$

Example 5.3.7

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

Exercise 5.3.7

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Answer

a. $P(B|D) = 0.6667$
b. $P(D|B) = 0.5$
c. No
d. No

Example 5.3.8

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

- R = red card is drawn,
- B = blue card is drawn,
- E = even-numbered card is drawn.

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

S has eight outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. $P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)
- $P(E) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(E|B) = \frac{2}{5}$. (There are five blue cards: $B1, B2, B3, B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E) = \frac{2}{3}$. (There are three even-numbered cards: $R2, B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.
- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = frac14$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, which means that G and H are independent.

Exercise 5.3.8

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let **A** be the event that a fan is rooting for the away team.

Let **B** be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Answer

- $P(\text{B}|\text{A}) = 0.67$
- $P(\text{B}) = 0.25$

So $P(\text{B})$ does not equal $P(\text{B}|\text{A})$ which means that $\text{B} and \text{A}$ are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND A}) = 0.20$, not 0.

Example 5.3.9

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 **F** be the event that a student is female. Let **L** be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(\text{F}) = 0.60$; $P(\text{L}) = 0.50$
- $P(\text{F AND L}) = 0.45$
- $P(\text{L}|\text{F}) = 0.75$

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(\text{F}|\text{L})$ yet, so you cannot use the second condition.

Solution 1

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

Solution 2

Check whether $P(\text{L}|\text{F})$ equals $P(\text{L})$. We are given that $P(\text{L}|\text{F}) = 0.75$, but $P(\text{L}) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Exercise 5.3.9

Mark is deciding which route to take to work. His choices are \(\text{I} =\) the Interstate and \(\text{F} =\) Fifth Street.

- $P(\text{I}) = 0.44$ and $P(\text{F}) = 0.55$
- $P(\text{I AND F}) = 0$ because Mark will take only one route to work.

What is the probability of $P(\text{I OR F})$?

Answer

Because $P(\text{I AND F}) = 0$,

$P(\text{I OR F}) = P(\text{I}) + P(\text{F}) - P(\text{I AND F}) = 0.44 + 0.56 - 0 = 1$

Example 5.3.10

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

Exercise 5.3.10

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let $textT$ be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(\text{T})$.
b. Compute $P(\text{T}|\text{F})$.
c. Are $textT$ and F independent?.
d. Are F and S mutually exclusive?
e. Are F and S independent?

Answer

a. $P(\text{T}) = \frac{1}{4}$
b. $P(\text{T}|\text{F}) = \frac{1}{2}$
c. No
d. No
e. Yes

References

1. Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).
2. Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events **A** and **B** are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

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

Contributors

- Template:ContribOpenStax

Exercise 3.3.11

E and **F** are mutually exclusive events. $P(\text{E}) = 0.4$; $P(\text{F}) = 0.5$. Find $P(\text{E}|\text{F})$.

Exercise 3.3.12

J and **K** are independent events. $P(\text{J}|\text{K}) = 0.3$. Find $P(\text{J})$.

Answer

$P(\text{J}) = 0.3$

Exercise 3.3.13

U and **V** are mutually exclusive events. $P(\text{U}) = 0.26$; $P(\text{V}) = 0.37$. Find:

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

Exercise 3.3.14

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

Answer

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

$0.1 = (0.4)P(\text{R})$

$P(\text{R}) = 0.25$

Bringing It Together

Exercise 3.3.16

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 3.3.17

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. $P(\text{C}) =$ _____
b. $P(\text{P}|\text{C}) =$ _____
c. $P(\text{P}|\text{C'}) =$ _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Answer

a. $P(\text{C}) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise 3.3.18

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 3.3.19

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

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 5.3: Independent and Mutually Exclusive Events is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.4: Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

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

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{5.4.1}$$

This rule may also be written as:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} \tag{5.4.2}$$

(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

$$P(A|B) = P(A). \tag{5.4.3}$$

and Equation 5.4.1 becomes

$$P(A \text{ AND } B) = P(A)P(B). \tag{5.4.4}$$

The Addition Rule

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

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) \tag{5.4.5}$$

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

$$P(A \text{ AND } B) = 0. \tag{5.4.6}$$

and Equation 5.4.5 becomes

$$P(A \text{ OR } B) = P(A) + P(B). \tag{5.4.7}$$

Example 5.4.1

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

- Klaus can only afford one vacation. The probability that he chooses $\mathbf{A}$ is $P(\mathbf{A}) = 0.6$ and the probability that he chooses $\mathbf{B}$ is $P(\mathbf{B}) = 0.35$.
- $P(\mathbf{A} \text{ AND } \mathbf{B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\mathbf{A} \text{ OR } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. $\mathbf{A}$ = the event Carlos is successful on his first attempt. $P(\mathbf{A}) = 0.65$. $\mathbf{B}$ = the event Carlos is successful on his second attempt. $P(\mathbf{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are $\mathbf{A}$ and $\mathbf{B}$ independent?
d. Are $\mathbf{A}$ and $\mathbf{B}$ mutually exclusive?

Solutions

a. The problem is asking you to find $P(\mathbf{A} \text{ AND } \mathbf{B}) = P(\mathbf{B} \text{ AND } \mathbf{A})$. Since $P(\mathbf{B}|\mathbf{A}) = 0.90$: $P(\mathbf{B} \text{ AND } \mathbf{A}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\text{A OR B})$.

$$P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B}) = 0.65 + 0.65 - 0.585 = 0.715 \tag{5.4.8}$$

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

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

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \tag{5.4.9}$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \tag{5.4.10}$$

So, $P(\text{B AND A})$ is **not** equal to $P(\text{B})P(\text{A})$.

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

To be mutually exclusive, $P(\text{A AND B})$ must equal zero.

Exercise 5.4.1

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(\text{C}) = 0.75$. D = the event Helen makes the second shot. $P(\text{D}) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Answer

$$P(\text{D}|\text{C}) = 0.85 \tag{5.4.11}$$

$$P(\text{C AND D}) = P(\text{D AND C}) \tag{5.4.12}$$

$$P(\text{D AND C}) = P(\text{D}|\text{C})P(\text{C}) = (0.85)(0.75) = 0.6375 \tag{5.4.13}$$

Helen makes the first and second free throws with probability 0.6375.

Example 5.4.2

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.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{5.4.14}$$

https://socialsci.libretexts.org/@go/page/19052

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{5.4.15}$$

$$0.0667 \neq 0.0996 \tag{5.4.16}$$

Exercise 5.4.2

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Answer

$$P = \frac{200 - 140 - 40}{200} = \frac{20}{200} = 0.1 \tag{5.4.17}$$

Example 5.4.3

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: $\mathbf{M}$ = math class, $\mathbf{S}$ = speech class, $\mathbf{M|S}$ = math given speech

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

Answer

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

Exercise 5.4.3

A student goes to the library. Let events $\mathbf{B}$ = the student checks out a book and $\mathbf{D}$ = the student check out a DVD. Suppose that $P(\text{B}) = 0.40, P(\text{D}) = 0.30$ and $P(\text{D|B}) = 0.5$.

a. Find $P(\text{B AND D})$.
b. Find $P(\text{B OR D})$.

Answer

a. $P(\text{B AND D}) = P(\text{D|B})P(\text{B}) = (0.5)(0.4) = 0.20$.
b. $P(\text{B OR D}) = P(\text{B}) + P(\text{D}) - P(\text{B AND D}) = 0.40 + 0.30 - 0.20 = 0.50$

Example 5.4.4

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 $\mathbf{B}$ = woman develops breast cancer and let $\mathbf{N}$ = tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative?
d. What is the probability that the woman has breast cancer or tests negative?
e. Are having breast cancer and testing negative independent events?
f. Are having breast cancer and testing negative mutually exclusive?

Answers

a. $P(\text{B}) = 0.143; P(\text{N}) = 0.85$
b. $P(\text{N}|\text{B}) = 0.02$
c. $P(\text{B AND N}) = P(\text{B})P(\text{N}|\text{B}) = (0.143)(0.02) = 0.0029$
d. $P(\text{B OR N}) = P(\text{B}) + P(\text{N}) - P(\text{B AND N}) = 0.143 + 0.85 - 0.0029 = 0.9901$
e. No. $P(\text{N}) = 0.85; P(\text{N}|\text{B}) = 0.02$. So, $P(\text{N}|\text{B})$ does not equal $P(\text{N})$.
f. No. $P(\text{B AND N}) = 0.0029$. For B and N to be mutually exclusive, $P(\text{B AND N})$ must be zero

Exercise 5.4.4

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Answer

Let A = student is a senior going to college.

Let B = student plays sports.

$P(\text{B}) = \frac{140}{200}$

$P(\text{B}|\text{A}) = \frac{50}{140}$

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

$P(\text{A AND B}) = (\frac{140}{200})(\frac{50}{140}) = \frac{1}{4}$

Example 5.4.5

Refer to the information in Example 5.4.4. 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

Exercise 5.4.5

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(\text{B}) = 0.40, P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B}')$.
b. Find $P(\text{D AND B})$.
c. Find $P(\text{B}|\text{D})$.
d. Find $P(\text{D AND B}')$.
e. Find $P(\text{D}|\text{B}')$.

Answer

a. $P(\text{B}') = 0.60$
b. $P(\text{D AND B}) = P(\text{D}|\text{B})P(\text{B}) = 0.20$
c. $P(\text{B}|\text{D}) = \frac{P(\text{B AND D})}{P(\text{D})} = \frac{(0.20)}{(0.30)} = 0.66$
d. $P(\text{D AND B}') = P(\text{D}) - P(\text{D AND B}) = 0.30 - 0.20 = 0.10$
e. $P(\text{D}|\text{B}') = P(\text{D AND B}')P(\text{B}') = (P(\text{D}) - P(\text{D AND B}))(0.60) = (0.10)(0.60) = 0.06$

References

1. DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://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 http://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 http://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 http://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 http://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).

Chapter 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: $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- $\mathbf{C}$ = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- $\mathbf{L}$ = Latino Californians

Suppose that one Californian is randomly selected.

Exercise 3.4.5

Find $P(\mathbf{C})$.

Exercise 3.4.6

Find $P(\mathbf{L})$.

Answer

0.376

Exercise 3.4.7

Find $P(\text{C}|\text{L})$.

Exercise 3.4.8

In words, what is $\text{C}|\text{L}$?

Answer

$\text{C}|\text{L}$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise 3.4.9

Find $P(\text{L AND C})$

Exercise 3.4.10

In words, what is L AND C?

Answer

L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Exercise 3.4.11

Are L and C independent events? Show why or why not.

Exercise 3.4.12

Find $P(\text{L OR C})$.

Answer

0.6492

Exercise 3.4.13

In words, what is L OR C?

Exercise 3.4.14

Are L and C mutually exclusive events? Show why or why not.

Answer

No, because $P(\text{L AND C})$ does not equal 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events 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. $P(\text{A AND B}) = P(\text{A})P(\text{B})$

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

This page titled 5.4: Two Basic Rules of Probability is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.5: 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 5.5.1

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 $P(\text{Person is a car phone user})$.
b. Find $P(\text{person had no violation in the last year})$.
c. Find $P(\text{Person had no violation in the last year AND was a car phone user})$.
d. Find $P(\text{Person is a car phone user OR person had no violation in the last year})$.
e. Find $P(\text{Person is a car phone user GIVEN person had a violation in the last year})$.
f. Find $P(\text{Person had no violation last year GIVEN person was not a car phone user})$

Answer

a. $\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$
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 car phone users.)

Exercise 5.5.1

Table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising}|\text{no injury in the last year})$?

Answer

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$

b. $P(\text{athlete stretches before exercising}|\text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example 5.5.2

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: $P(\text{___}|\text{___}) = \text{___}$.
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. $P(\text{M}|\text{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

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

d.

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

Exercise 5.5.2

Table shows a random sample of 200 cyclists and the routes they prefer. Let $\text{M} =$ males and $\text{H} =$ hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Answer

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$
$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$
$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example 5.5.3

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 5.5.4

Table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

a. Find $P(\text{2009 AND Robbery})$.
b. Find $P(\text{2010 AND Burglary})$.
c. Find $P(\text{2010 OR Burglary})$.
d. Find $P(\text{2011}|\text{Rape})$.
e. Find $P(\text{Vehicle}|\text{2008})$.

Answer

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Exercise 5.5.3

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

1. “Blood Types.” American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).
2. Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.
3. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
4. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).
5. “Human Blood Types.” Unite Blood Services, 2011. Available online at http://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 http://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).

Chapter Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. Table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Exercise 3.5.4

Find *P*(musician is a female).

Exercise 3.5.5

Find $P(\text{musician is a male AND had private instruction})$.

Answer

$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$

Exercise 3.5.6

Find *P*(musician is a female OR is self taught).

Exercise 3.5.7

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Answer

$P(\text{being a female musician AND learning music in school}) = \frac{38}{130} = \frac{19}{65} = 0.29$

$P(\text{being a female musician})\text{P}(\text{learning music in school}) = \left(\frac{72}{130}\right)\left(\frac{62}{130}\right) = \frac{4,464}{16,900} = \frac{1,116}{4,225} = 0.26$

No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

Bringing it Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Exercise 3.5.8

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Exercise 3.5.9

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Answer

$\frac{35,065}{100,450}$

Exercise 3.5.10

Find the probability that the person was Latino.

Exercise 3.5.11

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

Exercise 3.5.12

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

Exercise 3.5.13

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4,715}{15,273}$.

Exercise 3.5.14

Prove that smoking level/day and ethnicity are dependent events.

Glossary

contingency table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

This page titled 5.5: Contingency Tables is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.6: Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A *tree diagram* is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 5.6.1: 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.

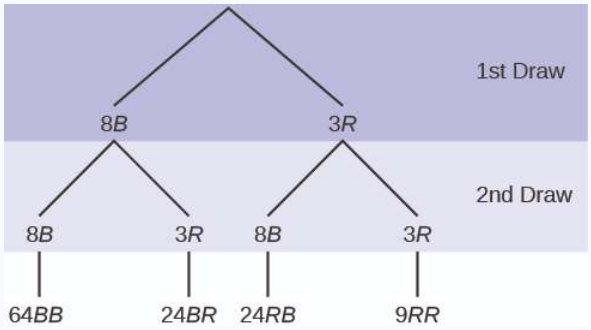


Figure 5.6.1: Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as $R1$, $R2$, and $R3$ and each blue ball as $B1$, $B2$, $B3$, $B4$, $B5$, $B6$, $B7$, and $B8$. Then the nine RR outcomes can be written as:

$R1R1$ $R1R2$ $R1R3$ $R2R1$ $R2R2$ $R2R3$ $R3R1$ $R3R2$ $R3R3$

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

Exercise 5.6.1

In a standard deck, there are 52 cards. 12 cards are face cards (event **F**) and 40 cards are not face cards (event **N**). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(\mathbf{FF})$.

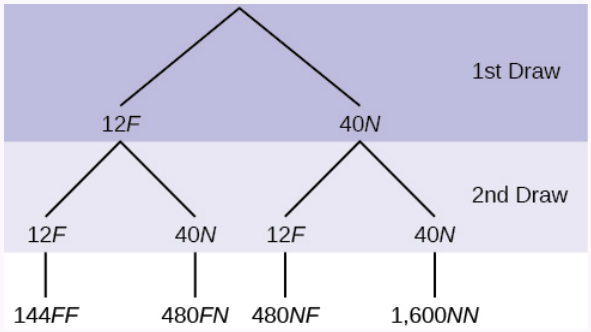


Figure 5.6.2:

Answer

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(\text{FF}) = \frac{144}{144+480+480+1,600} = \frac{144}{2,704} = \frac{9}{169} \tag{5.6.1}$$

1. a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...
2. b. Using the tree diagram, calculate *P*(*RR*).
3. c. Using the tree diagram, calculate $P(\text{RB OR BR})$.
4. d. Using the tree diagram, calculate $P(\text{R on 1st draw AND B on 2nd draw})$.
5. e. Using the tree diagram, calculate *P*(*R* on 2nd draw GIVEN *B* on 1st draw).
6. Using the tree diagram, calculate $P(\text{BB})$.
7. g. Using the tree diagram, calculate $P(\text{B on the 2nd draw given R on the first draw})$.

Solution

a. *B*1*R*1; *B*1*R*2; *B*1*R*3; *B*2*R*1; *B*2*R*2; *B*2*R*3; *B*3*R*1; *B*3*R*2; *B*3*R*3; *B*4*R*1; *B*4*R*2; *B*4*R*3; *B*5*R*1; *B*5*R*2; *B*5*R*3; *B*6*R*1; *B*6*R*2; *B*6*R*3 *B*7*R*1; *B*7*R*2; *B*7*R*3; *B*8*R*1; *B*8*R*2; *B*8*R*3

b. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. $P(\text{R on 1st draw AND B on 2nd draw}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. *P*(*R* on 2nd draw GIVEN *B* on 1st draw) = *P*(*R* on 2nd|*B* on 1st) = $\frac{24}{88} = \frac{3}{11}$ This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88} = \frac{3}{11}$

f. $P(\text{BB}) = \frac{64}{121}$

g. $P(\text{B on 2nd draw}|\text{R on 1st draw}) = \frac{8}{11}$. There are 9 + 24 outcomes that have R on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Example 5.6.2: 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, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

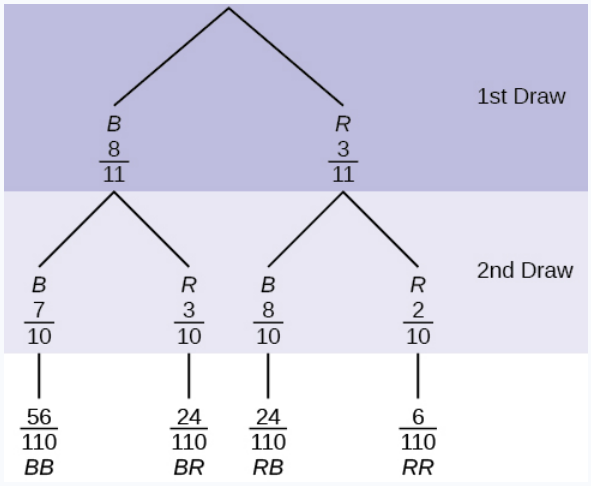


***Figure* 5.6.3:** *Total* $= \frac{56+24+24+6}{110} = \frac{110}{110} = 1$

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.

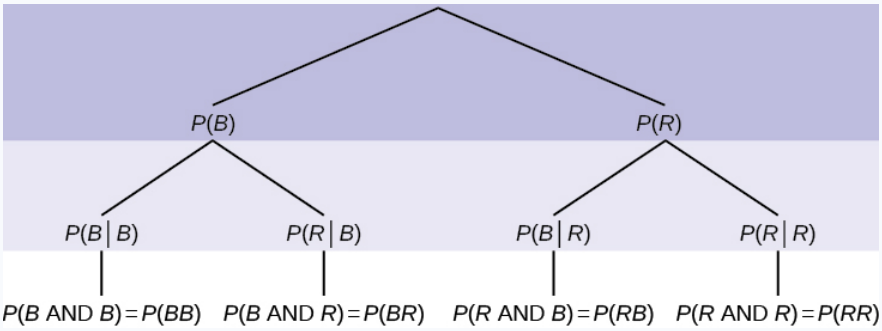
Calculate the following probabilities using the tree diagram.

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: $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = (\_\_\_)(\_\_\_) = \frac{24}{100}$
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. $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$
e. $P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$
f. Using the tree diagram, $P(\text{B on 2nd}|\text{R on 1st}) = P(\text{R}|\text{B}) = \frac{8}{10}$.

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



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

Exercise 5.6.2

In a standard deck, there are 52 cards. Twelve cards are face cards (**F**) and 40 cards are not face cards (**N**). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

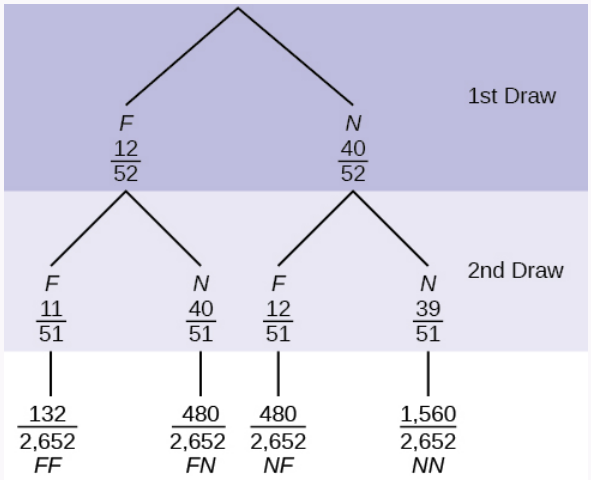


Figure 5.6.4*:*

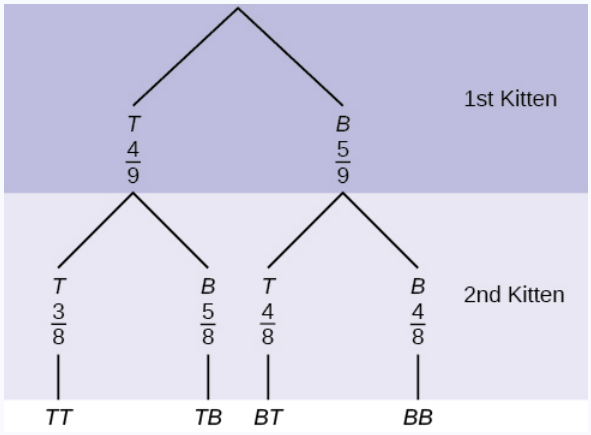
a. Find $P(\text{FN OR NF})$.
b. Find $P(\text{N}|\text{F})$.
c. Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
d. Find $P(\text{at least on face card})$.
Hint: "At least one face card" means one or two face cards.

Answer

a. $P(\text{FN OR NF}) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(\text{N}|\text{F}) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480+480+1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132+480+480)}{2,652} = \frac{1,092}{2,652}$

Example 5.6.3

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. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
b. What is the probability that one kitten of each coloring is selected?
a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$
c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Answer

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Exercise 5.6.3

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Answer

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right) + \left(\frac{3}{7}\right)\left(\frac{4}{6}\right)$

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 5.6.4

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:

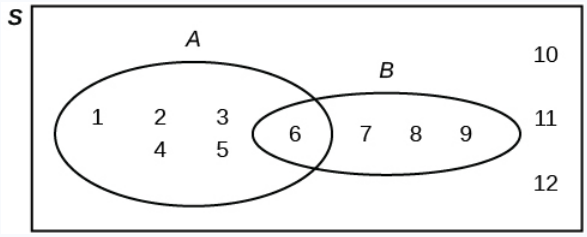


Figure* 5.6.5*:

Exercise 5.6.4

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event **C** = {green, blue, purple} and event **P** = {red, yellow, blue}. Then **C AND P** = {blue} and **C OR P** = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

Answer

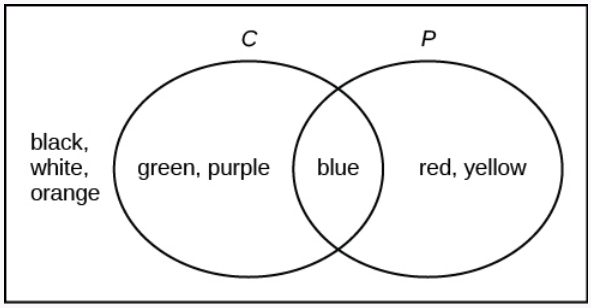


Figure* 5.6.6*:

Example 5.6.5

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:

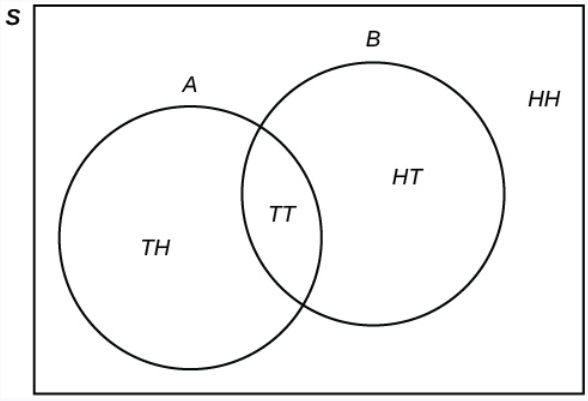
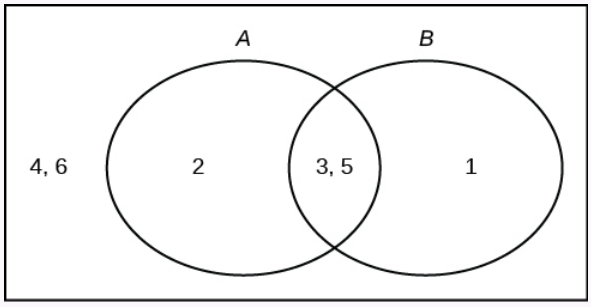


Figure* 5.6.7*:

Exercise 5.6.5

Roll a fair, six-sided die. Let $\mathbf{A}$ = a prime number of dots is rolled. Let $\mathbf{B}$ = an odd number of dots is rolled. Then $\mathbf{A} = \{2, 3, 5\}$ and $\mathbf{B} = \{1, 3, 5\}$. Therefore, $\mathbf{A}$ AND $\mathbf{B} = \{3, 5\}$. $\mathbf{A}$ OR $\mathbf{B} = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

Answer



Example 5.6.6: 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.

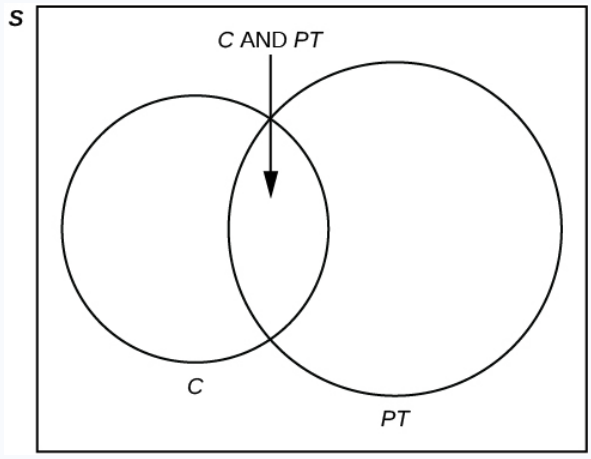


Figure* 5.6.8*:

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

Exercise 5.6.6

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let $\mathbf{W}$ = works a second job and $\mathbf{S}$ = spouse also works.

Answer

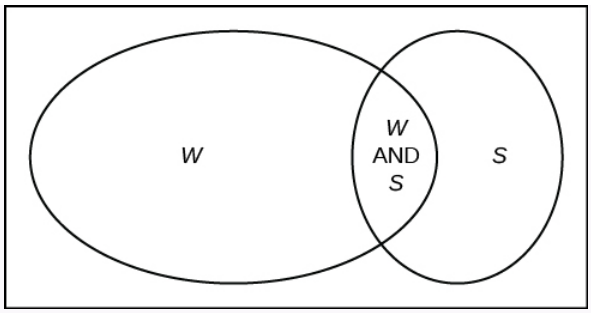


Figure* 5.6.9*:

Example 5.6.7

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.

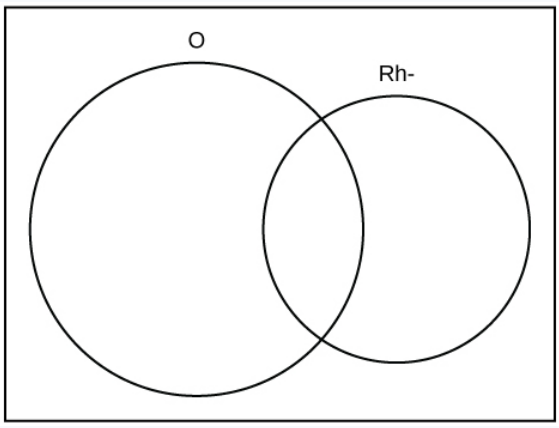


Figure* 5.6.10*:

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 **O =** African American with Type O blood and **R =** African American with Rh- factor.

a. $P(\text{O}) =$ _____
b. $P(\text{R}) =$ _____
c. $P(\text{O AND R}) =$ _____
d. $P(\text{O OR R}) =$ _____
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Answer

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Exercise 5.6.7

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.

c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Answer

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.

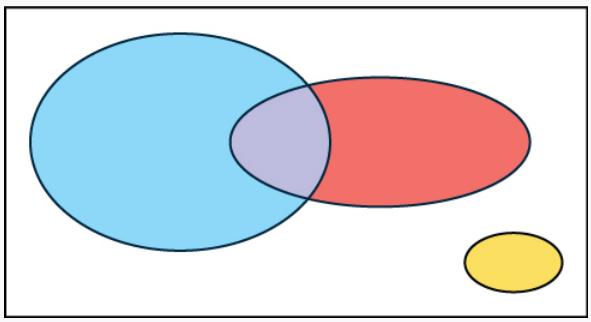


Figure* 5.6.11*:

b. $P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8$.

c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

1. Data from Clara County Public H.D.
2. Data from the American Cancer Society.
3. Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).
4. Data from the Federal Highway Administration, part of the United States Department of Transportation.
5. Data from the United States Census Bureau, part of the United States Department of Commerce.
6. Data from USA Today.
7. “Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).
8. “Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/dat..._datasets.html (accessed May 2, 2013).

Chapter 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 5.6: Tree and Venn Diagrams is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.7: Probability Topics (Worksheet)

This page titled 5.7: Probability Topics (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

5.8: Probability Topics (Exericses)

This page titled 5.8: Probability Topics (Exericses) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

6: The Normal Distribution

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ).

6.1: Prelude to The Normal Distribution

6.2: The Standard Normal Distribution

The Standard Normal Distribution (Exercises)

6.3: Using the Normal Distribution

6.4: Normal Distribution - Lap Times (Worksheet)

6.5: Normal Distribution - Pinkie Length (Worksheet)

6.6: The Normal Distribution (Exercises)

Contributors

- Template:ContribOpenStax

This page titled 6: The Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

6.1: Prelude to The Normal Distribution

Skills to Develop

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

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.



Figure* 6.1.1*: *If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)*

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.1.1}$$

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.

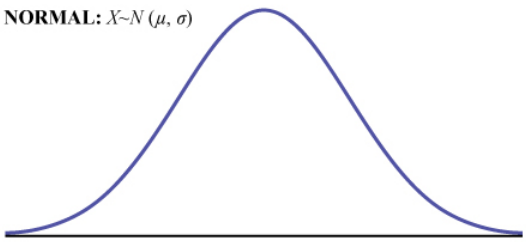


Figure* 6.1.2*: The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

COLLABORATIVE CLASSROOM ACTIVITY

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

- $X \sim N(\mu, \sigma)$
- $\mu =$ the mean $\sigma =$ the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x \cdot \mu)^2}{2\sigma^2}} \tag{6.1.2}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

Contributors

- Template:ContribOpenStax

This page titled 6.1: Prelude to The Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

6.2: The Standard Normal Distribution

Z-Scores

The standard normal distribution is a normal distribution of standardized values called *z-scores*. A z-score is measured in units of the standard deviation.

Definition: Z-Score

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:

$$z = \frac{x - \mu}{\sigma} \tag{6.2.1}$$

The *z*-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$\begin{aligned} x &= \mu + (z)(\sigma) \\ &= 5 + (3)(2) = 11 \end{aligned}$$

The *z*-score is three.

Since the mean for the standard normal distribution is zero and the standard deviation is one, then the transformation in Equation 6.2.1 produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

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

Example 6.2.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then (via Equation 6.2.1):

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

This means that $x = 17$ is **two** standard deviations (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Exercise 6.2.1

What is the z-score of x, when $x = 1$ and $X \sim N(12, 3)$?

Answer

$$z = \frac{1 - 12}{3} \approx -3.67$$

Example 6.2.2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $x = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. This z-score tells you that $x = 10$ is 2.5 standard deviations to the right of the mean five.

b. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$$

where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of their own) standard deviations to the right of their respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Exercise 6.2.2

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is -1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

Answer

1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the *Empirical Rule* says the following:

- About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and –1, respectively.
- The z-scores for +2σ and –2σ are +2 and –2, respectively.
- The z-scores for +3σ and –3σ are +3 and –3 respectively.

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

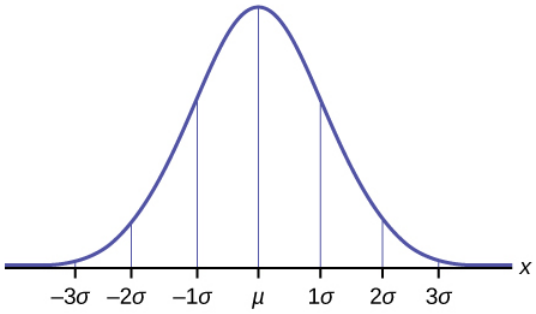


***Figure* 6.2.1**

Example 6.2.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when $x = 168$ cm is $z =$ _____. This z-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. –0.32, 0.32, left, 170
b. 177.98, 1.27, right

Exercise 6.2.3

Use the information in Example **6.2.3** to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z =$ _____. This z-score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = –2$. What is the male's height? The z-score ($z = –2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answer

Solve the equation $z = \frac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$z = \frac{176 - 170}{6.28}$, This z-score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.

Answer

Solve the equation $z = \frac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$X = 157.44$ cm, The z-score($z = –2$) tells you that the male's height is two standard deviations to the left of the mean.

Example 6.2.4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Answer

- The z-score (Equation 6.2.1) for $x = 160.58$ is $z = -1.5$.
- The z-score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Exercise 6.2.4

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let $X =$ a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Answer

The z-score (Equation 6.2.1) for $x_1 = 325$ is $z_1 = -1.14$.

The z-score (Equation 6.2.1) for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example 6.2.5

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between -1σ = (-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 of the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- 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 of the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50–18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Exercise 6.2.5

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Answer

between 20 and 30.

Example 6.2.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____ respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.
b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Exercise 6.2.6

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer a

About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.

Answer b

About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.

Answer c

About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

Summary

A z-score is a standardized value. Its distribution is the standard normal, \(Z \sim N(0, 1)\). The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

$Z \sim N(0, 1)$

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

mean = 0; standard deviation = 1

To find the 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

$Z \sim N(0, 1)$

Glossary

Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as \(Z \sim N(0, 1)\).

z-score

the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

References

1. "Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...reportid=11960 (accessed May 14, 2013).
2. "The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).
3. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at http://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 http://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 https://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

- Template:ContribOpenStax

This page titled 6.2: The Standard Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

The Standard Normal Distribution (Exercises)

Exercise 6.2.7

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ _____.

Answer

ounces of water in a bottle

Exercise 6.2.8

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise 6.2.9

$X \sim N(1, 2)$

$\sigma =$ _____

Answer

2

Exercise 6.2.10

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ _____.

Exercise 6.2.11

$X \sim N(-4, 1)$

What is the median?

Answer

–4

Exercise 6.2.12

$X \sim N(3, 5)$

$\sigma =$ _____

Exercise 6.2.13

$X \sim N(-2, 1)$

$\mu =$ _____

Answer

–2

Exercise 6.2.14

What does a z-score measure?

Exercise 6.2.15

What does standardizing a normal distribution do to the mean?

Answer

The mean becomes zero.

Exercise 6.2.16

Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

Exercise 6.2.17

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Answer

$z = 2$

Exercise 6.2.18

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Exercise 6.2.19

What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

Answer

$z = 2.78$

Exercise 6.2.20

What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

Exercise 6.2.21

Suppose $X \sim N(2, 6)$. What value of x has a z-score of three?

Answer

$x = 20$

Exercise 6.2.22

Suppose $X \sim N(8, 1)$. What value of x has a z-score of –2.25?

Exercise 6.2.23

Suppose $X \sim N(9, 5)$. What value of x has a z-score of –0.5?

Answer

$x = 6.5$

Exercise 6.2.24

Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

Exercise 6.2.25

Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

Answer

$x = 1$

Exercise 6.2.26

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Exercise 6.2.27

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Answer

$x = 1.97$

Exercise 6.2.28

Suppose $X \sim N(-1, 12)$. What is the z-score of $x = 2$?

Exercise 6.2.29

Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Answer

$z = -1.67$

Exercise 6.2.30

Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

Exercise 6.2.31

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Answer

$z \approx -0.33$

Exercise 6.2.32

In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.33

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

0.67, right

Exercise 6.2.34

In a normal distribution, $x = -2$ and $z = 6$. This tells you that $z = -2$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.35

In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

3.14, left

Exercise 6.2.36

In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.37

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Answer

about 68%

Exercise 6.2.38

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise 6.2.39

About what percent of x values lie between the second and third standard deviations (both sides)?

Answer

about 4%

Exercise 6.2.40

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise 6.2.41

Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution (i.e., –3).

Answer

between –5 and –1

Exercise 6.2.42

Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

Exercise 6.2.43

About what percent of x values lie between the mean and three standard deviations?

Answer

about 50%

Exercise 6.2.44

About what percent of x values lie between the mean and one standard deviation?

Exercise 6.2.45

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Answer

about 27%

Exercise 6.2.46

About what percent of x values lie between the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

Exercise 6.2.47

Define the random variable X in words. $X =$ _____.

Answer

The lifetime of a Sunshine CD player measured in years.

Exercise 6.2.48

$X \sim$ _____(_____,_____)

The Standard Normal Distribution (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

6.3: Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.

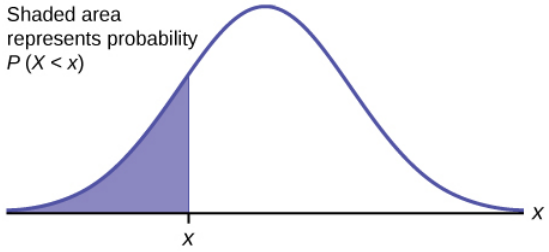


***Figure* 6.3.1.**

The area to the right is then $P(X > x) = 1–P(X < x)$. Remember, $P(X < x) =$ Area to the left of the vertical line through x. $P(X < x) = 1–P(X < x) =$ Area to the right of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.To calculate the probability, use the probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

Example 6.3.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Exercise 6.3.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer

$1 - 0.012 = 0.988$

Example 6.3.2

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.
c. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Answer

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

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

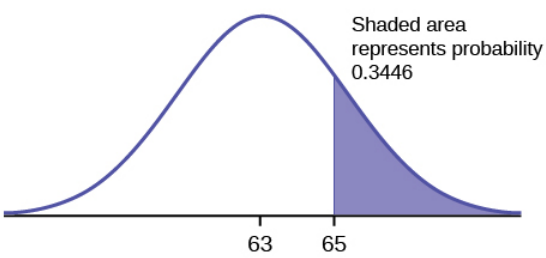


***Figure* 6.3.2.**

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

Go into `2nd DISTR` .

After pressing `2nd DISTR` , press `2:normalcdf` .

The syntax for the instructions are as follows:

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1` , the `EE` key (a 2nd key) and then `99` . Or, you can enter `10^ 99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Historical Note

The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65-63}{5} = 0.4$$

Area to the left is 0.6554.

$$P(x > 65) = P(z > 0.4) = 1-0.6554 = 0.3446$$

Calculate the z-score:

*Press `2nd Distr`

*Press `3:invNorm` (*Enter the area to the left of z followed by)
*Press `ENTER` .
For this Example, the steps are
`2nd Distr`
`3:invNorm` (.6554) `ENTER`

The answer is 0.3999 which rounds to 0.4.

Answer

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

$\text{normalcdf}(0, 85, 63, 5) = 1$ (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Answer

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let $k =$ the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a critical value.

$k = 69.4$

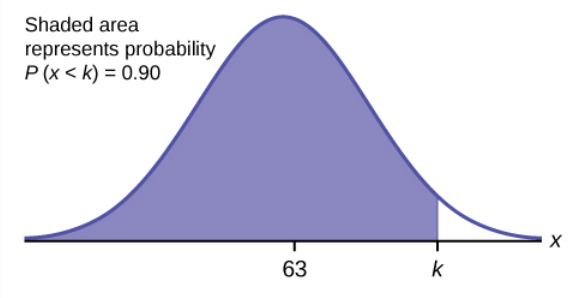


***Figure* 6.3.3.**

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

`invNorm` in `2nd DISTR` . invNorm(area to the left, mean, standard deviation)

For this problem, $\text{invNorm}(0.90, 63, 5) = 69.4$

Answer

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

Exercise 6.3.2

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a randomly selected golfer scored less than 65.

Answer

$\text{normalcdf}(10^{99}, 65, 68, 3) = 0.1587$

Example 6.3.3

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Answer

a. Let $X =$ the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

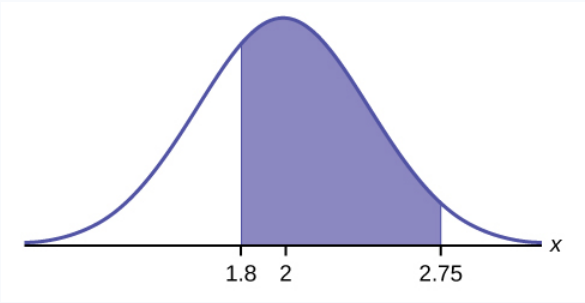


***Figure* 6.3.4.**

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

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b.

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile, k**, where $P(x < k) = 0.25$.

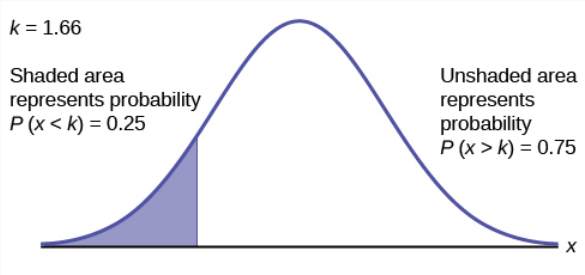


***Figure* 6.3.5.**

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

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

Exercise 6.3.3

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Answer

$$\text{normalcdf}(66, 70, 68, 3) = 0.4950$$

Example 6.3.4

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$

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.

Use the information in Example to answer the following questions.

Exercise 6.3.4

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old.

70.

Answer

Let $X =$ a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

To find the 30th percentile, find k such that $P(x < k) = 0.30$.
$\text{invNorm}(0.30, 36.9, 13.9) = 29.6$ years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years. Find $P(x < 27)$
(Note that $\text{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$. The two answers differ only by 0.0040.)

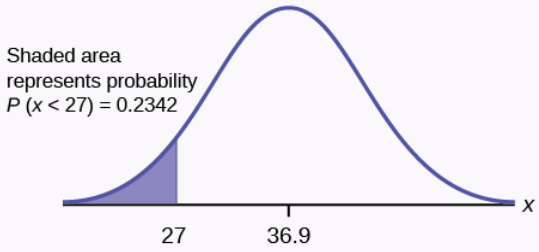


***Figure* 6.3.6.**

$\text{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$

Example 6.3.5

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. 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.

$$\text{invNorm}(0.75, 36.9, 13.9) = Q_3 = 46.2754$$

$$\text{invNorm}(0.25, 36.9, 13.9) = Q_1 = 27.5246$$

$$IQR = Q_3 - Q_1 = 18.7508$$

b.

Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")

$0.40 =$ the area to the right.

Area to the left $= 1–0.40 = 0.60$.

The area to the left of $k = 0.60$.

$\text{invNorm}(0.60, 36.9, 13.9) = 40.4215$.

$k = 40.42$.

Forty percent of the ages that range from 13 to 55+ are at least 40.42 years.

Exercise 6.3.5

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Answer

a. $Q_1 = 25^{\text{th}}$ percentile $= \text{invNorm}(0.25, 81, 15) = 70.9$
$Q_3 = 75^{\text{th}}$ percentile $= \text{invNorm}(0.75, 81, 15) = 91.9$
b. The middle 50% of the scores are between 70.9 and 91.1.

Example 6.3.6

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
c. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$

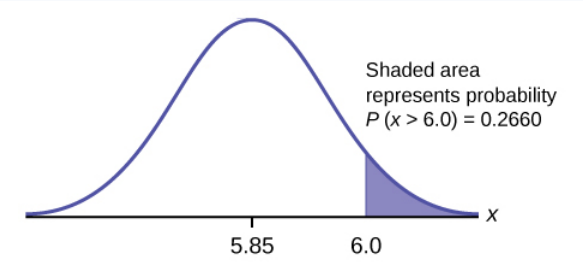


***Figure* 6.3.7.**

Answer

b.

$1–0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

Find $k1$, the 40th percentile, and $k2$, the 60th percentile ($0.40 + 0.20 = 0.60$).

$k1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79$ cm

$k2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91$ cm

Answer

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Exercise 6.3.6

Using the information from Example, answer the following:

a. The middle 45% of mandarin oranges from this farm are between _____ and _____.
b. Find the 16th percentile and interpret it in a complete sentence.

Answer a

The middle area $= 0.40$, so each tail has an area of 0.30.

$-0.40 = 0.60$

The tails of the graph of the normal distribution each have an area of 0.30.

Find $k1$, the 30th percentile and $k2$, the 70th percentile $(0.40 + 0.30 = 0.70)$.

$k1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72$ cm

$k2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98$ cm

Answer b

$\text{normalcdf}(5, 10^{99}, 5.85, 0.24) = 0.9998$

References

1. “Naegele’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).
2. “403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).
3. “Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at http://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).

Chapter Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

- Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.
- Standard Normal Distribution: $Z \sim N(0, 1)$.
- Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)
- Calculator function for the k^{th} percentile: $k = \text{invNorm}$ (area to the left of k, mean, standard deviation)

Exercise 6.3.7

How would you represent the area to the left of one in a probability statement?

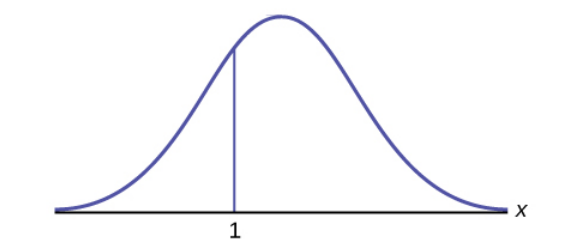


***Figure* 6.3.8.**

Answer

$P(x < 1)$

Exercise 6.3.8

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

Exercise 6.3.9

How would you represent the area to the left of three in a probability statement?

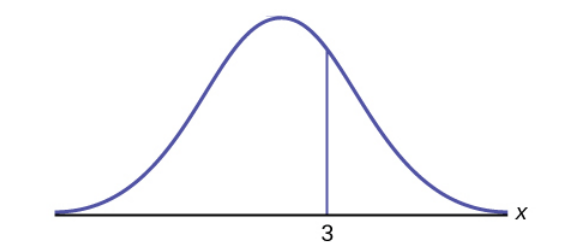


***Figure* 6.3.10.**

Exercise 6.3.10

What is the area to the right of three?

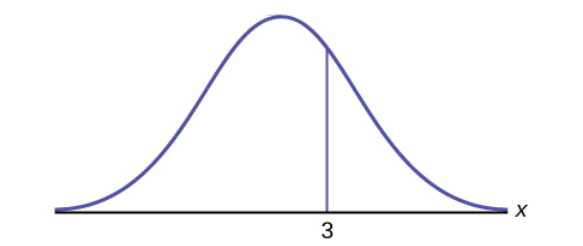


***Figure* 6.3.11.**

Answer

1–$P(x < 3)$ or $P(x > 3)$

Exercise 6.3.11

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise 6.3.12

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Answer

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise 6.3.13

Find the probability that $x > 56$.

Exercise 6.3.14

Find the probability that $x < 30$.

Answer

0.0013

Exercise 6.3.15

Find the 80th percentile.

Exercise 6.3.16

Find the 60th percentile.

Answer

56.03

Exercise 6.3.17

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise 6.3.18

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Answer

0.1186

Exercise 6.3.19

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Exercise 6.3.20

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

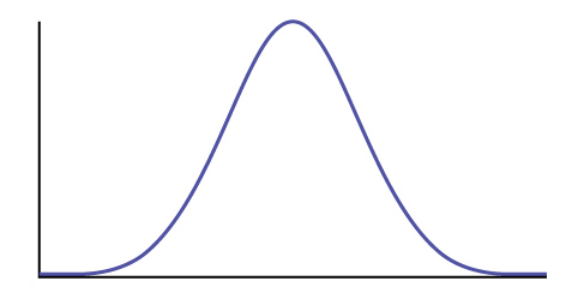


***Figure* 6.3.12.**

$P(0 < x <$ _____$) =$ _____ (Use zero for the minimum value of x.)

Answer

a. Check student's solution.
b. 3, 0.1979

Exercise 6.3.21

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

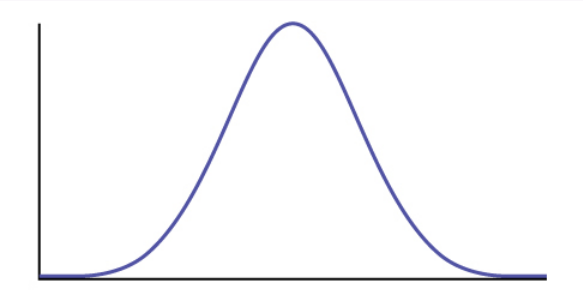


***Figure* 6.3.13.**

$P($_____ $< x <$ _____$) =$ _____

Exercise 6.3.22

Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.

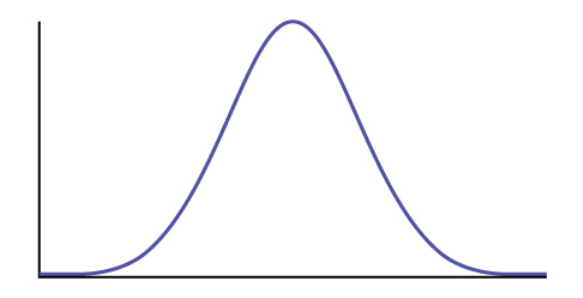


***Figure* 6.3.14.**

$P(x < k) =$ _____ Therefore, $k =$ _____

Answer

a. Check student's solution.

b. 0.70, 4.78 years

Contributors

- Template:ContribOpenStax

This page titled 6.3: Using the Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

6.4: Normal Distribution - Lap Times (Worksheet)

This page titled 6.4: Normal Distribution - Lap Times (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

6.5: Normal Distribution - Pinkie Length (Worksheet)

This page titled 6.5: Normal Distribution - Pinkie Length (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

6.6: The Normal Distribution (Exercises)

This page titled 6.6: The Normal Distribution (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

7: Sampling Distributions

Template:TopShafer

A statistic, such as the sample mean or the sample standard deviation, is a number computed from a sample. Since a sample is random, every statistic is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. As a random variable it has a mean, a standard deviation, and a probability distribution. The probability distribution of a statistic is called its sampling distribution. Typically sample statistics are not ends in themselves, but are computed in order to estimate the corresponding population parameters. This chapter introduces the concepts of the mean, the standard deviation, and the sampling distribution of a sample statistic, with an emphasis on the sample mean

7.1: The Mean and Standard Deviation of the Sample Mean

7.2: The Sampling Distribution of the Sample Mean

7.3: The Sample Proportion

7.4: Sampling Distributions (Exercises)

7: Sampling Distributions is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

7.1: The Mean and Standard Deviation of the Sample Mean

Skills to Develop

- To become familiar with the concept of the probability distribution of the sample mean.
- To understand the meaning of the formulas for the mean and standard deviation of the sample mean.

Suppose we wish to estimate the mean μ of a population. In actual practice we would typically take just one sample. Imagine however that we take sample after sample, all of the same size n, and compute the sample mean $\bar{x}$ each time. The sample mean x is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. We will write $\bar{X}$ when the sample mean is thought of as a random variable, and write x for the values that it takes. The random variable $\bar{X}$ has a mean, denoted $\mu_{\bar{X}}$, and a standard deviation, denoted $\sigma_{\bar{X}}$. Here is an example with such a small population and small sample size that we can actually write down every single sample.

Example 7.1.1

A rowing team consists of four rowers who weigh **152**, **156**, **160**, and **164** pounds. Find all possible random samples with replacement of size two and compute the sample mean for each one. Use them to find the probability distribution, the mean, and the standard deviation of the sample mean $\bar{X}$.

Solution:

The following table shows all possible samples with replacement of size two, along with the mean of each:

Sample	Mean		Sample	Mean		Sample	Mean		Sample	Mean
152, 152	152		156, 152	154		160, 152	156		164, 152	158
152, 156	154		156, 156	156		160, 156	158		164, 156	160
152, 160	156		156, 160	158		160, 160	160		164, 160	162
152, 164	158		156, 164	160		160, 164	162		164, 164	164

The table shows that there are seven possible values of the sample mean $\bar{X}$. The value $\bar{x} = 152$ happens only one way (the rower weighing **152** pounds must be selected both times), as does the value $\bar{x} = 164$, but the other values happen more than one way, hence are more likely to be observed than **152** and **164** are. Since the **16** samples are equally likely, we obtain the probability distribution of the sample mean just by counting:

$$\begin{array}{c|ccccccc} \bar{x} & 152 & 154 & 156 & 158 & 160 & 162 & 164 \\ \hline P(\bar{x}) & \frac{1}{16} & \frac{2}{16} & \frac{3}{16} & \frac{4}{16} & \frac{3}{16} & \frac{2}{16} & \frac{1}{16} \end{array} \tag{7.1.1}$$

Now we apply the formulas from Section 4.2 to $\bar{X}$. For $\mu_{\bar{X}}$, we obtain.

$$\mu_{\bar{X}} = \sum \bar{x} P(\bar{x}) \tag{7.1.2}$$

$$= 152(116) + 154(216) + 156(316) + 158(416) + 160(316) + 162(216) + 164(116) = 158 \tag{7.1.3}$$

For $\sigma_{\bar{X}}$, we first compute $\sum \bar{x}^2 P(\bar{x})$:

$$152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 164^2\left(\frac{1}{16}\right) \tag{7.1.4}$$

which is $24,974$, so that

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - \mu_{\bar{x}}^2} = \sqrt{24,974 - 158^2} = \sqrt{10} \tag{7.1.5}$$

The mean and standard deviation of the population $\{152, 156, 160, 164\}$ in the example are $\mu = 158$ and $\sigma = \sqrt{20}$. The mean of the sample mean $\bar{X}$ that we have just computed is exactly the mean of the population. The standard deviation of the sample mean $\bar{X}$

that we have just computed is the standard deviation of the population divided by the square root of the sample size: $\sqrt{10} = \sqrt{20}/\sqrt{2}$. These relationships are not coincidences, but are illustrations of the following formulas.

Definition: Sample mean and sample standard deviation

Suppose random samples of size n are drawn from a population with mean μ and standard deviation σ. The mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$ satisfy

$$\mu_{\bar{X}} = \mu \tag{7.1.6}$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.1.7}$$

Equation 7.1.6 says that if we could take every possible sample from the population and compute the corresponding sample mean, then those numbers would center at the number we wish to estimate, the population mean μ. Equation 7.1.7 says that averages computed from samples vary less than individual measurements on the population do, and quantifies the relationship.

Example 7.1.2

The mean and standard deviation of the tax value of all vehicles registered in a certain state are $\mu = \$13,525$ and $\sigma = \$4,180$. Suppose random samples of size 100 are drawn from the population of vehicles. What are the mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$?

Solution

Since $n = 100$, the formulas yield

$$\mu_{\bar{X}} = \mu = \$13,525 \tag{7.1.8}$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418 \tag{7.1.9}$$

Key Takeaway

- The sample mean is a random variable; as such it is written $\bar{X}$, and $\bar{x}$ stands for individual values it takes.
- As a random variable the sample mean has a probability distribution, a mean $\mu_{\bar{X}}$, and a standard deviation $\sigma_{\bar{X}}$.
- There are formulas that relate the mean and standard deviation of the sample mean to the mean and standard deviation of the population from which the sample is drawn.

7.1: The Mean and Standard Deviation of the Sample Mean is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

7.2: The Sampling Distribution of the Sample Mean

skills to develop

- To learn what the sampling distribution of $\overline{X}$ is when the sample size is large.
- To learn what the sampling distribution of $\overline{X}$ is when the population is normal.

In Example 6.1.1, we constructed the probability distribution of the sample mean for samples of size two drawn from the population of four rowers. The probability distribution is:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

(7.2.1)

Figure 7.2.1 shows a side-by-side comparison of a histogram for the original population and a histogram for this distribution. Whereas the distribution of the population is uniform, the sampling distribution of the mean has a shape approaching the shape of the familiar bell curve. This phenomenon of the sampling distribution of the mean taking on a bell shape even though the population distribution is not bell-shaped happens in general. Here is a somewhat more realistic example.

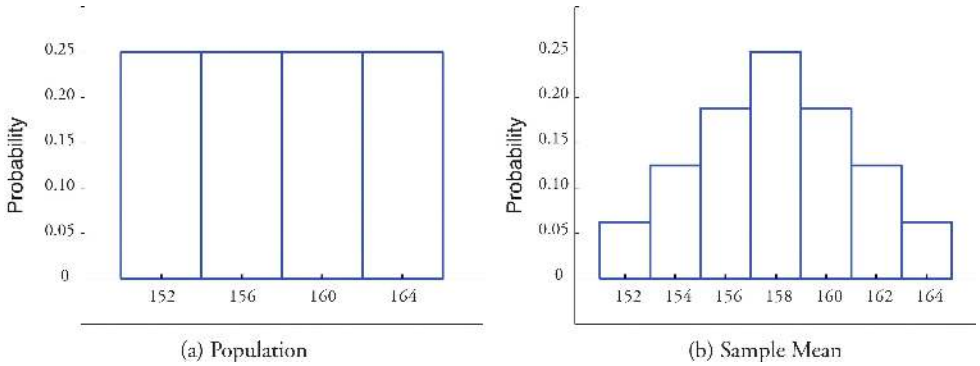


***Figure* 7.2.1**: *Distribution of a Population and a Sample Mean*

Suppose we take samples of size 1, 5, 10, or 20 from a population that consists entirely of the numbers 0 and 1, half the population 0, half 1, so that the population mean is 0.5. The sampling distributions are:

$n = 1$:

$\bar{x}$	0	1
$P(\bar{x})$	0.5	0.5

(7.2.2)

$n = 5$:

$\bar{x}$	0	0.2	0.4	0.6	0.8	1
$P(\bar{x})$	0.03	0.16	0.31	0.31	0.16	0.03

(7.2.3)

$n = 10$:

$\bar{x}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(\bar{x})$	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00

(7.2.4)

$n = 20$:

$\bar{x}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
$P(\bar{x})$	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.07	0.12	0.16	0.18

(7.2.5)

$\bar{x}$	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1
$P(\bar{x})$	0.16	0.12	0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00

(7.2.6)

Histograms illustrating these distributions are shown in Figure 7.2.2.

$n = 1$

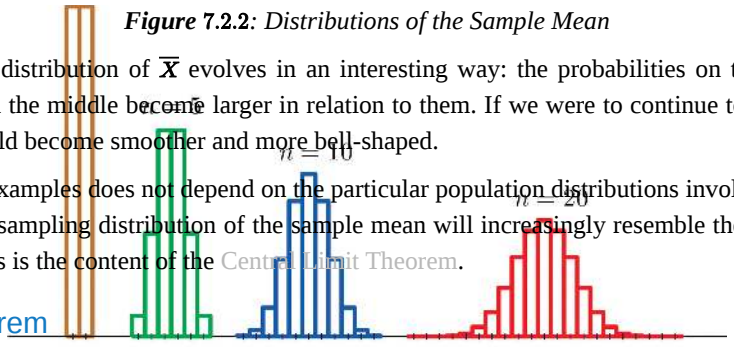


***Figure* 7.2.2**: *Distributions of the Sample Mean*

As n increases the sampling distribution of $\overline{X}$ evolves in an interesting way: the probabilities on the lower and the upper ends shrink and the probabilities in the middle become larger in relation to them. If we were to continue to increase n then the shape of the sampling distribution would become smoother and more bell-shaped.

What we are seeing in these examples does not depend on the particular population distributions involved. In general, one may start with any distribution and the sampling distribution of the sample mean will increasingly resemble the bell-shaped normal curve as the sample size increases. This is the content of the Central Limit Theorem.

The Central Limit Theorem

For samples of size 30 or more, the sample mean is approximately normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, where n is the sample size. The larger the sample size, the better the approximation. The Central Limit Theorem is illustrated for several common population distributions in Figure 7.2.3.

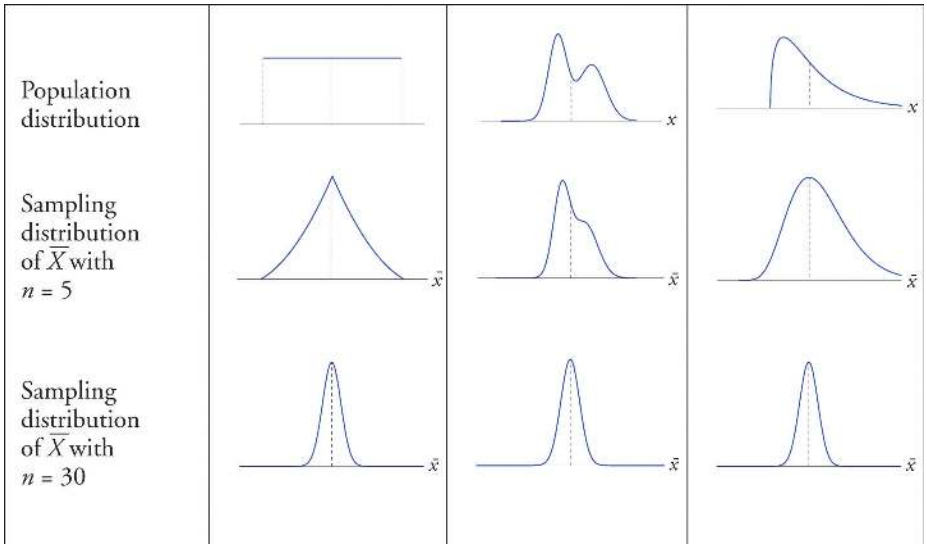


***Figure* 7.2.3**: *Distribution of Populations and Sample Means*

The dashed vertical lines in the figures locate the population mean. Regardless of the distribution of the population, as the sample size is increased the shape of the sampling distribution of the sample mean becomes increasingly bell-shaped, centered on the population mean. Typically by the time the sample size is 30 the distribution of the sample mean is practically the same as a normal distribution.

The importance of the Central Limit Theorem is that it allows us to make probability statements about the sample mean, specifically in relation to its value in comparison to the population mean, as we will see in the examples. But to use the result properly we must first realize that there are two separate random variables (and therefore two probability distributions) at play:

1. X, the measurement of a single element selected at random from the population; the distribution of X is the distribution of the population, with mean the population mean μ and standard deviation the population standard deviation σ;
2. $\overline{X}$, the mean of the measurements in a sample of size n; the distribution of $\overline{X}$ is its sampling distribution, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Example 7.2.1

Let $\overline{X}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40.

1. Find the mean and standard deviation of $\overline{X}$.
2. Find the probability that $\overline{X}$ assumes a value between 110 and 114.
3. Find the probability that $\overline{X}$ assumes a value greater than 113.

Solution:

1. By the formulas in the previous section

$$\mu_{\overline{X}} = \mu = 112 \text{ and } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685 \tag{7.2.7}$$

2. Since the sample size is at least 30, the Central Limit Theorem applies: $\overline{X}$ is approximately normally distributed. We compute probabilities using Figure 5.3.1 in the usual way, just being careful to use $\sigma_{\overline{X}}$ and not σ when we standardize:

$$\begin{aligned} P(110 < \overline{X} < 114) &= P\left(\frac{110 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{114 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{110 - 112}{5.65685} < Z < \frac{114 - 112}{5.65685}\right) \\ &= P(-0.35 < Z < 0.35) \\ &= 0.6368 - 0.3632 \\ &= 0.2736 \end{aligned}$$

3. Similarly

$$\begin{aligned} P(\overline{X} > 113) &= P\left(Z > \frac{113 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z > \frac{113 - 112}{5.65685}\right) \\ &= P(Z > 0.18) \\ &= 1 - P(Z < 0.18) \\ &= 1 - 0.5714 \\ &= 0.4286 \end{aligned}$$

Note that if inthe above example we had been asked to compute the probability that the value of a single randomly selected element of the population exceeds 113, that is, to compute the number $P(X > 113)$, we would not have been able to do so, since we do not know the distribution of X, but only that its mean is 112 and its standard deviation is 40. By contrast we could compute $P(\overline{X} > 113)$ even without complete knowledge of the distribution of X because the Central Limit Theorem guarantees that $\overline{X}$ is approximately normal.

Example 7.2.2

The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Solution:

The sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 2.61$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$, so

$$\begin{aligned} P(2.51 < \overline{X} < 2.71) &= P\left(\frac{2.51 - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < Z < \frac{2.71 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(\frac{2.51 - 2.61}{0.05} < Z < \frac{2.71 - 2.61}{0.05}\right) \\ &= P(-2 < Z < 2) \\ &= P(Z < 2) - P(Z < -2) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \end{aligned}$$

Normally Distributed Populations

The Central Limit Theorem says that no matter what the distribution of the population is, as long as the sample is "large," meaning of size 30 or more, the sample mean is approximately normally distributed. If the population is normal to begin with then the sample mean also has a normal distribution, regardless of the sample size.

For samples of any size drawn from a normally distributed population, the sample mean is normally distributed, with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, where n is the sample size.

The effect of increasing the sample size is shown in Figure $7.2.4$.

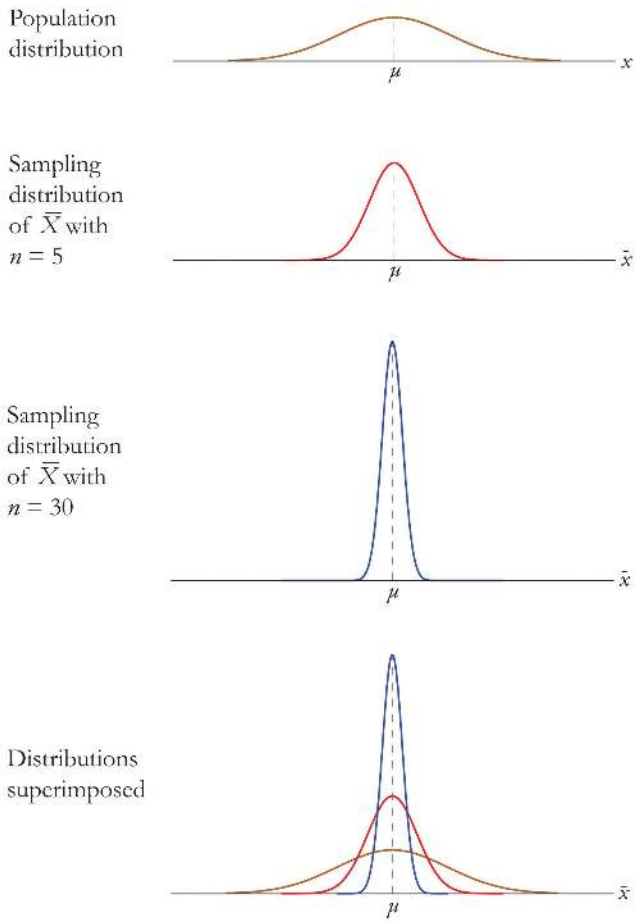


***Figure* 7.2.4**: *Distribution of Sample Means for a Normal Population*

Example 7.2.3

A prototype automotive tire has a design life of $38,500$ miles with a standard deviation of $2,500$ miles. Five such tires are manufactured and tested. On the assumption that the actual population mean is $38,500$ miles and the actual population standard deviation is $2,500$ miles, find the probability that the sample mean will be less than $36,000$ miles. Assume that the distribution of lifetimes of such tires is normal.

Solution:

For simplicity we use units of thousands of miles. Then the sample mean $\overline{X}$ has mean $\mu_{\overline{X}} = \mu = 38.5$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{5}} = 1.11803$. Since the population is normally distributed, so is $\overline{X}$, hence

$$\begin{aligned} P(\overline{X} < 36) &= P\left(Z < \frac{36 - \mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{36 - 38.5}{1.11803}\right) \\ &= P(Z < -2.24) \\ &= 0.0125 \end{aligned}$$

That is, if the tires perform as designed, there is only about a 1.25% chance that the average of a sample of this size would be so low.

Example 7.2.4

An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal.

a. On the assumption that the manufacturer's claims are true, find the probability that a randomly selected battery of this type will last less than 48 months.
b. On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Solution:

a. Since the population is known to have a normal distribution

$$\begin{aligned} P(X < 48) &= P\left(Z < \frac{48-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{48-50}{6}\right) \\ &= P(Z < -0.33) \\ &= 0.3707 \end{aligned}$$

b. The sample mean has mean $\mu_{\overline{X}} = \mu = 50$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$. Thus

$$\begin{aligned} P(\overline{X} < 48) &= P\left(Z < \frac{48-\mu_{\overline{X}}}{\sigma_{\overline{X}}}\right) \\ &= P\left(Z < \frac{48-50}{1}\right) \\ &= P(Z < -2) \\ &= 0.0228 \end{aligned}$$

Key Takeaway

- When the sample size is at least 30 the sample mean is normally distributed.
- When the population is normal the sample mean is normally distributed regardless of the sample size.

7.2: The Sampling Distribution of the Sample Mean is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

7.3: The Sample Proportion

Skills to Develop

- To recognize that the sample proportion $\hat{p}$ is a random variable.
- To understand the meaning of the formulas for the mean and standard deviation of the sample proportion.
- To learn what the sampling distribution of $\hat{p}$ is when the sample size is large.

Often sampling is done in order to estimate the proportion of a population that has a specific characteristic, such as the proportion of all items coming off an assembly line that are defective or the proportion of all people entering a retail store who make a purchase before leaving. The population proportion is denoted p and the sample proportion is denoted $\hat{p}$. Thus if in reality 43% of people entering a store make a purchase before leaving,

$$p = 0.43$$

if in a sample of 200 people entering the store, 78 make a purchase,

$$\hat{p} = \frac{78}{200} = 0.39.$$

The sample proportion is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. Viewed as a random variable it will be written $\hat{P}$. It has a mean $\mu_{\hat{P}}$ and a standard deviation $\sigma_{\hat{P}}$. Here are formulas for their values.

mean and standard deviation of the sample proportion

Suppose random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p. The mean $\mu_{\hat{P}}$ and standard deviation $\sigma_{\hat{P}}$ of the sample proportion $\hat{P}$ satisfy

$$\mu_{\hat{P}} = p \tag{7.3.1}$$

and

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} \tag{7.3.2}$$

where $q = 1 - p$.

The Central Limit Theorem has an analogue for the population proportion $\hat{p}$. To see how, imagine that every element of the population that has the characteristic of interest is labeled with a 1, and that every element that does not is labeled with a 0. This gives a numerical population consisting entirely of zeros and ones. Clearly the proportion of the population with the special characteristic is the proportion of the numerical population that are ones; in symbols,

$$p = \frac{\text{number of 1s}}{N} \tag{7.3.3}$$

But of course the sum of all the zeros and ones is simply the number of ones, so the mean μ of the numerical population is

$$\mu = \frac{\sum x}{N} = \frac{\text{number of 1s}}{N} \tag{7.3.4}$$

Thus the population proportion p is the same as the mean μ of the corresponding population of zeros and ones. In the same way the sample proportion $\hat{p}$ is the same as the sample mean $\bar{x}$. Thus the Central Limit Theorem applies to $\hat{p}$. However, the condition that the sample be large is a little more complicated than just being of size at least 30.

The Sampling Distribution of the Sample Proportion

For large samples, the sample proportion is approximately normally distributed, with mean $\mu_{\hat{P}} = p$ and standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

A sample is large if the interval $[p - 3\sigma_{\hat{p}},\ p + 3\sigma_{\hat{p}}]$ lies wholly within the interval $[0, 1]$.

In actual practice p is not known, hence neither is $\sigma_{\hat{p}}$. In that case in order to check that the sample is sufficiently large we substitute the known quantity $\hat{p}$ for p. This means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{7.3.5}$$

lies wholly within the interval $[0,1]$. This is illustrated in the examples.

Figure 7.3.1 shows that when $p=0.1$, a sample of size 15 is too small but a sample of size 100 is acceptable.

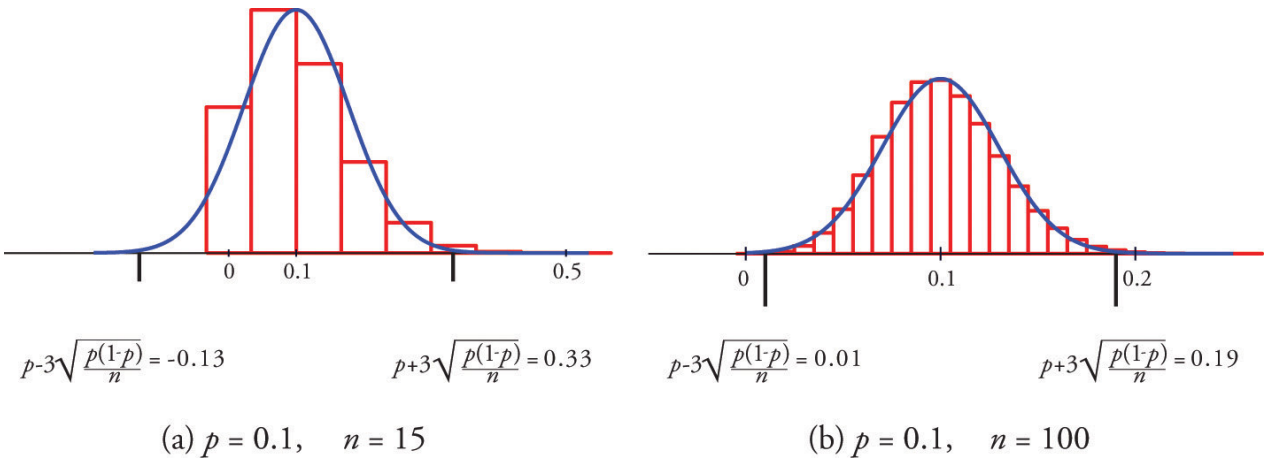


***Figure* 7.3.1**: *Distribution of Sample Proportions*

Figure 7.3.2 shows that when $p=0.5$ a sample of size 15 is acceptable.

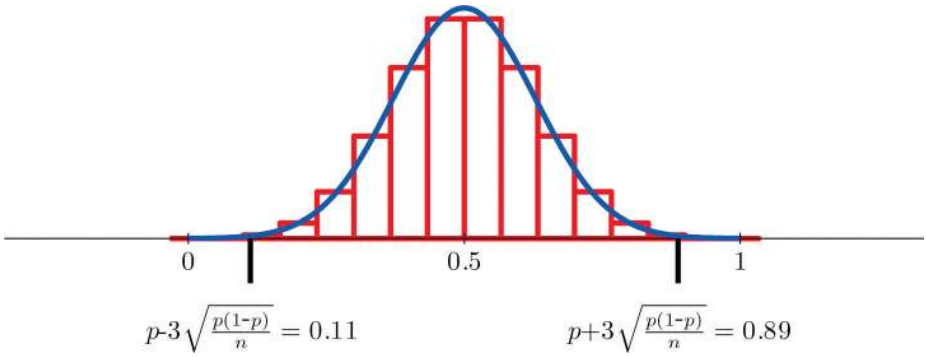


***Figure* 7.3.2**: *Distribution of Sample Proportions for* $p=0.5$ *and* $n=15$

Example 7.3.1

Suppose that in a population of voters in a certain region 38% are in favor of particular bond issue. Nine hundred randomly selected voters are asked if they favor the bond issue.

1. Verify that the sample proportion $\hat{p}$ computed from samples of size 900 meets the condition that its sampling distribution be approximately normal.
2. Find the probability that the sample proportion computed from a sample of size 900 will be within 5 percentage points of the true population proportion.

Solution:

1. The information given is that $p=0.38$, hence $q=1-p=0.62$. First we use the formulas to compute the mean and standard deviation of $\hat{p}$:

$$\mu_{\hat{p}}=p=0.38 \text{ and } \sigma_{\hat{p}}=\sqrt{\frac{pq}{n}}=\sqrt{\frac{(0.38)(0.62)}{900}}=0.01618$$

Then $3\sigma_{\hat{P}} = 3(0.01618) = 0.04854 \approx 0.05$ so

\[\left [\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\, \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]=[0.38-0.05,0.38+0.05]=[0.33,0.43] \nonumber\]

which lies wholly within the interval $[0, 1]$, so it is safe to assume that $\hat{p}$ is approximately normally distributed.

2. To be within 5 percentage points of the true population proportion 0.38 means to be between $0.38 - 0.05 = 0.33$ and $0.38 + 0.05 = 0.43$. Thus

$$\begin{aligned} P(0.33 < \hat{P} < 0.43) &= P\left(\frac{0.33 - \mu_{\hat{P}}}{\sigma_{\hat{P}}} < Z < \frac{0.43 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(\frac{0.33 - 0.38}{0.01618} < Z < \frac{0.43 - 0.38}{0.01618}\right) \\ &= P(-3.09 < Z < 3.09) \\ &= P(3.09) - P(-3.09) \\ &= 0.9990 - 0.0010 \\ &= 0.9980 \end{aligned}$$

Example 7.3.2

An online retailer claims that 90% of all orders are shipped within 12 hours of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within 12 hours.

1. Compute the sample proportion of items shipped within 12 hours.
2. Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use $p = 0.90$, corresponding to the assumption that the retailer's claim is valid.
3. Assuming the retailer's claim is true, find the probability that a sample of size 121 would produce a sample proportion so low as was observed in this sample.
4. Based on the answer to part (c), draw a conclusion about the retailer's claim.

Solution:

1. The sample proportion is the number x of orders that are shipped within 12 hours divided by the number n of orders in the sample:

$$\hat{p} = \frac{x}{n} = \frac{102}{121} = 0.84$$

2. Since $p = 0.90$, $q = 1 - p = 0.10$, and $n = 121$,

$$\sigma_{\hat{P}} = \sqrt{\frac{(0.90)(0.10)}{121}} = 0.02\overline{7}$$

hence

$$[p - 3\sigma_{\hat{P}},\, p + 3\sigma_{\hat{P}}] = [0.90 - 0.08, 0.90 + 0.08] = [0.82, 0.98]$$

Because

$$[0.82, 0.98] \subset [0, 1]$$

it is appropriate to use the normal distribution to compute probabilities related to the sample proportion $\hat{P}$.

3. Using the value of $\hat{P}$ from part (a) and the computation in part (b),

$$
\begin{aligned}
P(\hat{P} \leq 0.84) &= P\left(Z \leq \frac{0.84 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\
&= P\left(Z \leq \frac{0.84 - 0.90}{0.0\overline{27}}\right) \\
&= P(Z \leq -2.20) \\
&= 0.0139
\end{aligned}
$$

4. The computation shows that a random sample of size 121 has only about a 1.4% chance of producing a sample proportion as the one that was observed, $\hat{p} = 0.84$, when taken from a population in which the actual proportion is 0.90. This is so unlikely that it is reasonable to conclude that the actual value of p is less than the 90% claimed.

Key Takeaway

- The sample proportion is a random variable $\hat{P}$.
- There are formulas for the mean $\mu_{\hat{P}}$, and standard deviation $\sigma_{\hat{P}}$ of the sample proportion.
- When the sample size is large the sample proportion is normally distributed.

7.3: The Sample Proportion is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

7.4: Sampling Distributions (Exercises)

7.4: Sampling Distributions (Exercises) is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

8: Power

Template:TopHeblLane

Power is the probability of correctly rejecting a false null hypothesis.

Topic hierarchy

8.1: Introduction to Power
8.2: Example Calculations
8.3: Power Demo
8.4: Power Demo II
8.5: Factors Affecting Power
8.6: Statistical Literacy
8.7: Power (Exercises)

Contributors

- Template:ContribHeblLane

This page titled 8: Power is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.1: Introduction to Power

Skills to Develop

- Define power
- Identify situations in which it is important to estimate power

Suppose you work for a foundation whose mission is to support researchers in mathematics education and your role is to evaluate grant proposals and decide which ones to fund. You receive a proposal to evaluate a new method of teaching high-school algebra. The research plan is to compare the achievement of students taught by the new method with the achievement of students taught by the traditional method. The proposal contains good theoretical arguments why the new method should be superior and the proposed methodology is sound. In addition to these positive elements, there is one important question still to be answered: Does the experiment have a high probability of providing strong evidence that the new method is better than the standard method if, in fact, the new method is actually better? It is possible, for example, that the proposed sample size is so small that even a fairly large population difference would be difficult to detect. That is, if the sample size is small, then even a fairly large difference in sample means might not be significant. If the difference is not significant, then no strong conclusions can be drawn about the population means. It is not justified to conclude that the null hypothesis that the population means are equal is true just because the difference is not significant. Of course, it is not justified to conclude that this null hypothesis is false. Therefore, when an effect is not significant, the result is inconclusive. You may prefer that your foundation's money be used to fund a project that has a higher probability of being able to make a strong conclusion.

Power is defined as the probability of correctly rejecting a false null hypothesis. In terms of our example, it is the probability that given there is a difference between the population means of the new method and the standard method, the sample means will be significantly different. The probability of failing to reject a false null hypothesis is often referred to as β. Therefore power can be defined as:

$$\text{power} = 1 - \beta \tag{8.1.1}$$

It is very important to consider power while designing an experiment. You should avoid spending a lot of time and/or money on an experiment that has little chance of finding a significant effect.

Contributor

- Template:ContribHeblLane

This page titled 8.1: Introduction to Power is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.2: Example Calculations

skills to develop

- Compute power using the binomial distribution
- Compute power using the normal distribution
- Use a power calculator to compute power for the t distribution

In the "Shaking and Stirring Martinis" case study, the question was whether Mr. Bond could tell the difference between martinis that were stirred and martinis that were shaken. For the sake of this example, assume he can tell the difference and is able to correctly state whether a martini had been shaken or stirred 0.75 of the time. Now, suppose an experiment is being conducted to investigate whether Mr. Bond can tell the difference. Specifically, is Mr. Bond correct more than 0.50 of the time? We know that he is (that's an assumption of the example). However, the experimenter does not know and asks Mr. Bond to judge 16 martinis. The experimenter will do a significance test based on the binomial distribution. Specifically, if a one tailed test is significant at the 0.05 level, then he or she will conclude that Mr. Bond can tell the difference. The probability value is computed assuming the null hypothesis is true ($\pi = 0.50$). Therefore, the experimenter will determine how many times Mr. Bond is correct, and compute the probability of being correct that many or more times given that the null hypothesis is true. The question is: what is the probability the experimenter will correctly reject the null hypothesis that $\pi = 0.50$? In other words, what is the power of this experiment?

The binomial distribution for $N = 16$ and $\pi = 0.50$ is shown in Figure $8.2.1$. The probability of being correct on 11 or more trials is 0.105 and the probability of being correct on 12 or more trials is 0.038. Therefore, the probability of being correct on 12 or more trials is less than 0.05. This means that the null hypothesis will be rejected if Mr. Bond is correct on 12 or more trials and will not be rejected otherwise.

Binomial Calculator

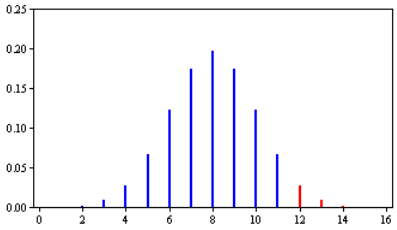


Figure $8.2.1$***:*** *The binomial distribution for* $N = 16$ *and* $\pi = 0.50$

We know that Mr. Bond is correct 0.75 of the time. (Obviously the experimenter does not know this or there would be no need for an experiment.) The binomial distribution with $N = 16$ and $\pi = 0.75$ is shown in Figure $8.2.2$.

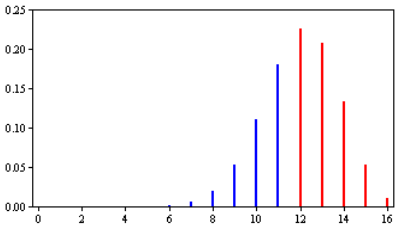


Figure $8.2.2$***:*** *The binomial distribution for* $N = 16$ *and* $\pi = 0.75$

The probability of being correct on 12 or more trials is 0.63. Therefore, the power of the experiment is 0.63.

To sum up, the probability of being correct on 12 or more trials given that the null hypothesis is true is less than 0.05. Therefore, if Mr. Bond is correct on 12 or more trials, the null hypothesis will be rejected. Given Mr. Bond's true ability to be correct on 0.75 of the trials, the probability he will be correct on 12 or more trials is 0.63. Therefore power is 0.63.

In the section on Testing a Single Mean for significance, the first example was based on the assumption that the experimenter knew the population variance. Although this is rarely true in practice, the example is very useful for pedagogical purposes. For the same

reason, the following example assumes the experimenter knows the population variance. Power calculators are available for situations in which the experimenter does not know the population variance.

Suppose a math achievement test were known to have a mean of 75 and a standard deviation of 10. A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean for the new method is 80. The researcher plans to sample 25 subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. What is the probability that the researcher will correctly reject the false null hypothesis that the population mean for the new method is 75 or lower? The following shows how this probability is computed.

The researcher assumes that the population standard deviation with the new method is the same as with the old method (10) and that the distribution is normal. Since the population standard deviation is assumed to be known, the researcher can use the normal distribution rather than the t distribution to compute the p value. Recall that the standard error of the mean (σ_M) is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{8.2.1}$$

which is equal to $10/5 = 2$ in this example. As can be seen in Figure 8.2.3, if the null hypothesis that the population mean equals 75 is true, then the probability of a sample mean being greater than or equal to 78.29 is 0.05. Therefore, the experimenter will reject the null hypothesis if the sample mean, M, is 78.29 or larger.

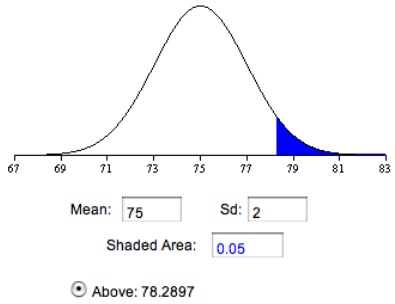


***Figure** 8.2.3: The sampling distribution of the mean if the null hypothesis is true. (Figure created with the Inverse Normal Calculator)*

The question, then, is what is the probability the experimenter gets a sample mean greater than 78.29 given that the population mean is 80? Figure 8.2.4shows that this probability is 0.80.

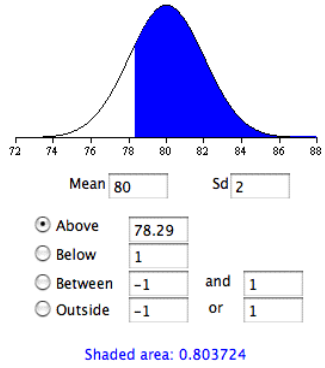


***Figure** 8.2.4: The sampling distribution of the mean if the population mean is 80. The test is significant if the sample mean is 78.29 or higher. (Figure created with the Normal Calculator)*

Therefore, the probability that the experimenter will reject the null hypothesis that the population mean of the new method is 75 is 0.80. In other words, $power = 0.80$.

Calculation of power is more complex for t tests and for Analysis of Variance. The power calculator computes power for a t test of independent groups. Calculators for other types of designs are linked to below:

Contributor

- Template:ContribHeblLane

This page titled 8.2: Example Calculations is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.3: Power Demo

Skills to Develop

- State the effect of effect size on power
- State the effect of one- versus two-tailed tests on power
- State the effect of the standard deviation on power
- State the effect of α level on power

Instructions

This simulation illustrates the effect of

a. sample size
b. the difference between population mean and hypothesized mean
c. the standard deviation
d. the type of test (one-tailed or two)
e. significance level on the power of a two-sample t test

You specify the difference between the population mean and hypothesized mean by either entering text in the box or by moving the slider. The population standard deviation can be entered in the box or specified by the pop-up menu. Finally, you specify the number of tails of the test. A power graph for the significance level 0.10, 0.05, and 0.01 significance levels as a function of sample size is displayed.

1. Examine the power curves. The X axis shows sample size, the Y axis shows power. Note the effect of sample size and significance level on power.
2. The default difference between the population mean and hypothesized mean is 1.55. Use the slider to change this value and note the effect on the power curves.
3. Change the population standard deviation (sd) and notice its effect on power.
4. Compare the power of one-tailed and two-tailed tests when the difference between the population mean and hypothesized mean is 2. Determine which is higher.
5. Set the difference between the population mean and hypothesized mean to zero. Since the null hypothesis is true, the Y axis no shows the **Type I** error rate, not power. Note the effect of sample size on the **Type I** error rate.
6. Determine the effect of changing the standard deviation on the **Type I** error rate.

Illustrated Instructions

Video Demo

The video demonstration begins by changing the population mean and hypothesized mean difference to 2 and then the population standard deviation to 3.5. Notice how the power curves change with each adjustment. The video by switching between one-tailed and two tailed tests.

Contributor

- Template:ContribHeblLane

This page titled 8.3: Power Demo is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.4: Power Demo II

Skills to Develop

- State the effect of effect size on power
- State the effect of one- versus two-tailed tests on power
- State the effect of the standard deviation on power
- State the effect of α level on power

Instructions

This demonstration shows various aspects of power. The graph displays power for a one-sample Z-test of the null hypothesis that the population mean is 50. The red distribution is the sampling distribution of the mean assuming the null hypothesis is true. The blue distribution is the sampling distribution of the mean based on the "true" mean. The default value of the true mean is 70. The cutoff for significance is based on the red distribution. For a one-tailed test at the 0.05 level, the cutoff point is determined so that 5% of the area is to the right of the cutoff. For the default values, the cutoff is 58.22. Therefore, any $\text{sample mean} = 58.22$ would be significant.

The shaded area of the blue distribution shows power. It is the probability that the sample mean will be 58.22 if the true mean is 70. For the default values it is 0.991.

You specify the true mean by either entering text in the box. The population standard deviation can be specified by the pop-up menu. Finally, you can specify the number of tails of the test and the significance level. Power and the cut-off points are displayed.

1. Notice the effect on power of changing the true mean to 60. Look at the distributions to find the area that represents power.
2. Change the sample size and notice its effect on the cutoff point and on power.
3. Change the population standard deviation (sd) and notice its effect on the cutoff point and on power.
4. Compare the power of one-tailed and two-tailed tests. What is the power for each when the true mean is 55. What about 45? (The one-tailed tests are based on the expectation that the $\text{population mean is} > 50$.)
5. Set the value of the true mean to 50. Does power make sense in this situation?
6. Determine the effect of the significance level on the red and blue distributions. On what (in addition to power) does it have its effect?

Illustrated Instructions

Video Demo

The video demonstration starts by changing the true mean to 55 and then the standard to 28. Notice how the probability of rejecting the null hypothesis varies with these changes.

The video concludes by changing the test to two-tailed and decreasing the significance level to 0.01. Again notice the changes in the probability of rejection.

Contributor

- Template:ContribHeblLane

This page titled 8.4: Power Demo II is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.5: Factors Affecting Power

Skills to Develop

- State five factors affecting power
- State what the effect of each of the factors is

Several factors affect the power of a statistical test. Some of the factors are under the control of the experimenter, whereas others are not. The following example will be used to illustrate the various factors.

Suppose a math achievement test were known to be normally distributed with a mean of 75 and a standard deviation of σ. A researcher is interested in whether a new method of teaching results in a higher mean. Assume that although the experimenter does not know it, the population mean μ for the new method is larger than 75. The researcher plans to sample N subjects and do a one-tailed test of whether the sample mean is significantly higher than 75. In this section, we consider factors that affect the probability that the researcher will correctly reject the false null hypothesis that the population mean is 75 or lower. In other words, factors that affect power.

Sample Size

Figure 8.5.1 shows that the larger the sample size, the higher the power. Since sample size is typically under an experimenter's control, increasing sample size is one way to increase power. However, it is sometimes difficult and/or expensive to use a large sample size.

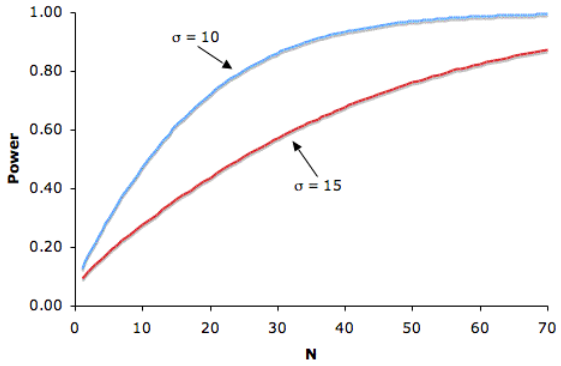


Figure 8.5.1***:*** *The relationship between sample size and power for* $H_0: \mu = 75$*, real* $\mu = 80$*, one-tailed* $\alpha = 0.05$*, for* σ*'s of* 10 *and* 15

Standard Deviation

Figure 8.5.1 also shows that power is higher when the standard deviation is small than when it is large. For all values of N, power is higher for the standard deviation of 10 than for the standard deviation of 15 (except, of course, when $N = 0$). Experimenters can sometimes control the standard deviation by sampling from a homogeneous population of subjects, by reducing random measurement error, and/or by making sure the experimental procedures are applied very consistently.

Difference between Hypothesized and True Mean

Naturally, the larger the effect size, the more likely it is that an experiment would find a significant effect. Figure 8.5.2 shows the effect of increasing the difference between the mean specified by the null hypothesis (75) and the population mean μ for standard deviations of 10 and 15.

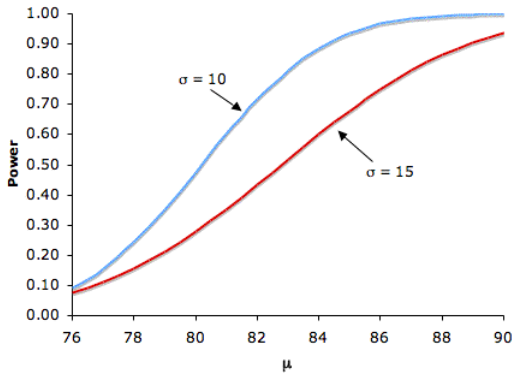


Figure 8.5.2***:*** *The relationship between* μ *and power for* $H_0 : \mu = 75$*, one-tailed* $\alpha = 0.05$*, for* $\sigma's$ *of* 10 *and* 15

Significance Level

There is a trade-off between the significance level and power: the more stringent (lower) the significance level, the lower the power. Figure 8.5.3 shows that power is lower for the 0.01 level than it is for the 0.05 level. Naturally, the stronger the evidence needed to reject the null hypothesis, the lower the chance that the null hypothesis will be rejected.

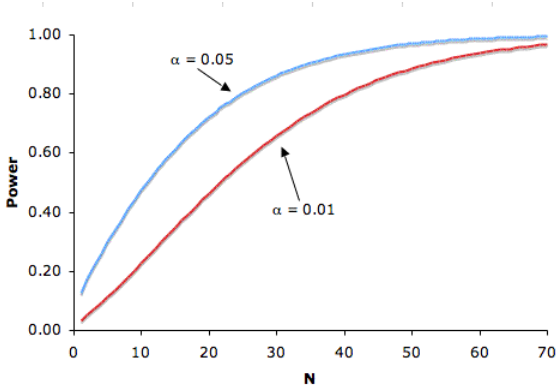


Figure 8.5.3***:*** *The relationship between significance level and power with one-tailed tests:* $\mu = 75$*, real* $\mu = 80$*, and* $\sigma = 10$

One- versus Two-Tailed Tests

Power is higher with a one-tailed test than with a two-tailed test as long as the hypothesized direction is correct. A one-tailed test at the 0.05 level has the same power as a two-tailed test at the 0.10 level. A one-tailed test, in effect, raises the significance level.

Contributor

- Template:ContribHeblLane

This page titled 8.5: Factors Affecting Power is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.6: Statistical Literacy

Skills to Develop

- Design of Studies for Alzheimer's Drug

A research design to compare three drugs for the treatment of Alzheimer's disease is described here. For the first two years of the study, researchers will follow the subjects with scans and memory tests.

Example 8.6.1: what do you think?

The data could be analyzed as a between-subjects design or as a within-subjects design. What type of analysis would be done for each type of design and how would the choice of designs affect power?

Solution

For a between-subjects design, the subjects in the different conditions would be compared after two years. For a within-subjects design, the change in subjects' scores in the different conditions would be compared. The latter would be more powerful.

Contributor

- Template:ContribHeblLane

This page titled 8.6: Statistical Literacy is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

8.7: Power (Exercises)

General Questions

Q1

Define power in your own words.

Q2

List 3 measures one can take to increase the power of an experiment. Explain why your measures result in greater power.

Q3

Population 1 mean = 36

Population 2 mean = 45

Both population standard deviations are 10.

Sample size (per group) 16.

What is the probability that a t test will find a significant difference between means at the 0.05 level? Give results for both one- and two-tailed tests. Hint: the power of a one-tailed test at 0.05 level is the power of a two-tailed test at 0.10.

Q4

Rank order the following in terms of power. n is the sample size per group.

	Population 1 Mean	n	Population 2 Mean	Standard Deviation
a	29	20	43	12
b	34	15	40	6
c	105	24	50	27
d	170	2	120	10

Q5

Alan, while snooping around his grandmother's basement stumbled upon a shiny object protruding from under a stack of boxes . When he reached for the object a genie miraculously materialized and stated: "You have found my magic coin. If you flip this coin an infinite number of times you will notice that heads will show 60% of the time." Soon after the genie's declaration he vanished, never to be seen again. Alan, excited about his new magical discovery, approached his friend Ken and told him about what he had found. Ken was skeptical of his friend's story, however, he told Alan to flip the coin 100 times and to record how many flips resulted with heads.

a. What is the probability that Alan will be able convince Ken that his coin has special powers by finding a p value below 0.05 (one tailed). Use the Binomial Calculator (and some trial and error).
b. If Ken told Alan to flip the coin only 20 times, what is the probability that Alan will not be able to convince Ken (by failing to reject the null hypothesis at the 0.05 level)?

Contributor

- Template:ContribHeblLane

This page titled 8.7: Power (Exercises) is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

CHAPTER OVERVIEW

9: Effect Size

Template:TopHeblLane

Topic hierarchy

9.1: Prelude to Effect Size
9.2: Proportions
9.3: Difference Between Two Means
9.4: Proportion of Variance Explained
9.5: Statistical Literacy
9.6: Effect Size (Exercises)

Contributors

- Template:ContribHeblLane

This page titled 9: Effect Size is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.1: Prelude to Effect Size

Researchers often seek to learn more than whether the variable under investigation has an effect and/or the direction of the effect. This is particularly true for research that has practical applications. For example, an investigation of the efficacy of a pain-relief drug would seek to determine the extent of the relief and not merely whether there was any relief. Similarly, a study of a test-preparation course's efficacy would seek to determine how much the course raises students' test scores. Finally, a study of the relationship between exercise and blood pressure would seek to determine how much blood pressure decreases for a given amount of exercise. In all of these examples, a significance test would not be sufficient since it would only provide the researcher with information about the existence and direction of the effect. It would not provide any information about the size of the effect.

Before we proceed with a discussion of how to measure effect size, it is important to consider that for some research it is the presence or absence of an effect rather than its size that is important. A controversial example is provided by Bem (2011) who investigated precognition. Bem found statistically significant evidence that subjects' responses are affected by future events. That is, he rejected the null hypothesis that there is no effect. The important question is not the size of the effect but, rather, whether it exists at all. It would be truly remarkable if future events affect present responses even a little. It is important to note that subsequent research (Ritchie, Wiseman, & French, 2012) has failed to replicate Bem's results and the likelihood that the precognition effects he described are real is very low.

Bem, D. J. (201). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology, 100, 407–425.

Ritchie, S. J., Wiseman R., and French, C. C. (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. PLoS ONE 7.

Contributor

- Template:ContribHeblLane

This page titled 9.1: Prelude to Effect Size is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.2: Proportions

Skills to Develop

- Compute absolute risk reduction
- Compute relative risk reduction
- Compute number needed to treat

Often the interpretation of a proportion is self-evident. For example, the obesity rate for white non-Hispanic adults living in the United States was estimated by a study conducted between 2006 and 2008 to be 24%. This value of 24% is easily interpretable and indicates the magnitude of the obesity problem in this population.

Often the question of interest involves the comparison of two outcomes. For example, consider the analysis of proportions in the case study "Mediterranean Diet and Health." In this study, one group of people followed the diet recommended by the American Heart Association (AHA), whereas a second group followed the "Mediterranean Diet." One interesting comparison is between the proportions of people who were healthy throughout the study as a function of diet. It turned out that 0.79 of the people who followed the AHA diet and 0.90 of those who followed the Mediterranean diet were healthy. How is the effect size of diet best measured?

We will take the perspective that we are assessing the benefits of switching from the AHA diet to the Mediterranean diet. One way to assess the benefits is to compute the difference between the proportion who were not healthy on the AHA diet (0.21) with the proportion who were not healthy on the Mediterranean diet (0.10). Therefore, the difference in proportions is:

$$0.21 - 0.10 = 0.11 \tag{9.2.1}$$

This measure of the benefit is called the ***Absolute Risk Reduction*** (ARR).

To define ARR more formally, let C be the proportion of people in the control group with the ailment of interest and T be the proportion in the treatment group. ARR can then be defined as:

$$ARR = C - T \tag{9.2.2}$$

Alternatively, one could measure the difference in terms of percentages. For our example, the proportion of non-healthy people on the Mediterranean diet (0.10) is 52% lower than the proportion of non-healthy people on the AHA diet (0.21). This value is computed as follows:

$$\frac{(0.21 - 0.10)}{0.21} \times 100 = 52\% \tag{9.2.3}$$

This measure of the benefit is called the ***Relative Risk Reduction*** (RRR). The general formula for RRR is:

$$RRR = \frac{(C - T)}{C} \times 100 \tag{9.2.4}$$

where C and T are defined as before.

A third commonly used measure is the "*odds ratio*." For our example, the odds of being healthy on the Mediterranean diet are $90 : 10 = 9 : 1$; the odds on the AHA diet are $79 : 21 = 3.76 : 1$. The ratio of these two odds is $9/3.76 = 2.39$. Therefore, the odds of being healthy on the Mediterranean diet is 2.39 times the odds of being healthy on the AHA diet. Note that the odds ratio is the ratio of the odds and not the ratio of the probabilities.

A fourth measure is the number of people who need to be treated in order to prevent one person from having the ailment of interest. In our example, being treated means changing from the AHA diet to the Mediterranean diet. The number who need to be treated can be defined as:

$$N = \frac{1}{ARR} \tag{9.2.5}$$

For our example,

$$N = \frac{1}{0.11} = 9 \tag{9.2.6}$$

Therefore, one person who would otherwise not be healthy would be expected to stay healthy for every nine people changing from the AHA diet to the Mediterranean diet.

The obvious question is which of these measures is the best one. Although each measure has its proper uses, the RRR measure can exaggerate the importance of an effect, especially when the absolute risks are low. For example, if a drug reduced the risk of a certain disease from 1 in $1,000,000$ to 1 in $2,000,000$, the RRR is 50%. However, since the ARR is only 0.0000005, the practical reduction in risk is minimal.

Contributor

- Template:ContribHeblLane

This page titled 9.2: Proportions is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.3: Difference Between Two Means

Skills to Develop

- State how the inherent meaningfulness of the scales affects the type of measure that should be used
- Compute g
- Compute d
- State the effect of the variability of subjects on the size of standardized measures

When the units of a measurement scale are meaningful in their own right, then the difference between means is a good and easily interpretable measure of effect size. For example, a study conducted by Holbrook, Crowther, Lotter, Cheng and King in 2000 investigated the effectiveness of benzodiazepine for the treatment of insomnia. These researchers found that, compared to a placebo, this drug increased total sleep duration by a mean of 61.8 minutes. This difference in means shows clearly the degree to which benzodiazepine is effective. (It is important to note that the drug was found to sometimes have adverse side effects.)

When the dependent variable is measured on a ratio scale, it is often informative to consider the proportional difference between means in addition to the absolute difference. For example, if in the Holbrook et al. study the mean total sleep time for the placebo group was 120 minutes, then the 61.8-minute increase would represent a 51% increase in sleep time. On the other hand, if the mean sleep time for the placebo were 420 minutes, then the 61.8-minute increase would represent a 15% increase in sleep time.

It is interesting to note that if a log transformation is applied to the dependent variable, then equal percent changes on the original scale will result in equal absolute changes on the log scale. For example, suppose the mean sleep time increased 10% from 400 minutes to 440 in one condition and 10% from 300 to 330 minutes in a second condition. If we take the log base 10 of these values, we find that

$$\log_{10}(440) - \log_{10}(400) = 2.643 - 2.602 = 0.041 \tag{9.3.1}$$

Similarly,

$$\log_{10}(330) - \log_{10}(300) = 2.518 - 2.477 = 0.041 \tag{9.3.2}$$

Many times the dependent variable is measured on a scale that is not inherently meaningful. For example, in the "Animal Research" case study, attitudes toward animal research were measured on a 7-point scale. The mean rating of women on whether animal research is wrong was 1.47 scale units higher than the mean rating of men. However, it is not clear whether this 1.47-unit difference should be considered a large effect or a small effect, since it is not clear exactly what this difference means.

When the scale of a dependent variable is not inherently meaningful, it is common to consider the difference between means in standardized units. That is, effect size is measured in terms of the number of standard deviations the means differ by. Two commonly used measures are Hedges' g and Cohen's d. Both of these measures consist of the difference between means divided by the standard deviation. They differ only in that Hedges' g uses the version of the standard deviation formula in which you divide by $N-1$, whereas Cohen's d uses the version in which you divide by N. The two formulas are given below.

$$g = \frac{M_1 - M_2}{\sqrt{MSE}} \tag{9.3.3}$$

$$d = g\sqrt{\frac{N}{N-2}} \tag{9.3.4}$$

where M_1 is the mean of the first group, M_2 is the mean of the second group, MSE is the mean square error, and N is the total number of observations.

Standardized measures such as Cohen's d and Hedges' g have the advantage that they are scale free. That is, since the dependent variable is standardized, the original units are replaced by standardized units and are interpretable even if the original scale units do not have clear meaning. Consider the Animal Research case study in which attitudes were measured on a 7-point scale. On a rating of whether animal research is wrong, the mean for women was 5.353, the mean for men was 3.882, and MSE was 2.864. Hedges' g can be calculated to be 0.87. It is more meaningful to say that the means were 0.87 standard deviations apart than 1.47 scale units apart, since the scale units are not well-defined.

It is natural to ask what constitutes a large effect. Although there is no objective answer to this question, the guidelines suggested by Cohen (1988) stating that an effect size of 0.2 is a small effect, an effect size of 0.5 is a medium effect, and an effect size of 0.8 is a large effect have been widely adopted. Based on these guidelines, the effect size of 0.87 is a large effect.

It should be noted, however, that these guidelines are somewhat arbitrary and have not been universally accepted. For example, Lenth (2001) argued that other important factors are ignored if Cohen's definition of effect size is used to choose a sample size to achieve a given level of power.

Interpretational Issues

It is important to realize that the importance of an effect depends on the context. For example, a small effect can make a big difference if only extreme observations are of interest. Consider a situation in which a test is used to select students for a highly selective program. Assume that there are two types of students (red and blue) and that the mean for the red students is 52, the mean for the blue students is 50, both distributions are normal, and the standard deviation for each distribution is 10. The difference in means is therefore only 0.2 standard deviations and would generally be considered to be a small difference. Now assume that only students who scored 70 or higher would be selected for the program. Would there be a big difference between the proportion of blue and red students who would be able to be accepted into the program? It turns out that the proportion of red students who would qualify is 0.036 and the proportion of blue students is 0.023. Although this difference is small in absolute terms, the ratio of red to blue students who qualify is 1.6 : 1. This means that if 100 students were to be accepted and if equal numbers of randomly-selected red and blue students applied, 62% would be red and 38% would be blue. In most contexts this would be considered an important difference.

When the effect size is measured in standard deviation units as it is for Hedges' g and Cohen's d, it is important to recognize that the variability in the subjects has a large influence on the effect size measure. Therefore, if two experiments both compared the same treatment to a control but the subjects were much more homogeneous in Experiment 1 than in Experiment 2, then a standardized effect size measure would be much larger in the former experiment than in the latter. Consider two hypothetical experiments on the effect of an exercise program on blood pressure. Assume that the mean effect on systolic blood pressure of the program is $10 mm\ Hg$ and that, due to differences in the subject populations sampled in the two experiments, the standard deviation was 20 in Experiment 1 and 30 in Experiment 2. Under these conditions, the standardized measure of effect size would be 0.50 in Experiment 1 and 0.33 in Experiment 2. This standardized difference in effect size occurs even though the effectiveness of the treatment is exactly the same in the two experiments.

Reference

1. Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.
2. Lenth, R. V. (2001) Some Practical Guidelines for Effective Sample Size Determination. The American Statistician, 55, 187-193.

Contributor

- Template:ContribHeblLane

This page titled 9.3: Difference Between Two Means is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.4: Proportion of Variance Explained

Skills to Develop

- State the difference in bias between η^2 and ω^2
- Compute η^2 Compute ω^2
- Distinguish between ω^2 and partial ω^2
- State the bias in R^2 and what can be done to reduce it

Effect sizes are often measured in terms of the proportion of variance explained by a variable. In this section, we discuss this way to measure effect size in both ANOVA designs and in correlational studies.

ANOVA Designs

Responses of subjects will vary in just about every experiment. Consider, for example, the "Smiles and Leniency" case study. A histogram of the dependent variable "leniency" is shown in Figure 9.4.1. It is clear that the leniency scores vary considerably. There are many reasons why the scores differ. One, of course, is that subjects were assigned to four different smile conditions and the condition they were in may have affected their leniency score. In addition, it is likely that some subjects are generally more lenient than others, thus contributing to the differences among scores. There are many other possible sources of differences in leniency ratings including, perhaps, that some subjects were in better moods than other subjects and/or that some subjects reacted more negatively than others to the looks or mannerisms of the stimulus person. You can imagine that there are innumerable other reasons why the scores of the subjects could differ.

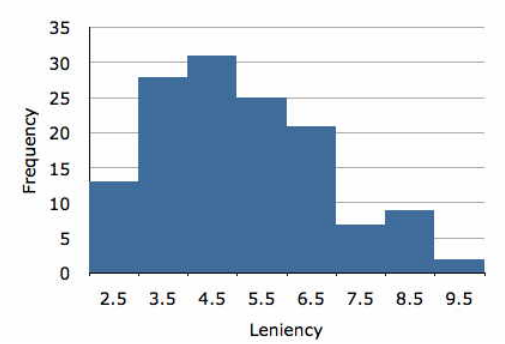


Figure* 9.4.1*: *Distribution of leniency scores*

One way to measure the effect of conditions is to determine the proportion of the variance among subjects' scores that is attributable to conditions. In this example, the variance of scores is 2.794. The question is how this variance compares with what the variance would have been if every subject had been in the same treatment condition. We estimate this by computing the variance within each of the treatment conditions and taking the mean of these variances. For this example, the mean of the variances is 2.649. Since the mean variance within the smile conditions is not that much less than the variance ignoring conditions, it is clear that "Smile Condition" is not responsible for a high percentage of the variance of the scores. The most convenient way to compute the proportion explained is in terms of the sum of squares "conditions" and the sum of squares total. The computations for these sums of squares are shown in the chapter on ANOVA. For the present data, the sum of squares for "Smile Condition" is 27.535 and the sum of squares total is 377.189. Therefore, the proportion explained by "Smile Condition" is:

$$\frac{27.535}{377.189} = 0.073 \tag{9.4.1}$$

Thus, 0.073 or 7.3% of the variance is explained by "Smile Condition."

An alternative way to look at the variance explained is as the proportion reduction in error. The sum of squares total (377.189) represents the variation when "Smile Condition" is ignored and the sum of squares error ($377.189 - 27.535 = 349.654$) is the variation left over when "Smile Condition" is accounted for. The difference between 377.189 and 349.654 is 27.535. This reduction in error of 27.535 represents a proportional reduction of $27.535/377.189 = 0.073$, the same value as computed in terms of proportion of variance explained.

This measure of effect size, whether computed in terms of variance explained or in terms of percent reduction in error, is called η^2 where η is the Greek letter eta. Unfortunately, η^2 tends to overestimate the variance explained and is therefore a biased estimate of the proportion of variance explained. As such, it is not recommended (despite the fact that it is reported by a leading statistics package).

An alternative measure, ω^2 (omega squared), is unbiased and can be computed from

$$\omega^2 = \frac{SSQ_{condition} - (k-1)MSE}{SSQ_{total} + MSE} \tag{9.4.2}$$

where MSE is the mean square error and k is the number of conditions. For this example, $k = 4$ and $\omega^2 = 0.052$.

It is important to be aware that both the variability of the population sampled and the specific levels of the independent variable are important determinants of the proportion of variance explained. Consider two possible designs of an experiment investigating the effect of alcohol consumption on driving ability. As can be seen in Table 9.4.1, **Design 1** has a smaller range of doses and a more diverse population than **Design 2**. What are the implications for the proportion of variance explained by Dose? Variation due to Dose would be greater in **Design 2** than **Design 1** since alcohol is manipulated more strongly than in **Design 1**. However, the variance in the population should be greater in **Design 1** since it includes a more diverse set of drivers. Since with **Design 1** the variance due to Dose would be smaller and the total variance would be larger, the proportion of variance explained by Dose would be much less using **Design 1** than using **Design 2**. Thus, the proportion of variance explained is not a general characteristic of the independent variable. Instead, it is dependent on the specific levels of the independent variable used in the experiment and the variability of the population sampled.

Table* 9.4.1*: *Design Parameters*

Design	Dose	Population
1	0.00	All Drivers between 16 and 80 Years of Age
	0.30	
	0.60	
2	0.00	Experienced Drivers between 25 and 30 Years of Age
	0.50	
	1.00	

Factorial Designs

In one-factor designs, the sum of squares total is the sum of squares condition plus the sum of squares error. The proportion of variance explained is defined relative to sum of squares total. In an $A \times B$ design, there are three sources of variation ($A, B, A \times B$) in addition to error. The proportion of variance explained for a variable (A, for example) could be defined relative to the sum of squares total ($SSQ_A + SSQ_B + SSQ_{A \times B} + SSQ_{error}$) or relative to $SSQ_A + SSQ_{error}$.

To illustrate with an example, consider a hypothetical experiment on the effects of age (6 and 12 years) and of methods for teaching reading (experimental and control conditions). The means are shown in Table 9.4.2. The standard deviation of each of the four cells

($Age \times Treatment$ combinations) is 5. (Naturally, for real data, the standard deviations would not be exactly equal and the means would not be whole numbers.) Finally, there were 10 subjects per cell resulting in a total of 40 subjects.

***Table* 9.4.2:** *Condition Means*

	Treatment	
Age	Experimental	Control
6	40	42
12	50	56

The sources of variation, degrees of freedom, and sums of squares from the analysis of variance summary table as well as four measures of effect size are shown in Table 9.4.3. Note that the sum of squares for age is very large relative to the other two effects. This is what would be expected since the difference in reading ability between 6- and 12-year-olds is very large relative to the effect of condition.

***Table* 9.4.3:** *ANOVA Summary Table*

Source	df	SSQ	η^2	partial η^2	ω^2	partial ω^2
Age	1	1440	0.567	0.615	0.552	0.586
Condition	1	160	0.063	0.151	0.053	0.119
A x C	1	40	0.016	0.043	0.006	0.015
Error	36	900				
Total	39	2540				

First, we consider the two methods of computing η^2, labeled η^2 and partial η^2. The value of η^2 for an effect is simply the sum of squares for this effect divided by the sum of squares total. For example, the η^2 for Age is $1440/2540 = 0.567$. As in a one-factor design, η^2 is the proportion of the total variation explained by a variable. Partial η^2 for Age is SSQ_{Age} divided by ($SSQ_{Age} + SSQ_{error}$), which is $1440/2340 = 0.615$.

As you can see, the partial η^2 is larger than η^2. This is because the denominator is smaller for the partial η^2. The difference between η^2 and partial η^2 is even larger for the effect of condition. This is because SSQ_{Age} is large and it makes a big difference whether or not it is included in the denominator.

As noted previously, it is better to use ω^2 than η^2 because η^2 has a positive bias. You can see that the values for ω^2 are smaller than for η^2. The calculations for ω^2 are shown below:

$$\omega^2 = \frac{SSQ_{effect} - df_{effect}MS_{error}}{SSQ_{total} + MS_{error}} \tag{9.4.3}$$

$$\omega^2_{partial} = \frac{SSQ_{effect} - df_{effect}MS_{error}}{SSQ_{effect} + (N - df_{effect})MS_{error}} \tag{9.4.4}$$

where N is the total number of observations.

The choice of whether to use ω^2 or the partial ω^2 is subjective; neither one is correct or incorrect. However, it is important to understand the difference and, if you are using computer software, to know which version is being computed. (Beware, at least one software package labels the statistics incorrectly).

Correlational Studies

In the section "Partitioning the Sums of Squares" in the Regression chapter, we saw that the sum of squares for Y (the criterion variable) can be partitioned into the sum of squares explained and the sum of squares error. The proportion of variance explained in multiple regression is therefore:

$$SSQ_{explained}/SSQ_{total} \tag{9.4.5}$$

In simple regression, the proportion of variance explained is equal to r^2; in multiple regression, it is equal to R^2.

In general, R^2 is analogous to η^2 and is a biased estimate of the variance explained. The following formula for adjusted R^2 is analogous to ω^2 and is less biased (although not completely unbiased):

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \tag{9.4.6}$$

where N is the total number of observations and p is the number of predictor variables.

Contributor

- Template:ContribHeblLane

This page titled 9.4: Proportion of Variance Explained is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.5: Statistical Literacy

Skills to Develop

- Health Effects of Coffee

This article describes some health effects of drinking coffee. Among the key findings were

a. women who drank four or more cups a day reduced their risk of endometrial cancer by 25% compared with those who drank less than one cup a day
b. men who drank six or more cups had a 60% lower risk of developing the most deadly form of prostate cancer than those who drank less than one cup a day

Example 9.5.1: what do you think?

What is the technical term for the measure of risk reduction reported? What measures of risk reduction cannot be determined from the article? What additional information would have been helpful for assessing risk reduction?

Solution

This is called the "relative risk reduction." The article does not provide information necessary to compute the absolute risk reduction, the odds ratio, or the number needed to treat. It would have been helpful if the article had reported the proportion of women drinking less than one cup a day who developed endometrial cancer as well as the analogous statistic for men and prostate cancer.

Contributor

- Template:ContribHeblLane

This page titled 9.5: Statistical Literacy is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

9.6: Effect Size (Exercises)

General Questions

Q1

If the probability of a disease is 0.34 without treatment and 0.22 with treatment then what is the

a. Absolute risk reduction
b. Relative risk reduction
c. Odds ratio
d. Number needed to treat

Q2

When is it meaningful to compute the proportional difference between means?

Q3

The mean for an experimental group is 12, the mean for the control group were 8, the MSE from the ANOVA is 16, and N, the number of observations is 20, compute g and d.

Q4

Two experiments investigated the same variables but one of the experiment had subject who differed greatly from each other where as the subjects in the other experiment were relatively homogeneous. Which experiment would likely have the larger value of g?

Q5

Why is ω^2 preferable to η^2?

Q6

What is the difference between η^2 and partial η^2?

Questions from Case Studies

The following questions are from the Teacher Ratings case study.

Q7

What are the values of d and g?

Q8

What are the values of ω^2 and η^2?

The following question is from the Smiles and Leniency case study.

Q9

What are the values of ω^2 and η^2?

The following question is from the Obesity and Bias case study.

Q10

For compute ω^2 and partial $\omega^2$for the effect of "Weight" in a "Weight x Relatedness" ANOVA.

Contributor

- Template:ContribHeblLane

This page titled 9.6: Effect Size (Exercises) is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

CHAPTER OVERVIEW

10: 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.

10.1: Prelude to Confidence Intervals

10.2: A Single Population Mean using the Normal Distribution

A Single Population Mean using the Normal Distribution (Exercises)

10.3: A Single Population Mean using the Student t-Distribution

10.4: A Population Proportion

10.5: Confidence Interval - Home Costs (Worksheet)

10.6: Confidence Interval -Place of Birth (Worksheet)

10.7: Confidence Interval -Women's Heights (Worksheet)

10.8: Confidence Intervals (Exercises)

10.9: Confidence Intervals (Summary)

Contributors

- Template:ContribOpenStax

This page titled 10: Confidence Intervals is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.1: Prelude to Confidence Intervals

Skills to Develop

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

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's *t* distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.



Figure 8.1.1. *Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)*

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of $\bar{x}$ and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1. \quad (10.1.1)$$

The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \quad (10.1.2)$$

and

$$\bar{x} + 0.2 = 2 + 0.2 = 2.2 \quad (10.1.3)$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is (1.8, 2.2). This 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean $\bar{x}$ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean $\bar{x}$. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, $\bar{x}$. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

COLLABORATIVE EXERCISE

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval $\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between _____ and _____.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

Contributors

- Template:ContribOpenStax

This page titled 10.1: Prelude to Confidence Intervals is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.2: A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, where the population standard deviation is known, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (EBM) is called the error bound for a population mean (abbreviated *EBM*). The sample mean $\bar{x}$ is the point estimate of the unknown population mean μ.

The confidence interval estimate will have the form:

$$(\text{point estimate} - \text{error bound}, \text{point estimate} + \text{error bound})$$

or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The **margin of error** (EBM) depends on the confidence level (abbreviated CL). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically,

$$\alpha + CL = 1.$$

Example 10.2.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is seven, and the error bound for the mean is 2.5: $\bar{x} = 7$ and $EBM = 2.5$

The confidence interval is (7 – 2.5, 7 + 2.5) and calculating the values gives (4.5, 9.5). If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Exercise 10.2.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2. What is the confidence interval estimate for the population mean?

Answer

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$. To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10$ in both tails, or 5% in each tail, of the normal distribution.

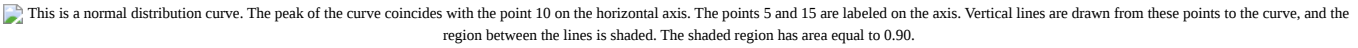

This is a normal distribution curve. The peak of the curve coincides with the point 10 on the horizontal axis. The points 5 and 15 are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 0.90.

***Figure* 10.2.1**

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the *z*-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far

left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is

$$\frac{\sigma}{\sqrt{n}}$$

This fraction is commonly called the "standard error of the mean" to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N(\mu_x, \frac{\sigma}{\sqrt{n}})$.
- When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0,1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

using a calculator, computer or a standard normal probability table.

`invNorm` $(0.975, 0, 1) = 1.96$

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0,1)$.

Calculating the Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EBM = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} = EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

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

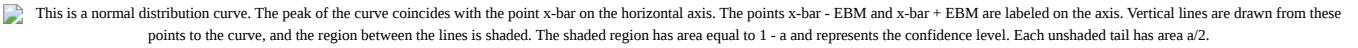

This is a normal distribution curve. The peak of the curve coincides with the point x-bar on the horizontal axis. The points x-bar - EBM and x-bar + EBM are labeled on the axis. Vertical lines are drawn from these points to the curve, and the region between the lines is shaded. The shaded region has area equal to 1 - a and represents the confidence level. Each unshaded tail has area a/2.

Figure 8.2.2.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 10.2.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Answer

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

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

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3; n = 36$$

The confidence level is 90% (CL = 0.90)

$$CL = 0.90$$

so

$$\alpha = 1–CL = 1–0.90 = 0.10$$

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

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

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

using $\text{invNorm}(0.95, 0, 1)$ on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

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

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

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

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

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `7:ZInterval` .
Press `ENTER` .
Arrow to `Stats` and press `ENTER` .
Arrow down and enter three for σ, 68 for $\bar{x}$, 36 for n, and .90 for `C-level` .
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Exercise 10.2.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.1347, 37.8653)

Example 10.2.3: Specific Absorption Rate

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$$\bar{x} = 1.024$$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.

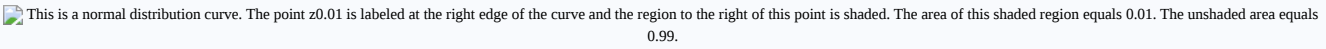

This is a normal distribution curve. The point z0.01 is labeled at the right edge of the curve and the region to the right of this point is shaded. The area of this shaded region equals 0.01. The unshaded area equals 0.99.

Figure 8.2.3.

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$$\bar{x} - EBM = 1.024–0.1431 = 0.8809$$

$$\bar{x} - EBM = 1.024–0.1431 = 1.1671$$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to 7:Z Interval.
- Press ENTER.
- Arrow to Stats and press ENTER.

- Arrow down and enter the following values:
  - σ : 0.337
  - $\bar{x}$: 1024
  - n : 30
  - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (to three decimal places) (0.881, 1.167).

Exercise 10.2.3

Table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Answer

$$\bar{x} = 0.940$$

$$\frac{\alpha}{2} = \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035$$

$$z_{0.035} = 1.812$$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$$\bar{x} - EBM = 0.940 - 0.1365 = 0.8035$$

$$\bar{x} + EBM = 0.940 + 0.1365 = 1.0765$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 10.2.4

Suppose we change the original problem in Example by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Answer

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

$$\bar{x} = 68$$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3; n = 36$; The confidence level is 95% (CL = 0.95).

$CL = 0.95$ so $\alpha = 1–CL = 1–0.95 = 0.05$

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

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1–0.025 = 0.975$.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x}–EBM = 68–0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the EBM is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



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

Figure 8.2.4.

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Exercise 10.2.4

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Answer

(33.37, 38.63)

Example 10.2.5

Suppose we change the original problem in Example to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Answer

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935.$

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25 : EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987.$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Exercise 10.2.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,

- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 10.2.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is $68: EBM = 68.82–68 = 0.82$.
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82–0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Exercise 10.2.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population mean when the population standard deviation is known is

$$EBM = \left(z\frac{a}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n. In Equation , z is $z\frac{a}{2}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 10.2.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2}$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Exercise 10.2.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Answer

35 students

References

1. “American Fact Finder.” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/...html?refresh=t (accessed July 2, 2013).
2. “Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at http://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 http://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 http://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 http://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

- Template:ContribOpenStax

This page titled 10.2: A Single Population Mean using the Normal Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

A Single Population Mean using the Normal Distribution (Exercises)

A Single Population Mean using the Normal Distribution (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

10.3: A Single Population Mean using the Student 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 σ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ. If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}, \qquad (10.3.1)$$

then the t-scores follow a Student's t-distribution with n–1 degrees of freedom. The t-score has the same interpretation as the z-score. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The degrees of freedom, n–1, come from the calculation of the sample standard deviation s. In [link], we used n deviations ($x - \bar{x}$ values) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other n–1 deviations. The other n–1 deviations can change or vary freely. We call the number n–1 the degrees of freedom (df).

> *For each sample size n, there is a different Student's t-distribution.*

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator,

you need to use a probability table for the Student's t-Distribution.) When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t-table (See [link]) 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:

- $EBM = (t_{\frac{\alpha}{2}})\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n$–1 degrees of freedom, and
- $s =$ sample standard deviation.

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM). \tag{10.3.2}$$

To calculate the confidence interval directly:

Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 10.3.1: Acupuncture

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution A

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

$$\bar{x} = 8.2267\backslash(s = 1.6722 \text{ n} = 15\backslash)$$

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

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

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.

$$EBM = (t_{\frac{\alpha}{2}})\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$$\bar{x}–EBM = 8.2267–0.9240 = 7.3$$

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

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq` .
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER` .
The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

Exercise 10.3.2

You do a study of hypnotherapy to determine how effective it is in increasing the number of hourse of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Answer

(8.1634, 9.8032)

Example 10.3.2: The Human Toxome Project

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution A

From the sample, you can calculate $\bar{x} = 127.45$ and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20–1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$, so

$$\alpha = 1–CL = 1–0.90 = 0.10 \frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05} \tag{10.3.3}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1–0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER` .
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER` .

The 90% confidence interval is (117.41, 137.49).

Exercise 10.3.2

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. 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.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A

$\bar{x} = 6.133$, \(s = 5.514, n = 15\), and $df = 15 - 1 = 14$

$CL = 0.98$, so $\alpha = 1 - CL = 1 - 0.98 = 0.02$

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

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.624\left(\frac{5.514}{\sqrt{15}}\right) - 3.736$$

$$\bar{x} - EBM = 6.133 - 3.736 = 2.397$$

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

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B

Enter the data as a list.

Press `STAT` and arrow over to `TESTS` .
Arrow down to `8:TInterval` .
Press `ENTER` .
Arrow to `Data` and press `ENTER` .
Arrow down and enter the name of the list where the data is stored.
Enter `Freq` : 1
Enter `C-Level` : 0.98

Arrow down to `Calculate` and press `Enter`.
The 98% confidence interval is (2.3965, 9,8702).

Reference

1. “America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).
2. Data from *Microsoft Bookshelf*.
3. Data from http://www.businessweek.com/.
4. Data from http://www.forbes.com/.
5. “Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at http://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 http://www.ewg.org/sites/humantoxome...tero%2Fnewborn (accessed July 2, 2013).
7. “Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at http://www.fec.gov/finance/disclosur...pPacList.shtml (accessed July 2, 2013).

Glossary

Degrees of Freedom (df)

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

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's \(t\)-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a "family of t-distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

Contributors

- Template:ContribOpenStax

This page titled 10.3: A Single Population Mean using the Student t-Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.4: A Population Proportion

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

$$X \sim B(n, p) \tag{10.4.1}$$

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,

$$P' = \frac{X}{n} \tag{10.4.2}$$

(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.

$$X \sim N(np, \sqrt{npq}) \tag{10.4.3}$$

If we divide the random variable, the mean, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n.)

$$\frac{X}{n} = P' - N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right) \tag{10.4.4}$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}} \tag{10.4.5}$$

P' follows a normal distribution for proportions:

$$\frac{X}{n} = P' - N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right) \tag{10.4.6}$$

The confidence interval has the form

$$(p'–EBP, p'+EBP). \tag{10.4.7}$$

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

$$EBP = (z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) \tag{10.4.8}$$

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

$$\sqrt{\frac{pq}{n}}. \tag{10.4.9}$$

However, in the error bound formula, we use

$$\sqrt{\frac{p'q'}{n}} \tag{10.4.10}$$

as the standard deviation, instead of

$$\sqrt{\frac{pq}{n}}. \tag{10.4.11}$$

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

Normal Distribution of Proportions

For the normal distribution of proportions, the z-score formula is as follows.

If

$$P' - N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{10.4.12}$$

then the z-score formula is

$$z = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \tag{10.4.13}$$

Example 10.4.1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution A

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Let $X =$ the number of people in the sample who have cell phones. X is binomial.

$$X \sim B(500, \frac{421}{500}). \tag{10.4.14}$$

To calculate the confidence interval, you must find p', q', and EBP.

- $n = 500$
- $x =$ the number of successes $= 421$

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842 \tag{10.4.15}$$

- $p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158 \tag{10.4.16}$$

Since $CL = 0.95$, then

$$\alpha = 1 - CL = 1 - 0.95 = 0.05 \left(\frac{\alpha}{2}\right) = 0.025. \tag{10.4.17}$$

Then

$$z_{\frac{\alpha}{2}} = z_{0.025 = 1.96} \tag{10.4.18}$$

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.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032 \tag{10.4.19}$$

$$p' - EBP = 0.842 - 0.032 = 0.81 \tag{10.4.20}$$

$$p' + EBP = 0.842 + 0.032 = 0.874 \tag{10.4.21}$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 421.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER` .
The confidence interval is (0.81003, 0.87397).

Exercise 10.4.1

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Answer

(0.3315, 0.4525)

Example 10.4.2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Answer

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

- $x = 300$ and
- $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600 \tag{10.4.22}$$

$$q' = 1 - p' = 1 - 0.600 = 0.400 \tag{10.4.23}$$

Since $CL = 0.90$, then

\[\alpha = 1 – CL = 1 – 0.90 = 0.10\left(\dfrac{\alpha}{2}\right) = 0.05\[

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645 \tag{10.4.24}$$

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.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036 \tag{10.4.25}$$

$$p'–EBP = 0.60 - 0.036 = 0.564 \tag{10.4.26}$$

$$p'+EBP = 0.60 + 0.036 = 0.636 \tag{10.4.27}$$

The confidence interval for the true binomial population proportion is $(p'–EBP, p'+EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Press `STAT` and arrow over to `TESTS` .

Arrow down to `A:1-PropZint` . Press `ENTER` .
Arrow down to xx and enter 300.
Arrow down to nn and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER` .

The confidence interval is (0.564, 0.636).

Exercise 10.4.2

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

Answer

(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%.

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

Solution A

Sixty-eight percent (68%) of students own an iPod and a smart phone.

$$p' = 0.68 \tag{10.4.28}$$

$$q' = 1 - p' = 1 - 0.68 = 0.32 \tag{10.4.29}$$

Since $CL = 0.97$, we know

$$\alpha = 1 - 0.97 = 0.03 \tag{10.4.30}$$

and

$$\frac{\alpha}{2} = 0.015. \tag{10.4.31}$$

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),

$$z_{0.05} = 2.17 \tag{10.4.32}$$

$$EPB = \left(z\frac{\alpha}{2}\right)\sqrt{\frac{p'q'}{n}} = 2.17\sqrt{\frac{0.68(0.32)}{300}} \approx 0.0269 \tag{10.4.33}$$

$$p' - EPB = 0.68 - 0.0269 = 0.6531 \tag{10.4.34}$$

$$p' + EPB = 0.68 + 0.0269 = 0.7069 \tag{10.4.35}$$

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.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 300*0.68.
Arrow down to n and enter 300.
Arrow down to C-Level and enter 0.97.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

"Plus Four" Confidence Interval for p

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation

of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$. Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example 10.4.3

A random sample of 25 statistics students was asked: "Have you smoked a cigarette in the past week?" Six students reported smoking within the past week. Use the "plus-four" method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution A

Six students out of 25 reported smoking within the past week, so $x = 6$ and $n = 25$. Because we are using the "plus-four" method, we will use $x = 6 + 2 = 8$ and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276 \tag{10.4.36}$$

$$q' = 1 - p' = 1 - 0.276 = 0.724 \tag{10.4.37}$$

Since $CL = 0.95$, we know $\alpha = 1 - 0.95 = 0.05$ and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96 \tag{10.4.38}$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$$

$$p' - EPB = 0.276 - 0.163 = 0.113 \tag{10.4.39}$$

$$p' + EPB = 0.276 + 0.163 = 0.439 \tag{10.4.40}$$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.

REMINDER

Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to x and enter eight.

Arrow down to n and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.113, 0.439).

Exercise 10.4.3

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the "plus-four" method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution A

Using "plus four," we have $x = 31 + 2 = 33$ and $n = 65 + 4 = 69$.

$$p' = 3369 \approx 0.478 \tag{10.4.41}$$

$$q' = 1 - p' = 1 - 0.478 = 0.522 \tag{10.4.42}$$

Since $CL = 0.96$, we know $\alpha = 1 - 0.96 = 0.04$ and $\frac{\alpha}{2} = 0.02$.

$$z_{0.02} = 2.054 \tag{10.4.43}$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (2.054)\left(\sqrt{\frac{(0.478)(0.522)}{69}}\right) - 0.124 \tag{10.4.44}$$

$$p' - EPB = 0.478 - 0.124 = 0.354 \tag{10.4.45}$$

$$p' + EPB = 0.478 + 0.124 = 0.602 \tag{10.4.46}$$

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

Solution B

Press STAT and arrow over to TESTS.

Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 33.
Arrow down to n and enter 69.
Arrow down to C-Level and enter 0.96.
Arrow down to Calculate and press ENTER.

The confidence interval is (0.355, 0.602).

Example 10.4.4

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the "plus four" method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution A

Using "plus-four," we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$.

$$p' = \frac{15}{54} \approx 0.278 \tag{10.4.47}$$

$$q' = 1 - p' = 1 - 0.241 = 0.722 \tag{10.4.48}$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$.

$$z_{0.05} = 1.645 \tag{10.4.49}$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100 \tag{10.4.50}$$

$$p' - EPB = 0.278 - 0.100 = 0.178 \tag{10.4.51}$$

$$p' + EPB = 0.278 + 0.100 = 0.378 \tag{10.4.52}$$

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 10.4.4

The Berkman Center Study referenced in Example talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the "plus-four" method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in Example.

Answer

Solution A

Using "plus-four," we have $x = 159 + 2 = 161$ and $n = 588 + 4 = 592$.

$$p' = \frac{161}{592} \approx 0.272 \tag{10.4.53}$$

$$q' = 1 - p' = 1 - 0.272 = 0.728 \tag{10.4.54}$$

Since $CL = 0.90$, we know $\alpha = 1 - 0.90 = 0.10$ and $\frac{\alpha}{2} = 0.05$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030 \tag{10.4.55}$$

$$p' - EPB = 0.272 - 0.030 = 0.242 \tag{10.4.56}$$

$$p' + EPB = 0.272 + 0.030 = 0.302 \tag{10.4.57}$$

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution B

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint. Press ENTER.
- Arrow down to x and enter 161.
- Arrow down to n and enter 592.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.
- The confidence interval is (0.242, 0.302).

Conclusion: The confidence interval for the larger sample is narrower than the interval from Example. Larger samples will always yield more precise confidence intervals than smaller samples. The "plus four" method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right) \tag{10.4.58}$$

Solving for n gives you an equation for the sample size.

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2} \tag{10.4.59}$$

Example 10.4.5

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Answer

From the problem, we know that $\mathbf{EBP = 0.03}$ (3%=0.03) and $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$ because the confidence level is 90%.

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that $q' = 1 - p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2} \tag{10.4.60}$$

gives

$$n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7 \tag{10.4.61}$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Exercise 10.4.5

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Answer

271 customers should be surveyed. Check the Real Estate section in your local

Glossary

Binomial Distribution

a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x = \binom{n}{x}) p^x q^{n-x}$.

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

- Template:ContribOpenStax

This page titled 10.4: A Population Proportion is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.5: Confidence Interval - Home Costs (Worksheet)

This page titled 10.5: Confidence Interval - Home Costs (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.6: Confidence Interval -Place of Birth (Worksheet)

This page titled 10.6: Confidence Interval -Place of Birth (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.7: Confidence Interval -Women's Heights (Worksheet)

This page titled 10.7: Confidence Interval -Women's Heights (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.8: Confidence Intervals (Exercises)

This page titled 10.8: Confidence Intervals (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

10.9: Confidence Intervals (Summary)

Chapter Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

$$\text{(lower bound, upper bound)} = \text{(point estimate}–EBM, \text{point estimate} + EBM)$$

The calculation of EBM depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding EBM increases as well. As the sample size increases, the EBM decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (EBM) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the EBM formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

Formula Review

$\bar{X} - N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

$$\text{(lower bound, upper bound)} = \text{(point estimate}–EBM, \text{point estimate} + EBM) \quad (10.9.1)$$

$$= \bar{x} - EBM, \bar{x} + EBM \quad (10.9.2)$$

$$= \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right) \quad (10.9.3)$$

$EBM = z\frac{\sigma}{\sqrt{n}} =$ the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

$CL =$ confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1–CL =$ the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of "EBM where $\alpha = 1–CL$".

$n = \frac{z^2\sigma^2}{EBM^2}$ the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

$$\text{(lower value, upper value)} = \text{(point estimate} - \text{error bound, point estimate} + \text{error bound)} \quad (10.9.4)$$

To find the error bound when you know the confidence interval

$$\text{error bound} = \text{upper value} - \text{point estimate} \quad (10.9.5)$$

OR

$$\text{error bound} = \frac{\text{upper value - lower value}}{2} \tag{10.9.6}$$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

Exercise 8.2.8

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. 244
b. 15
c. 50

Exercise 8.2.9

In words, define the random variables X and $\bar{X}$.

Exercise 8.2.10

Which distribution should you use for this problem?

Answer

$N\left(244, \frac{15}{\sqrt{50}}\right)$

Exercise 8.2.11

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

Exercise 8.2.12

What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Answer

As the sample size increases, there will be less variability in the mean, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Exercise 8.2.13

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Exercise 8.2.14

In words, define the random variables X and $\bar{X}$.

Answer

X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Exercise 8.2.15

Which distribution should you use for this problem?

Exercise 8.2.16

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$CI: (7.9441, 8.4559)$

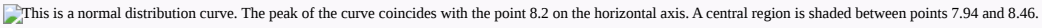


Figure 8.2.3.

$EBM = 0.26$

Exercise 8.2.17

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

Exercise 8.2.18

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Answer

The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

Exercise 8.2.19

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Exercise 8.2.20

Identify the following:

a. $\bar{x} =$ _____
b. $\sigma =$ _____
c. $n =$ _____

Answer

a. $\bar{x} = 2.2$
b. $\sigma = 0.2$
c. $n = 20$

Exercise 8.2.21

In words, define the random variable X.

Exercise 8.2.22

In words, define the random variable $\bar{X}$.

Answer

$\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

Exercise 8.2.23

Which distribution should you use for this problem?

Exercise 8.2.24

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

$EBM = 0.07$

$CI : (2.1264, 2.2736)$

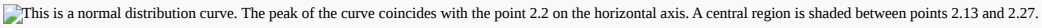


Figure 8.2.4.

Exercise 8.2.25

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

Exercise 8.2.26

In complete sentences, explain why the confidence interval in Exercise is larger than in Exercise.

Answer

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

Exercise 8.2.27

In complete sentences, give an interpretation of what the interval in Exercise means.

Exercise 8.2.28

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Answer

The confidence level would increase.

Exercise 8.2.29

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

Exercise 8.2.30

$\bar{x} =$ _____

Answer

30.4

Exercise 8.2.31

\(n =\) _____

Exercise 8.2.32

_____ $= 15$

Answer

σ

Exercise 8.2.33

In words, define the random variable $\bar{X}$.

Exercise 8.2.34

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.2.35

Is σ_x known?

Exercise 8.2.36

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

normal

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

Exercise 8.2.37

How much area is in both tails (combined)? $\alpha =$ _____

Exercise 8.2.38

How much area is in each tail? $\frac{\alpha}{2} =$ _____

Answer

0.025

Exercise 8.2.39

Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

Exercise 8.2.40

The 95% confidence interval is:_____.

Answer

(24.52,36.28)

Exercise 8.2.41

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

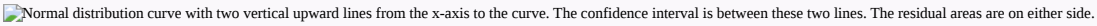


Figure 8.2.5.

Exercise 8.2.42

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

Exercise 8.2.43

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise 8.2.44

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Answer

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Chapter Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{10.9.7}$$

The t-score follows the Student's t-distribution with n–1 degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

$s =$ the standard deviation of sample values.

$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} =$ the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t-score in the Student's t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

$$= (\text{point estimate} - EBM, \text{point estimate} + EBM) \tag{10.9.8}$$

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}}\right) \tag{10.9.9}$$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

Exercise 8.3.3

Identify the following:

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

Exercise 8.3.4

Define the random variables X and $\bar{X}$ in words.

Answer

X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

Exercise 8.3.5

Which distribution should you use for this problem?

Exercise 8.3.6

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(1.3808, 1.6192)$

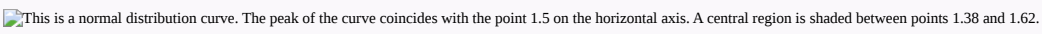


Figure 8.3.1.

$EBM = 0.12$

Exercise 8.3.7

Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Exercise 8.3.8

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____
d. $n - 1 =$_____

Answer

a. $\bar{x} = 151$
b. $s_x = 32$
c. $n = 108$
d. $n - 1 = 107$

Exercise 8.3.9

Define the random variable X in words.

Exercise 8.3.10

Define the random variable $\bar{X}$ in words.

Answer

$\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

Exercise 8.3.11

Which distribution should you use for this problem?

Exercise 8.3.12

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

Answer

$CI: (142.92, 159.08)$

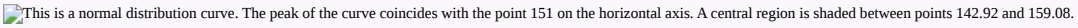


Figure 8.3.2.

$EBM = 8.08$

Exercise 8.3.13

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in Table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let $X =$ the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Exercise 8.3.14

a. $\bar{x} =$_____
b. $s_x =$_____
c. $n =$_____

Answer

a. 3.26

b. 1.02
c. 39

Exercise 8.3.15

Define the random variable $\bar{X}$ in words.

Exercise 8.3.16

What is $\bar{x}$ estimating?

Answer

μ

Exercise 8.3.17

Is σ_x known?

Exercise 8.3.18

As a result of your answer to Exercise, state the exact distribution to use when calculating the confidence interval.

Answer

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

Exercise 8.3.19

How much area is in both tails (combined)?

Exercise 8.3.20

How much area is in each tail?

Answer

0.025

Exercise 8.3.21

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.3.22

The 95% confidence interval is_____.

Answer

(2.93, 3.59)

Exercise 8.3.23

Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

This is a template of a normal distribution curve with the central region shaded to represent a confidence interval. The residual areas are on either side of the shaded region. Blanks indicate that students should label the confidence level, residual areas, and points that define the confidence interval.

Figure 8.3.3.

Exercise 8.3.24

In one complete sentence, explain what the interval means.

Answer

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Exercise 8.3.25

Using the same $\bar{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Answer

The error bound would become $EBM = 0.245$. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

Exercise 8.3.26

Using the same $\bar{x}$, s_x, and $n = 39$, how would the error bound change if the confidence level were reduced to 90%? Why?

References

1. Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).
2. Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at http://www.pewinternet.org/Reports/2...d-Privacy.aspx (accessed July 2, 2013).
3. Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//...al%20Media.pdf (accessed July 2, 2013).
4. Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/th...ement-age.aspx (accessed July 2, 2013).
5. The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).
6. Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/2...analytics-poll (accessed July 2, 2013).
7. “52% Say Big-Time College Athletics Corrupt Education Process.” Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/publ...cation_process (accessed July 2, 2013).

Chapter Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, $\frac{x}{n}$, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}\right)$$

The “plus four” method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n_4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = \frac{x}{n}$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p' \tag{10.9.10}$$

$$p' - N\left(p, \sqrt{\frac{pq}{n}}\right) \tag{10.9.11}$$

The variable p' has a binomial distribution that can be approximated with the normal distribution shown here.

$EBP =$ the error bound for a proportion $= z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$

Confidence interval for a proportion:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}}\right), p' + z\sqrt{\frac{p'q'}{n}}$$

$n = \frac{z_{\frac{\alpha}{2}}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error EBP.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$

$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} p' + q' = 1$

The confidence interval has the format $(p' - EBP, p' + EBP)$.

- $\bar{x}$ is a point estimate for μ
- p' is a point estimate for ρ
- s is a point estimate for σ

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

Exercise 8.4.6

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

Exercise 8.4.7

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Answer

It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

Exercise 8.4.8

Identify the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

Exercise 8.4.9

Define the random variables X and P' in words.

Answer

X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

Exercise 8.4.10

Which distribution should you use for this problem?

Exercise 8.4.11

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.5321, 0.6679)$

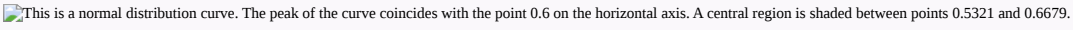


Figure 8.4.1.

EBM : 0.0679

Exercise 8.4.12

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

Exercise 8.4.13

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

Answer

X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

Exercise 8.4.14

Which distribution should you use for this problem?

Exercise 8.4.15

Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

Answer

CI : $(0.19432, 0.33068)$

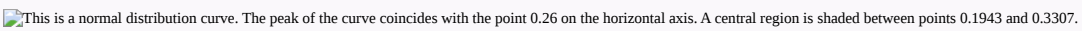


Figure 8.4.2.

EBM : 0.0707

Exercise 8.4.16

Suppose we want to lower the sampling error. What is one way to accomplish that?

Exercise 8.4.17

The sampling error given in the survey is ± 2. Explain what the ± 2 means.

Answer

The sampling error means that the true mean can be 2% above or below the sample mean.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

Exercise 8.4.18

Define the random variable X in words.

Exercise 8.4.19

Define the random variable P' in words.

Answer

P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Exercise 8.4.20

Which distribution should you use for this problem?

Exercise 8.4.21

Construct a 90% confidence interval, and state the confidence interval and the error bound.

Answer

CI : $(0.62735, 0.67265)$

EBM : 0.02265

Exercise 8.4.22

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

Exercise 8.4.23

What is being counted?

Answer

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

Exercise 8.4.24

In words, define the random variable X.

Exercise 8.4.25

Calculate the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

Answer

a. $x = 64$
b. $n = 80$

c. $p' = 0.8$

Exercise 8.4.26

State the estimated distribution of X. $X \sim$_____

Exercise 8.4.27

Define a new random variable P'. What is p' estimating?

Answer

p

Exercise 8.4.28

In words, define the random variable P'.

Exercise 8.4.29

State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Answer

$P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. $(0.72171, 0.87829)$.

Exercise 8.4.30

How much area is in both tails (combined)?

Exercise 8.4.31

How much area is in each tail?

Answer

0.04

Exercise 8.4.32

Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise 8.4.33

The 92% confidence interval is _____.

Answer

(0.72; 0.88)

Exercise 8.4.34

Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

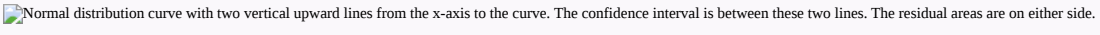


Figure 8.4.3.

Exercise 8.4.35

In one complete sentence, explain what the interval means.

Answer

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Exercise 8.4.36

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise 8.4.37

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Answer

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

Exercise 8.4.38

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Contributors

- Template:ContribOpenStax

This page titled 10.9: Confidence Intervals (Summary) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

11: Hypothesis Testing with One Sample

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

11.1: Prelude to Hypothesis Testing

11.2: Null and Alternative Hypotheses

Null and Alternative Hypotheses (Exercises)

11.3: Outcomes and the Type I and Type II Errors

Outcomes and the Type I and Type II Errors (Exercises)

11.4: Distribution Needed for Hypothesis Testing

Distribution Needed for Hypothesis Testing (Exercises)

11.5: Rare Events, the Sample, Decision and Conclusion

Rare Events, the Sample, Decision and Conclusion (Exercises)

11.6: Additional Information and Full Hypothesis Test Examples

11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)

11.8: Hypothesis Testing with One Sample (Exercises)

Contributors

- Template:ContribOpenStax

This page titled 11: Hypothesis Testing with One Sample is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

11.1: Prelude to Hypothesis Testing

CHAPTER OBJECTIVES

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

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.



Figure* 11.1.1*: *You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)*

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analysis of the data, to reject the null hypothesis. In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.
- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
- Make a decision and write a meaningful conclusion.

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Contributors

- Template:ContribOpenStax

This page titled 11.1: Prelude to Hypothesis Testing is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

11.2: Null and Alternative Hypotheses

The actual test begins by considering two hypotheses. They are called the null hypothesis and the alternative hypothesis. These hypotheses contain opposing viewpoints.

- *The null hypothesis* (H_0) is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.
- *The alternative hypothesis* (H_a) is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

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* 11.2.1*: *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 11.2.1

- H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
- H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Exercise 11.2.1

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Answer

- H_0: The drug reduces cholesterol by 25%. $p = 0.25$
- H_a: The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example 11.2.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

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

Exercise 11.2.2

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol ($=, \neq, \geq, <, \leq, >$) for the null and alternative hypotheses.

- $H_0 : \mu_6 6$
- $H_a : \mu_6 6$

Answer

- $H_0 : \mu = 66$
- $H_a : \mu \neq 66$

Example 11.2.3

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:

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

Exercise 11.2.3

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. $H_0 : \mu_4 5$
b. $H_a : \mu_4 5$

Answer

a. $H_0 : \mu \geq 45$
b. $H_a : \mu < 45$

Example 11.2.4

In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

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

Exercise 11.2.4

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol $(=, \neq, \geq, <, \leq, >)$ for the null and alternative hypotheses.

a. $H_0 : p_0 .40$
b. $H_a : p_0 .40$

Answer

a. $H_0 : p = 0.40$
b. $H_a : p > 0.40$

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Chapter Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality $(=, \leq \text{ or } \geq)$

2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., $(\neq, >, \text{or } <)$.
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

If H_a has:	equal $(=)$	greater than or equal to $(\geq)$	less than or equal to $(\leq)$
then H_a has:	not equal $(\neq)$ **or** greater than $(>)$ **or** less than $(<)$	less than $(<)$	greater than $(>)$

- If $\alpha \leq p$-value, then do not reject H_0.
- If$\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Contributors

- Template:ContribOpenStax

This page titled 11.2: Null and Alternative Hypotheses is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Null and Alternative Hypotheses (Exercises)

Exercise 5

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

Answer

The random variable is the mean Internet speed in Megabits per second.

Exercise 1

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Exercise 1

The American family has an average of two children. What is the random variable? Describe in words.

Answer

The random variable is the mean number of children an American family has.

Exercise 9.2.8

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Exercise 9.2.9

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

Answer

The random variable is the proportion of people picked at random in Times Square visiting the city.

Exercise 9.2.10

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Exercise 9.2.11

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Answer

a. $H_0 : p = 0.42$
b. $H_a : p < 0.42$

Exercise 9.2.12

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0:_____
b. H_a:_____

Exercise 9.2.13

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0:_____
b. H_a:_____

Answer

a. $H_0 : \mu = 15$
b. $H_a : \mu \neq 15$

Exercise

The National Institute 9.2.14 of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

1. H_0:_____
2. H_a:_____

Null and Alternative Hypotheses (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

11.3: Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject** H_0 when H_0 **is true (correct decision).**
2. The decision is to **reject** H_0 when H_0 **is true** (incorrect decision known as aType I error).
3. The decision is **not to reject** H_0 when, in fact, H_0 **is false** (incorrect decision known as a Type II error).
4. The decision is to **reject** H_0 when H_0 **is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

- $\alpha =$ 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.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The *Power of the Test* is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test. The following are examples of Type I and Type II errors.

Example 11.3.1: 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.)

Exercise 11.3.1

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Answer

- **Type I error**: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.
- **Type II error**: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Example 11.3.2

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 $= P(\text{Type I error})$.

β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Exercise 11.3.2

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Answer

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Example 11.3.3

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

- **Type I error**: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.
- **Type II error**: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Exercise 11.3.3

"Red tide" is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Answer

In this scenario, an appropriate null hypothesis would be H_0: the mean level of toxins is at most $800\mu g$, $H_0 : \mu_0 \leq 800\mu g$.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most $800\mu g$. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than $800\mu g$). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example 11.3.4

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

- **Type I**: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

- **Type II**: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Exercise 11.3.4

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Answer

Type I error: c

Type I error: b

Summary

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected. The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1-\beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

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

Glossary

Type 1 Error

The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

Contributors

- Template:ContribOpenStax

This page titled 11.3: Outcomes and the Type I and Type II Errors is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Outcomes and the Type I and Type II Errors (Exercises)

Exercise 9.3.5

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Answer

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

Exercise 9.3.6

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

Exercise 9.3.7

For Exercise 9.12, what are α and β in words?

Answer:

$\alpha =$ the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

$\beta =$ the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

Exercise 9.3.8

In words, describe $1 - \beta$ For Exercise 9.12.

Exercise 9.3.9

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Answer:

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Exercise 9.3.10

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

Exercise 9.3.11

The power of a test is 0.981. What is the probability of a Type II error?

Answer:

0.019

Exercise 9.3.12

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Exercise 9.3.13

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

Answer

0.998

Exercise 9.3.14

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

Outcomes and the Type I and Type II Errors (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

11.4: Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a single population mean, the distribution for the test is for *means*:

$$\bar{X} - N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \tag{11.4.1}$$

or

$$t_{df} \tag{11.4.2}$$

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

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

$$P' - N\left(p, \sqrt{\frac{p-q}{n}}\right) \tag{11.4.3}$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a Student's t-distribution (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a simple random sample that comes from a population that is approximately normally distributed. You use the sample standard deviation to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a binomial distribution which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1–p$.

Summary

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures

satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and
  $sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p) \mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Student's t-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t-distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

Contributors

- Template:ContribOpenStax

This page titled 11.4: Distribution Needed for Hypothesis Testing is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Distribution Needed for Hypothesis Testing (Exercises)

Exercise 9.4.1

Which two distributions can you use for hypothesis testing for this chapter?

Answer

A normal distribution or a Student's *t*-distribution

Exercise 9.4.2

Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.

Exercise 9.4.3

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

Answer

Use a Student's t-distribution

Exercise 9.4.4

A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Exercise 9.4.5

A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Answer

a normal distribution for a single population mean

Exercise 9.4.6

It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

Exercise 9.4.7

You are performing a hypothesis test of a single population mean using a Student's t-distribution. What must you assume about the distribution of the data?

Answer

It must be approximately normally distributed.

Exercise 9.4.8

You are performing a hypothesis test of a single population mean using a Student's t-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

Exercise 9.4.9

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

Answer

They must both be greater than five.

Exercise 9.4.10

You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

Exercise 9.4.11

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

Answer

binomial distribution

Distribution Needed for Hypothesis Testing (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

11.5: Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the null hypothesis). Then you gather sample data randomly. If the sample has properties that would be very *unlikely* to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an assumption—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the p-value. The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 11.5.1

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm. and the distribution of heights is normal.

- The null hypothesis could be $H_0 : \mu \leq 15$
- The alternate hypothesis is $H_a : \mu > 15$

The words **"is more than"** translates as a "$>$" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ **is known** ($\sigma = 0.5cm.$), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16.$$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

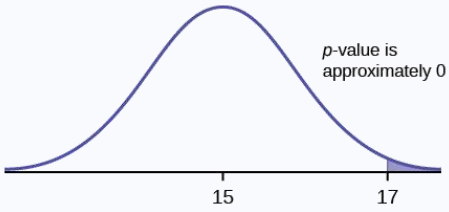


***Figure* 11.5.1**

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 11.5.1

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 $= 0.0013$

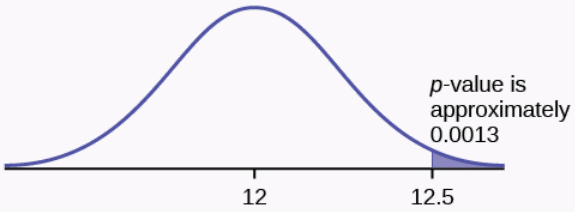


***Figure* 11.5.2**

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p-value and a preset or preconceived α (also called a "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a decision to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the alternative hypothesis, H_a, may be correct.
- If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.

When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses in terms of the given problem.

Example 11.5.2

When using the p-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

- If the p-value is low, the null must go.
- If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.

Answer

Reject the null hypothesis when **the p-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the p-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Exercise 11.5.2

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50, H_a : p > 0.50$
- $\alpha = 0.01$
- p-value $= 0.025$

Interpret the results and state a conclusion in simple, non-technical terms.

Answer

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support It's a Boy Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

Chapter Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the *p*-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

- $\alpha > p - value$, reject the null hypothesis
- $\alpha \leq p - value$, do not reject the null hypothesis

Glossary

Level of Significance of the Test

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

p-value

the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Contributors

- Template:ContribOpenStax

This page titled 11.5: Rare Events, the Sample, Decision and Conclusion is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Rare Events, the Sample, Decision and Conclusion (Exercises)

Exercise 9.5.3

When do you reject the null hypothesis?

Exercise 9.5.4

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

Answer

The outcome of winning is very unlikely.

Exercise 9.5.5

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

Exercise 9.5.6

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p-value is almost zero. State the null and alternative hypotheses and interpret the p-value.

Answer

$H_0 : \mu \geq 73$

$H_a : \mu \leq 73$

The p-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

Exercise 9.5.7

The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

Exercise 9.5.8

Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?

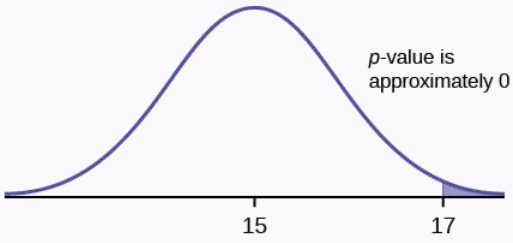


Figure 9.5.3.

Answer

The shaded region shows a low p-value.

Exercise 9.5.9

What should you do when $\alpha > p$-value?

Exercise 9.5.10

What should you do if $\alpha = p\text{-value}$?

Answer

Do not reject H_0.

Exercise 9.5.11

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Exercise 9.5.12

Is this a test of means or proportions?

Answer

means

Exercise 9.5.13

What symbol represents the random variable for this test?

Exercise 9.5.14

In words, define the random variable for this test.

Answer

the mean time spent in jail for 26 first time convicted burglars

Exercise 9.5.15

Is the population standard deviation known and, if so, what is it?

Exercise 9.5.16

Calculate the following:

a. $\bar{x}$ _____
b. σ _____
c. s_x _____
d. n _____

Answer

a. 3
b. 1.5
c. 1.8
d. 26

Exercise 9.5.17

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

Exercise 9.5.18

State the distribution to use for the hypothesis test.

Answer

$\bar{X} - N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

Exercise 9.5.19

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   $H0$: _____ H_a: _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:
   i. $\bar{x}$ = _____
   ii. s = _____
   iii. n = _____
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

Rare Events, the Sample, Decision and Conclusion (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

11.6: Additional Information and Full Hypothesis Test Examples

- In a hypothesis test problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use before collecting the sample data.
- If no level of significance is given, a common standard to use is $\alpha = 0.05$.
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The alternative hypothesis, H_a, tells you if the test is left, right, or two-tailed. It is the key to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- Thinking about the meaning of the p-value: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 ($\alpha = 0.05$ is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 11.6.1

$H_0 : \mu = 5, H_a : \mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:

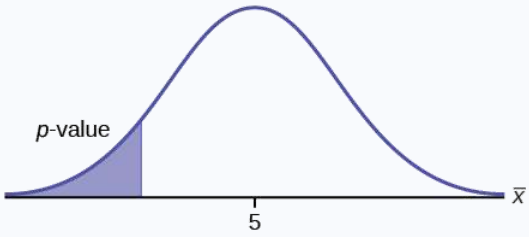


***Figure* 11.6.1**

Exercise 11.6.1

$H_0 : \mu = 10, H_a : \mu < 10$

Assume the p-value is 0.0935. What type of test is this? Draw the picture of the p-value.

Answer

left-tailed test

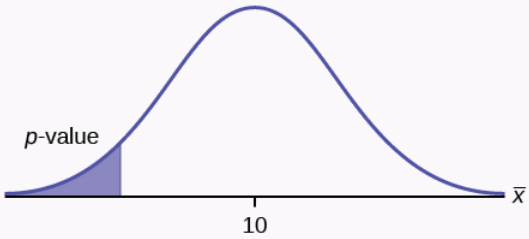


***Figure* 11.6.2**

Example 11.6.2

$H_0 : \mu \leq 0.2, H_a : \mu < 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:

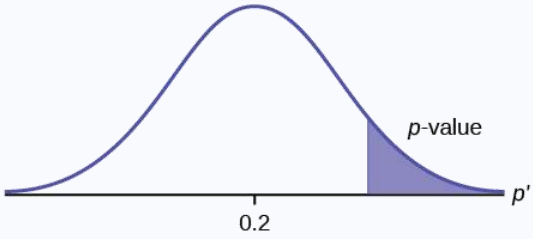


***Figure* 11.6.3**

Exercise 11.6.2

$H_0 : \mu \leq 1, H_a : \mu > 1$

Assume the p-value is 0.1243. What type of test is this? Draw the picture of the p-value.

Answer

right-tailed test

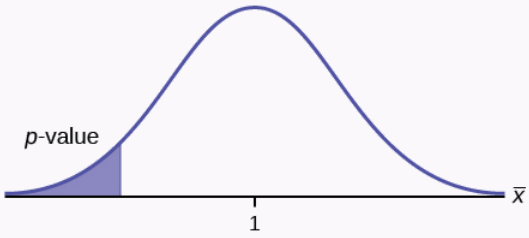


***Figure* 11.6.4**

Example 11.6.3

$H_0 : \mu = 50, H_a : \mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.

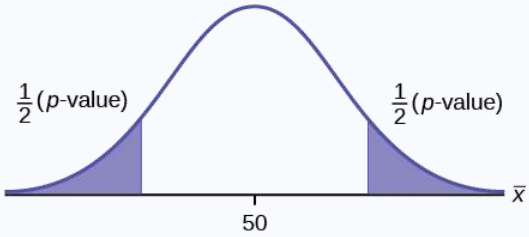


***Figure* 11.6.5**

Exercise 11.6.3

$H_0 : \mu = 0.5, H_a : \mu \neq 0.5$

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Answer

two-tailed test

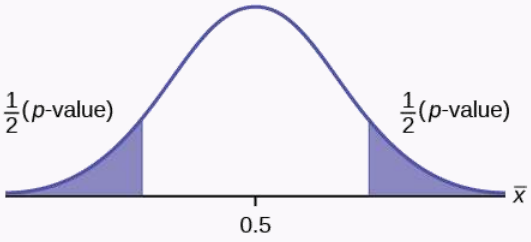


***Figure* 11.6.6**

Full Hypothesis Test Examples

Example 11.6.4

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.**Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$H_0 : \mu = 16.43, H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X} =$ the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

$\bar{X} - N\left(\mu, \frac{\sigma_x}{\sqrt{n}}\right)$ Therefore, $\bar{X} - N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the $p - \text{value}$ using the normal distribution for a mean:

$p\text{-value} = P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

$p\text{-value} = 0.0187$ (This is called the **actual level of significance**.) The $p - \text{value}$ is the area to the left of the sample mean is given as 16.

Graph:

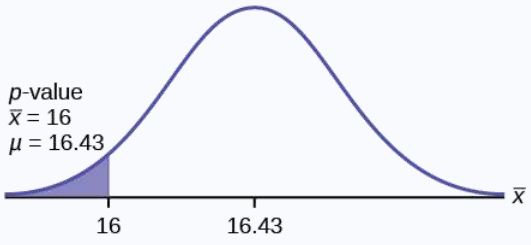


***Figure* 11.6.7**

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the $p-$value: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the $p-$value:

$\alpha = 0.05 p\text{-value} = 0.0187 \alpha > p\text{-value}$

Make a decision: Since $\alpha > p$-value, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

Press `STAT` and arrow over to `TESTS` . Press `1:Z-Test` . Arrow over to `Stats` and press `ENTER` . Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for *n*. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator not only calculates the *p*-value ($p = 0.0187$) but it also calculates the test statistic (*z*-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ ($p-$value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the *p*-value by doing a normal probability calculation using the central limit theorem:

$P(\bar{X} < 16)$ `2nd DISTR normcdf` $((-10^{99}, 16, 16.43, \frac{0.8}{\sqrt{15}})$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 11.6.4

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Press STAT and arrow over to TESTS. Press 1: Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as $<$, $\neq$, or $>$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (*z*-score) for the sample mean. Select $<$, $\neq$, or $>$ for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and *p*-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Answer

Since the problem is about a mean, this is a test of a single population mean.

- $H_0 : \mu = 40$
- $H_a : \mu > 40$

- $p = 0.0062$

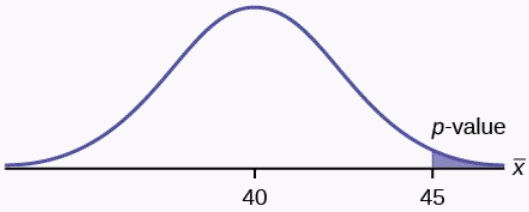


***Figure* 11.6.8**

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Historical Note

The traditional way to compare the two probabilities, α and the $p-$**value**, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x}-\mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$. For this problem, $\bar{x} = 16$, $\mu_x = 16.43$ from the null hypotheses is, $\sigma_x = 0.8$, and $n = 15$.) You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p-$**value**), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the $p-$**value**, 0.0187 is considerably smaller than \(\alpha = 0.05\). You can be confident about your decision to reject. The graph shows α, the $p-$**value**, and the test statistics and the critical value.

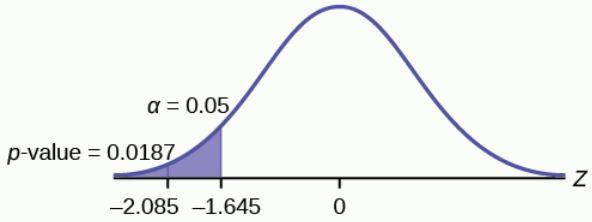


***Figure* 11.6.9**

Example 11.6.5

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Answer

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a test of a single population mean.

- $H_0 : \mu = 275$
- $H_a : \mu > 275$

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X} =$ the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

- $\bar{X} - N\left(275, \frac{55}{\sqrt{30}}\right)$
- $\bar{x} = 286.2$ pounds (from the data).
- $\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the *p*-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(\bar{x} > 286.2) = 0.1323.$$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

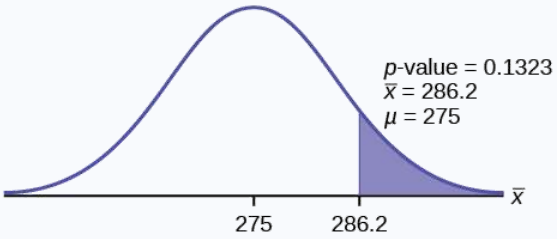


***Figure* 11.6.10**

Compare α and the $p - \text{value}$:

$\alpha = 0.025 p - value = 0.1323$

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The $p - \text{value}$ can easily be calculated.

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS` . Press `1:Z-Test` . Arrow over to `Data` and press `ENTER` . Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER` . Arrow down to `Calculate` and press `ENTER` . The calculator not only calculates the $p - \text{value}$) (\($p = 0.1331$), a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (*z*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER` . A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ ($p - \text{value}$). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 11.6.6

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Answer

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$H_0 : \mu = 65 \qquad H_a : \mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's t-distribution:

p-value $= P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

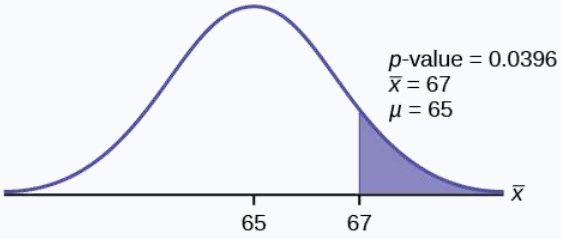


***Figure* 11.6.11**

Compare α and the \(p-\text{value}\):

Since \(α = 0.05\) and p-value $= 0.0396$. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with \(t = 1.9781\) (test statistic) and $p = 0.0396$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Exercise 11.6.6

It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : \mu = 5$
- $H_a : \mu < 5$

- $p = 0.0082$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

- Type I Error: To conclude that the stock price is growing slower than \$5 a week when, in fact, the stock price is growing at \$5 a week (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that the stock price is growing at a rate of \$5 a week when, in fact, the stock price is growing slower than \$5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example 11.6.7

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Answer

Set up the hypothesis test:

The 1% level of significance means that α = 0.01. This is a **test of a single population proportion**.

$H_0 : p = 0.50 \qquad H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $P' =$ the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

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

Therefore,

$$P' - N\left(0.5, \sqrt{\frac{0.5-0.5}{100}}\right)$$

where $p = 0.50, q = 1 - p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = P(p' < 0.47 or p' > 0.53) = 0.5485$$

where

$$x = 53, p' = \frac{x}{n} = \frac{53}{100} = 0.53$$

.

Interpretation of the \(p\text{-value})\: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure).

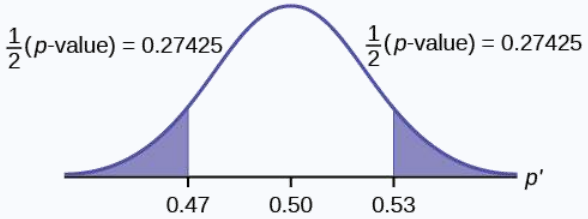


Figure 11.6.12

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50–0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value $= 0.5485$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

Press STAT and arrow over to TESTS. Press 5:1-PropZTest. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to Prop and arrow to not equals p_0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). Prop not equals .5 is the alternate hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with \(z = 0.6\) (test statistic) and $p = 0.5485$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Exercise 11.6.7

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Answer

Since the problem is about percentages, this is a test of single population proportions.

- $H_0 : p = 0.85$
- $H_a : p \neq 0.85$
- $p = 0.7554$

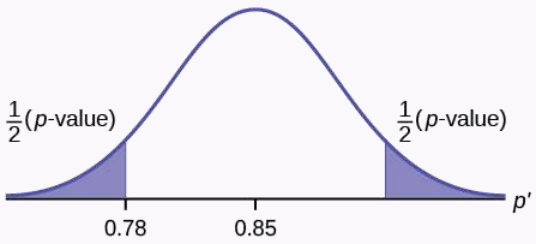


Figure 11.6.13

Because $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example 11.6.8

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Answer

Set up the Hypothesis Test:

$H_0 : p = 0.30, H_a : p \neq 0.30$

Determine the distribution needed:

The **random variable** is $P' =$ proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' - N\left(0.30, \sqrt{\frac{(0.30 \cdot 0.70)}{150}}\right)$

Exercise 9.6.8.2

a. The value that helps determine the p-value is p'. Calculate p'.

Answer

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43, n = 150$

$p' = 43150$

Exercise 9.6.8.3

b. What is a **success** for this problem?

Answer

b. A success is having three cell phones in a household.

Exercise 9.6.8.4

c. What is the level of significance?

Answer

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

Exercise 9.6.8.5

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the p-value.

Answer

d. p-value $= 0.7216$

Exercise 9.6.8.6

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Answer

e. Assuming that $\alpha = 0.05, \alpha < p$-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Exercise 11.6.8

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State

the null and alternative hypothesis, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Answer

- $H_0 : p = 0.92$
- $H_a : p < 0.92$
- p-value $= 0.0046$

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

- Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).
- Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example 11.6.9

My dog has so many fleas,

They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!
Before his shampoo
I counted 42.

At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out

Use the new shampoo or go without?

Answer

Set up the hypothesis test:

$H_0 : p \leq 0.25 \qquad H_a : p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\bar{X}$ or P' represents.

$P'=$ The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)1-0.25}{42}}\right)$$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

$$p\text{-value} = 0.0103$$

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

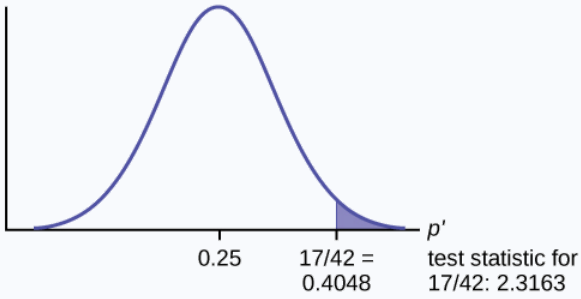


***Figure* 11.6.14**

Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

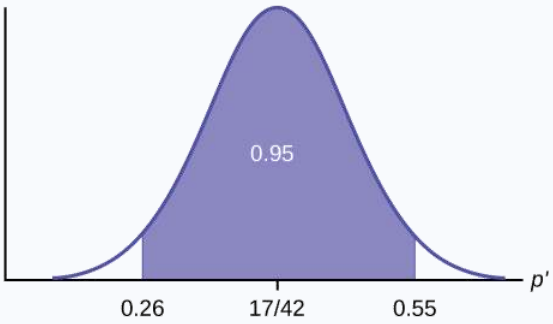


***Figure* 11.6.15**

Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

This test result is not very definitive since the p-value *is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.*

Example 11.6.10

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Answer

Let's follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. $H_0 : \mu \leq 1$
   b. $H_a : \mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

Figure 9.6.7.

Figure 9.6.8.

Figure 9.6.9.

Figure 9.6.10.

4. **State the Conclusions**: Since the p-value$^*(p = 0.036)$ is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example 11.6.11

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Answer

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. $H_0 : p \leq 0.00034$
   b. $H_a : p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420,019$. The sample is sufficiently large because we have $np = 420,019(0.00034) = 142.8$, $nq = 420,019(0.99966) = 419,876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are

   Figure 9.6.11.

   Figure 9.6.12.
4. Since the p-value $= 0.0073$ is greater than our alpha value $= 0.005$, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example 11.6.12

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Answer

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion z-test. The hypotheses for the test will be
   a. $H_0 : p = 0.00078$
   b. $H_a : p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

   Figure 9.6.13.

   Figure 9.6.14.
4. Since the p-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County,

Kentucky is different from the national average proportion.

Chapter Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1-\beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

Exercise 9.6.8

Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.6.9

Assume $H_0 : \mu \leq 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.6.10

Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

Exercise 9.6.11

Draw the general graph of a left-tailed test.

Exercise 9.6.12

Draw the graph of a two-tailed test.

Answer

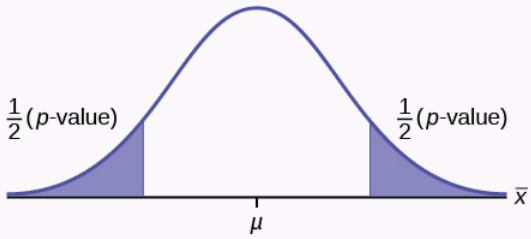


***Figure* 11.6.16**

Exercise 9.6.13

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise 9.6.14

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

Answer

a right-tailed test

Exercise 9.6.15

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

Exercise 9.6.16

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

Answer

a left-tailed test

Exercise 9.6.17

If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

Exercise 9.6.18

Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a left-tailed test.

Exercise 9.6.19

Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise 9.6.20

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Answer

This is a two-tailed test.

References

1. Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.
2. Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.
3. Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).
4. Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).
5. Data from *Growing by Degrees* by Allen and Seaman.
6. Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.
7. Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).
8. Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).
9. Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.
10. Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)
11. Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).
12. Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

13. Data from Toastmasters International. Available online at http://toastmasters.org/artisan/deta...eID=429&Page=1.
14. Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).
15. Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).
16. “Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DA...t_da_2006w.pdf.
17. Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).
18. Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information...sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, $\sum X$. If the size n of the sample is sufficiently large, then $\bar{X} - N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\sum X - N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Contributors

- Template:ContribOpenStax

This page titled 11.6: Additional Information and Full Hypothesis Test Examples is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)

This page titled 11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

11.8: Hypothesis Testing with One Sample (Exercises)

This page titled 11.8: Hypothesis Testing with One Sample (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

12: Hypothesis Testing with Two Samples

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded. To compare two means or two proportions, you work with two groups. The groups are classified either as independent or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

Topic hierarchy

12.1: Prelude to Hypothesis Testing with Two Samples
12.2: Two Population Means with Unknown Standard Deviations
12.3: Two Population Means with Known Standard Deviations
12.4: Comparing Two Independent Population Proportions
12.5: Matched or Paired Samples
12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet)
12.7: Hypothesis Testing with Two Samples (Exercises)

Contributors

- Template:ContribOpenStax

This page titled 12: Hypothesis Testing with Two Samples is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.1: Prelude to Hypothesis Testing with Two Samples

Skills to Develop

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

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.



Figure 10.1.1. If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or matched pairs. Independent groups consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and *p*-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Test of the two population proportions by testing one population mean of differences.

Contributors

- Template:ContribOpenStax

This page titled 12.1: Prelude to Hypothesis Testing with Two Samples is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.2: Two Population Means with Unknown Standard Deviations

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:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{12.2.1}$$

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

$$\frac{(\bar{x} - \bar{x}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{12.2.2}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_1, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means. μ_1 and μ_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

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{12.2.3}$$

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 12.2.1: 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 **12.2.1**. Each populations has a normal distribution.

Table **12.2.1**

	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, μ_g is the population mean for girls and μ_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:

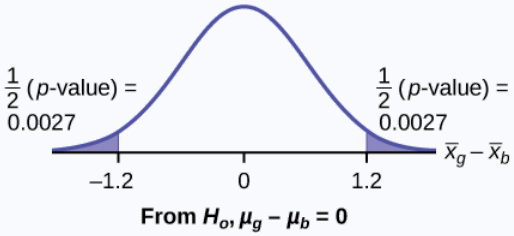


Figure* 12.2.1*: *Normal distribution curve representing the difference in the average amount of time girls and boys play sports all day*

$$s_g = 0.866 \tag{12.2.4}$$

$$s_b = 1 \tag{12.2.5}$$

So,

$$\bar{x}_g - \bar{x}_b = 2 - 3.2 = -1.2 \tag{12.2.6}$$

Half the $p\text{-value}$ 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 $p\text{-value}$ 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).

Exercise 10.2.1

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 **12.2.2**

	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 12.2.2

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.

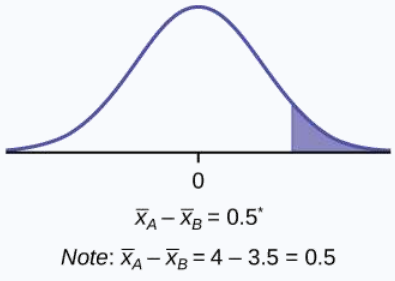


Figure 10.2.2.

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.

Exercise 12.2.2

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

Answer

a. They are unknown.
b. The p-value $= 0.0878$. At the 5% level of significance, there is insufficient evidence to conclude that the workers of Company A stay longer with the company.

Example 12.2.3

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 12.2.3 and Table 12.2.4.

Table **12.2.3***: 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 **12.2.4***: 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 $= 0.0011$

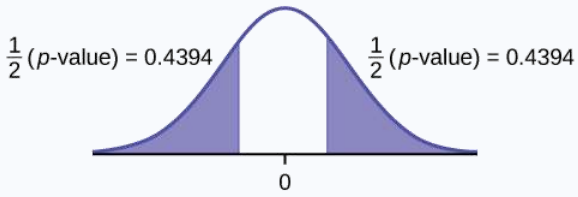


Figure 10.2.3.

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 μ_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 12.2.5: *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 12.2.4

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

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

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

$$d = 0.384$$

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 12.2.5

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.

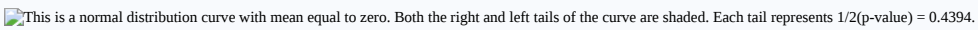


Figure 10.2.4.

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 http://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 http://en.wikipedia.org/wiki/List_of...enators_by_age (accessed June 17, 2013).
5. "Sectoring by Industry Groups." Nasdaq. Available online at http://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).

Chapter 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:

$$SE = \sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}} \tag{12.2.7}$$

Test statistic (*t*-score):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \tag{12.2.8}$$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2} \tag{12.2.9}$$

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:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{12.2.10}$$

where

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{12.2.11}$$

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 σ 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 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

- Template:ContribOpenStax

This page titled 12.2: Two Population Means with Unknown Standard Deviations is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.3: 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:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{12.3.1}$$

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \tag{12.3.2}$$

The test statistic (z-score) is:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{12.3.3}$$

Example 12.3.1

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 12.3.1, the distribution is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right) \tag{12.3.4}$$

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:

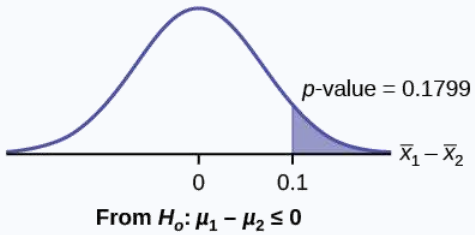


Figure 10.3.1.

$\bar{X}_1 - \bar{X}_2 = 3–2.9 = 0.1$

Compare α and the p-value: $\alpha = 0.05$ and p-value $= 0.1799$. 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 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 12.3.1

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 12.3.2: 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:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}\right] \tag{12.3.5}$$

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 $p\text{-value} = 0.4040$)

Graph:

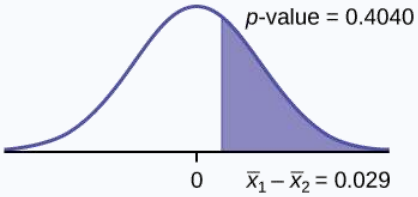


Figure 10.3.2.

Compare α and the p-value: $\alpha = 0.05$ and $p\text{-value} = 0.4040$. 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 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 http://www.census.gov/prod/cen2010/b...c2010br-02.pdf
2. Hinduja, Sameer. “Sexting Research and Gender Differences.” Cyberbulling Research Center, 2013. Available online at http://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 http://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).

Chapter Review

- A hypothesis test of two population means from independent samples where the population standard deviations are known (typically approximated with the sample standard deviations), will have these characteristics:
- Random variable: $\bar{X}_1 - \bar{X}_2 =$ the difference of the means
- Distribution: normal distribution

Formula Review

Normal Distribution:

$$\bar{X}_1 - \bar{X}_2 =\sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{12.3.6}$$

Generally $\mu_1 - \mu_2 = 0$.

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{12.3.7}$$

Generally $\mu_1 - \mu_2 = 0$.

where:

σ_1 and σ_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. μ_1 and μ_2 are the population means

Contributors

- Template:ContribOpenStax

This page titled 12.3: Two Population Means with Known Standard Deviations is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.4: Comparing 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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{12.4.1}$$

The distribution for the differences is:

$$P\_A - P_b' N \left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{12.4.2}$$

The test statistic (z-score) is:

$$z = \frac{(p_A' - p_B') - (p_A - p_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{12.4.3}$$

Example 12.4.1

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:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 8001 - p_c = 0.92 \tag{12.4.4}$$

$$P'_A - P'_B - N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right] \quad (12.4.5)$$

$P'_A - P'_B$ 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:

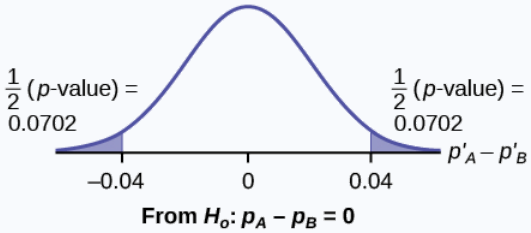


Figure 10.4.1.

$P'_A - P'_B = 0.1–0.06 = 0.04$.

Half the p-value is below -0.04, and half is above 0.04.

Compare α 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 12.4.1

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 12.4.2: 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:

$$p_C = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077 \tag{12.4.6}$$

$$1 - p_C = 0.923 \tag{12.4.7}$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right) \tag{12.4.8}$$

$p'_F - p'_M$ 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:

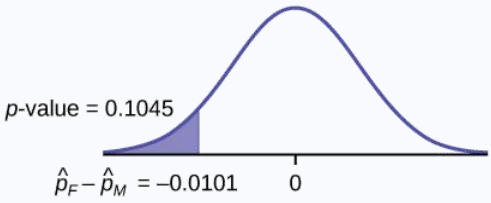


Figure 10.4.2.

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 12.4.3

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:

$$p_C = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927 \tag{12.4.9}$$

$$1 - p_C = 0.9073 \tag{12.4.10}$$

Therefore,

$$p'_W - p'_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right) \tag{12.4.11}$$

$p'_W - p'_A$ 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:

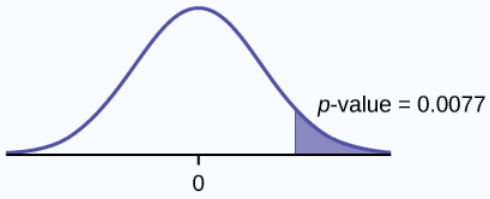


Figure 10.4.3.

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 12.4.3

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 $= 0.00086$

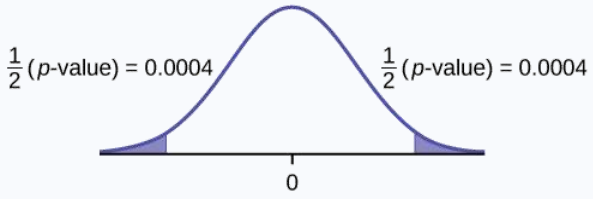


Figure 10.4.4.

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 http://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 http://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).

Chapter 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:

$$p_c = \frac{x_F + x_M}{n_F + n_M} \tag{12.4.12}$$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{12.4.13}$$

where the null hypothesis is $H_0 : p_A = p_B$ or $H_0 : p_A - p_B = 0$.

Test Statistic (z-score):

$$z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{12.4.14}$$

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

- Template:ContribOpenStax

This page titled 12.4: Comparing Two Independent Population Proportions is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.5: 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, μ_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:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{12.5.1}$$

Example 12.5.1

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 μ_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

$$H_0 : \mu_d \geq 0 \tag{12.5.2}$$

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. μ_d is the population mean of the differences.

$$H_a : \mu_d < 0 \tag{12.5.3}$$

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:

$$p\text{-value} = 0.0095 \tag{12.5.4}$$

Graph:

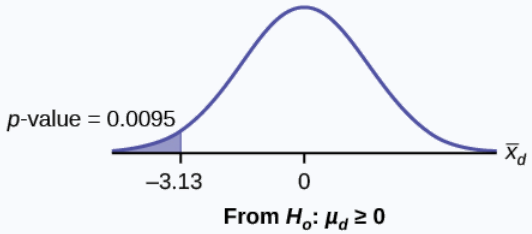


Figure 10.5.1.

$\bar{X}_d$ 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 α 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 μ_0, the name of the list where you put the data, and `1` for Freq:.

Arrow down to μ: and arrow over to $<$ μ_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 12.5.1

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 12.5.2

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.

$$\bar{x}_d = 21.3 \tag{12.5.5}$$

and

$$s_d = 46.7 \tag{12.5.6}$$

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:

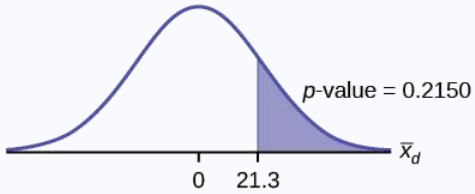


Figure 10.5.2.

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 12.5.2

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 12.5.3

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:

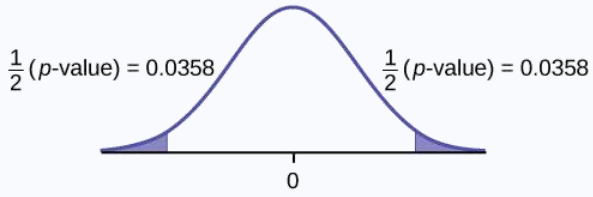


Figure 10.5.3.

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 12.5.3

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.

Chapter 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):

$$t = \frac{\bar{x}_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{12.5.7}$$

where:

$\bar{x}_d$ is the mean of the sample differences. μ_d is the mean of the population differences. s_d is the sample standard deviation of the differences. n is the sample size.

Contributors

- Template:ContribOpenStax

This page titled 12.5: Matched or Paired Samples is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet)

This page titled 12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

12.7: Hypothesis Testing with Two Samples (Exercises)

This page titled 12.7: Hypothesis Testing with Two Samples (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

13: 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.

13.1: Prelude to The Chi-Square Distribution
13.2: Facts About the Chi-Square Distribution
13.3: Goodness-of-Fit Test
13.4: Test of Independence
13.5: Test for Homogeneity
13.6: Comparison of the Chi-Square Tests
13.7: Test of a Single Variance
13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet)
13.9: Lab 2: Chi-Square Test of Independence (Worksheet)
13.10: The Chi-Square Distribution (Exercises)

Template:ContribOpenStax

This page titled 13: The Chi-Square Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.1: Prelude to The Chi-Square Distribution

CHAPTER OBJECTIVES

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

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests.

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.



Figure* 13.1.1*: *The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)*

In this chapter, you will learn the three major applications of the chi-square distribution:

a. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
b. the test of independence, which determines if events are independent, such as in the movie example
c. the test of a single variance, which tests variability, such as in the coffee example

Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see [link]). TI-83+ and TI-84 calculator instructions are included in the text.

COLLABORATIVE CLASSROOM EXERCISE

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

This page titled 13.1: Prelude to The Chi-Square Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.2: Facts About the Chi-Square Distribution

The notation for the chi-square distribution is:

$$\chi \sim \chi^2_{df} \tag{13.2.1}$$

where df = degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is

$$\sigma = \sqrt{2(df)}. \tag{13.2.2}$$

The random variable is shown as χ^2, but may be any upper case letter. The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + \ldots + (Z_k)^2 \tag{13.2.3}$$

a. The curve is nonsymmetrical and skewed to the right.
b. There is a different chi-square curve for each df.

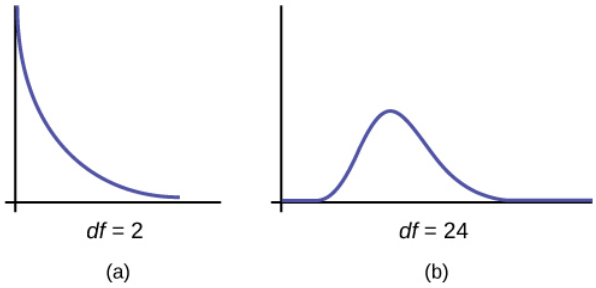


***Figure* 13.2.1**

c. The test statistic for any test is always greater than or equal to zero.
d. When $df > 90$, the chi-square curve approximates the normal distribution. For $\chi \sim \chi^2_{1,000}$ the mean, $\mu = df = 1,000$ and the standard deviation, $\mu = \sqrt{2(1,000)}$. Therefore, $X \sim N(1,000, 44.7)$, approximately.
e. The mean, μ, is located just to the right of the peak.

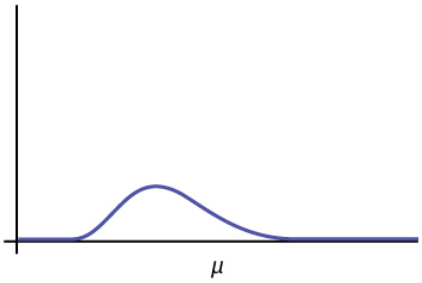


***Figure* 13.2.2**

References

1. Data from *Parade Magazine*.
2. “HIV/AIDS Epidemiology Santa Clara County.”Santa Clara County Public Health Department, May 2011.

Chapter 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

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + ... + (Z_{df})^2 \quad (13.2.4)$$

chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2(df)}$ Chi-Square distribution population standard deviation

Exercise 11.2.1

If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

Answer

mean $= 25$ and standard deviation $= 7.0711$

Exercise 11.2.2

If $df > 90$, the distribution is _____. If $df = 15$, the distribution is _____.

Exercise 11.2.3

When does the chi-square curve approximate a normal distribution?

Answer

when the number of degrees of freedom is greater than 90

Exercise 11.2.4

Where is μ located on a chi-square curve?

Exercise

Is it more likely the *df* is 90, 20, or two in the graph?

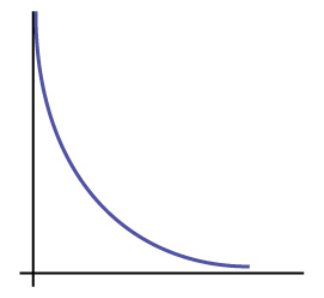


Figure 11.2.3.

Answer

$df = 2$

This page titled 13.2: Facts About the Chi-Square Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.3: Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data "fit" a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is:

$$\sum_{k} \frac{(O-E)^2}{E} \tag{13.3.1}$$

where:

- $O =$ observed values (data)
- $E =$ expected values (from theory)
- $k =$ the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is $df = (\text{number of categories} - 1)$.

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

The expected value for each cell needs to be at least five in order for you to use this test.

Example 11.3.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to 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.

Exercise 11.3.1.1

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Answer

a. **No.** Notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in Table and Table.

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

Exercise 11.3.1.2

b. What is the number of degrees of freedom (df)?

Answer

b. There are four "cells" or categories in each of the new tables.

$df = \text{number of cells} - 1 = 4 - 1 = 3$

Exercise 11.3.1

how many are produced. The number of expected defects is listed in Table.

Number produced	Number defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

A random sample was taken to determine the actual number of defects. Tableshows the results of the survey.

Number produced	Number defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Answer

H_0:The number of defaults fits expectations.

H_a:The number of defaults does not fit expectations.

$df = 4$

Example 11.3.2

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in 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 χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values $(12, 12, 12, 12, 12)$
- Observed (O) values $(15, 12, 9, 9, 15)$
- $(O–E)$
- $(O–E)^2$
- \(\frac{(O - E)^{2}}{E})

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get $p\text{-value} = 0.5578$.)

The dfs are the $\text{number of cells} - 1 = 5 - 1 = 4$

Press 2nd DISTR . Arrow down to \(\chi^{2}\)cdf . Press ENTER . Enter \((3,10^99,4)\) . Rounded to four decimal places, you should see 0.5578, which is the p-value.

Next, complete a graph like the following one with the proper labeling and shading. (You should shade the right tail.)

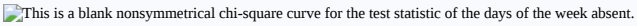


Figure 11.3.1.

The decision is not to reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example Example has the calculator instructions. The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF . Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw . Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into STAT EDIT and arrow up to the list name area of the particular list. Press CLEAR and then arrow down. The list will be cleared. Alternatively, you can press STAT and press 4 (for ClrList). Enter the list name and press ENTER .

Exercise 11.3.2

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students were asked on which night of the week they did the most homework. The results were distributed as in Table.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Answer

$df = 6$

p-value $= 0.6093$

We decline to reject the null hypothesis. There is not enough evidence to support that students do not do the majority of their homework equally throughout the week.

Example 11.3.3

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as in Table.

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the data in Table.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

Exercise 11.3.3.1

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Answer

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are shown in Table.

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
over 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter $0.10 * 600$.

H_0: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where \(df = (\text{the number of cells}) - 1 = 5 - 1 = 4\).

Note 11.3.3.1

$df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:

This is a nonsymmetric chi-square curve with values of 0, 4, and 29.65 labeled on the horizontal axis. The value 4 coincides with the peak of the curve. A vertical upward line extends from 29.65 to the curve, and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.3.2.

Probability statement: $p\text{-value} = P(\chi^2 > 29.65) = 0.000006$

Compare α and the p-value:

$\alpha = 0.01$

$p\text{-value} = 0.000006$

So, $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600, .16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:χ2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to four places, you should see `5.77E-6 = .000006` (rounded to six decimal places), which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Exercise 11.3.3

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in Table.

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in Table.

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, does it appear that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the p-value?

Answer

p-value $= 0.0036$

We reject the null hypothesis that the distributions are the same. There is sufficient evidence to conclude that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole.

Example 11.3.4

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5% significance level.

Answer

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is \({*HH*, *HT*, *TH*, *TT*}\). Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (\(20 *HH*, 27 *HT*, 30 *TH*, 23 *TT*\)) fit the expected distribution?"

Random Variable: Let $X =$ the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the **number of cells is three**. Since $X =$ the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

H_0: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



This is a nonsymmetrical chi-square curve with values of 0 and 2.14 labeled on the horizontal axis. A vertical upward line extends from 2.14 to the curve and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.3.3.

Probability statement: p-value $= P(\chi^2 > 2.14) = 0.3430$

Compare α and the p-value:

$\alpha = 0.05$

p-value $= 0.3430$

$\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Press STAT and ENTER . Make sure you clear lists L1 , L2 , and L3 if they have data in them. Into L1 , put the observed frequencies 20 , 57 , 23 . Into L2 , put the expected frequencies 25 , 50 , 25 . Arrow over to list L3 and up to the name area "L3" . Enter (L1-L2)^2/L2 and ENTER . Press 2nd QUIT . Press 2nd LIST and arrow over to MATH . Press 5 . You should see "sum" . Enter L3 . Rounded to two decimal places, you should see 2.14 . Press 2nd DISTR . Arrow down to 7:χ2cdf (or press 7). Press ENTER . Enter 2.14,1E99,2) . Rounded to four places, you should see .3430 , which is the p-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press STAT TESTS and Chi2 GOF . Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw . Make sure you clear any lists before you start.

Exercise 11.3.4

Students in a social studies class hypothesize that the literacy rates across the world for every region are 82%. Table shows the actual literacy rates across the world broken down by region. What are the test statistic and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed Regions	99.0
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91.0
Eastern Asia	93.8
Southern Asia	61.9
South-Eastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Answer

$df = 9$

χ^2 \text{test statistic} = 26.38\)

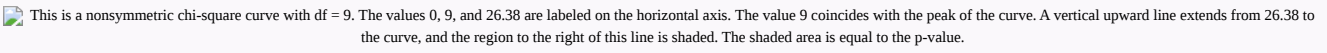

This is a nonsymmetric chi-square curve with df = 9. The values 0, 9, and 26.38 are labeled on the horizontal axis. The value 9 coincides with the peak of the curve. A vertical upward line extends from 26.38 to the curve, and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.3.4.

Press STAT and ENTER . Make sure you clear lists L1, L2, and L3 if they have data in them. Into L1, put the observed frequencies 99, 99.5, 67.3, 62.5, 91, 93.8, 61.9, 91.9, 84.5, 66.4 . Into L2 , put the expected frequencies 82, 82, 82, 82, 82, 82, 82, 82, 82, 82 . Arrow over to list L3 and up to the name area "L3" . Enter (L1-L2)^2/L2 and ENTER . Press 2nd QUIT . Press 2nd LIST and arrow over to MATH . Press 5 . You should see "sum" . Enter L3 . Rounded to two decimal places, you should see 26.38 . Press 2nd DISTR . Arrow down to 7:χ2cdf (or press 7). Press ENTER . Enter 26.38,1E99,9) . Rounded to four places, you should see .0018 , which is the *p*-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF . To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press calculate or draw. Make sure you clear any lists before you start.

References

1. Data from the U.S. Census Bureau
2. Data from the College Board. Available online at http://www.collegeboard.com.
3. Data from the U.S. Census Bureau, Current Population Reports.
4. Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I.S. Ockene, "Association between Eating Patterns and Obesity in a Free-living US Adult Population." *American Journal of Epidemiology* volume 158, no. 1, pages 85-92.
5. Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, "Prevalence of Obesity in the United States, 2009–2010." NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).
6. Stevens, Barbara J., "Multi-family and Commercial Solid Waste and Recycling Survey." Arlington Count, VA. Available online at http://www.arlingtonva.us/department.../file84429.pdf (accessed May 24,2013).

Chapter Review

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_k \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where:

O: observed values

E: expected value

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

Determine the appropriate test to be used in the next three exercises.

Exercise 11.3.5

An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

Exercise 11.3.6

An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

Answer

a goodness-of-fit test

Exercise 11.3.7

A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises: A teacher predicts that the distribution of grades on the final exam will be and they are recorded in Table.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 20 is in Table.

Grade	Frequency
A	7
B	7
C	5
D	1

Exercise 11.3.8

$df =$ _____

Answer

3

Exercise 11.3.9

State the null and alternative hypotheses.

Exercise 11.3.10

χ^2test statistic = _____

Answer

2.04

Exercise 11.3.11

p-value = _____

Exercise 11.3.12

At the 5% significance level, what can you conclude?

Answer

We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in Table.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is as in Table.

Ethnicity	Percentage of total county population	Number expected (round to two decimal places)
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Exercise 11.3.13

If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.

Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.3.14

H_0: _____

Answer

H_0: the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise 11.3.15

H_a: _____

Exercise 11.3.16

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

right-tailed

Exercise 11.3.17

degrees of freedom = _____

Exercise 11.3.18

χ^2test statistic = _____

Answer

88,621

Exercise 11.3.19

p-value = _____

Exercise 11.3.20

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



Figure 11.3.5.

Let $\alpha = 0.05$

Decision: _____

Reason for the Decision: _____

Conclusion (write out in complete sentences): _____

Answer

Graph: Check student's solution.

Decision: Reject the null hypothesis.

Reason for the Decision: p-value $< \alpha$

Conclusion (write out in complete sentences): The make-up of AIDS cases does not fit the ethnicities of the general population of Santa Clara County.

Exercise 11.3.21

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

This page titled 13.3: Goodness-of-Fit Test is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.4: Test of Independence

Tests of independence involve using a contingency table of observed (data) values.

The test statistic for a *test of independence* is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{13.4.1}$$

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 $\frac{(O-E)^2}{E}$.

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.

The expected value for each cell needs to be at least five in order for you to use this test.

Example 11.4.1

Suppose $A =$ a speeding violation in the last year and $B =$ a cell phone user while driving. If A and B are independent then $P(A \text{ AND } B) = P(A)P(B)$. $A \text{ AND } B$ is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell phone while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let $y =$ expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right) \tag{13.4.2}$$

Solve for y : $y = \frac{(70)(305)}{755} = 28.3$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternative hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example, then the null hypothesis is:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

$df = (\text{number of columns} - 1)(\text{number of rows} - 1)$

The following formula calculates the **expected number** (E):

$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$

Exercise 11.4.1

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Answer

About 16 students are expected to be music students and on the honor roll.

Example 11.4.2

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 is a **sample** of the adult volunteers and the number of hours they volunteer per week.

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 result are in Table.

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

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57 \quad (13.4.3)$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$df = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$

Graph:


Nonsymmetrical chi-square curve with values of 0 and 12.99 on the x-axis representing the test statistic of number of hours worked by volunteers of different types. A vertical upward line extends from 12.99 to the curve and the area to the right of this is equal to the p-value.

Figure 11.4.1.

Probability statement: $p\text{-value} = P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0113$. $\alpha > p\text{-value}$.

Make a decision: Since $\alpha > p\text{-value}$, reject H_0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the example in Table, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Press the `MATRX` key and arrow over to `EDIT` . Press `1:[A]` . Press `3 ENTER 3 ENTER` . Enter the table values by row from Table. Press `ENTER` after each. Press `2nd QUIT` . Press `STAT` and arrow over to `TESTS` . Arrow down to `C:\(\chi^{2}\)-TEST` . Press `ENTER` . You should see `Observed:[A] and Expected:[B]` . Arrow down to `Calculate` . Press `ENTER` . The test statistic is 12.9909 and the $p\text{-value} = 0.0113$. Do the procedure a second time, but arrow down to `Draw` instead of `calculate` .

Exercise 11.4.2

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table shows the results:

Industry Sector	2000	2010	2020	Total
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Answer

- H_0: The number of jobs is independent of the year.
- H_a: The number of jobs is dependent on the year.

$df = 12$

Figure 11.4.2.

Press the MATRX key and arrow over to EDIT . Press 1:[A] . Press 3 ENTER 3 ENTER . Enter the table values by row. Press ENTER after each. Press 2nd QUIT . Press STAT and arrow over to TESTS . Arrow down to C:\(\chi^{2}\)-TEST . Press ENTER . You should see Observed:[A] and Expected:[B] . Arrow down to Calculate . Press ENTER . The test statistic is 227.73 and the p-value $= 5.90E-42 = 0$. Do the procedure a second time but arrow down to Draw instead of calculate .

Example 11.4.3

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.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09 \quad (13.4.4)$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

d. 8

Exercise 11.4.3

Refer back to the information in Note. How many service providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

Answer

12,727, 14,965

References

1. DiCamilo, Mark, Mervin Field, “Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs.” The Field Poll, released Feb. 14, 2013. Available online at http://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).

Chapter Review

To assess whether two factors are independent or not, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least 5.

Formula Review

Test of Independence

- The number of degrees of freedom is equal to $(\text{number of columns - 1})(\text{number of rows - 1})$.
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where $O =$ observed values, $E =$ expected values, $i =$ the number of rows in the table, and $j =$ the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Determine the appropriate test to be used in the next three exercises.

Exercise 11.4.4

A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

Answer

a test of independence

Exercise 11.4.5

The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

Exercise 11.4.6

A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.

Answer

a test of independence

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. Table shows the results. The railroad wants to know if a passenger’s choice in ticket class is independent of the distance they must travel.

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

Exercise 11.4.7

State the hypotheses.

H_0: _____

H_a: _____

Exercise 11.4.8

df = _____

Answer

8

Exercise 11.4.9

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

Exercise 11.4.10

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

Answer

6.6

Exercise 11.4.11

What is the test statistic?

Exercise 11.4.12

What is the p-value?

Answer

0.0435

Exercise 11.4.13

What can you conclude at the 5% level of significance?

Use the following information to answer the next eight exercises: An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650 whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

Exercise 11.4.14

Complete the table.

Smoking Levels by Ethnicity (Observed)

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Answer

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1-10	9,886	2,745	12,831	8,378	7,650	41,490
11-20	6,514	3,062	4,932	10,680	9,877	35,065
21-30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

Exercise 11.4.15

State the hypotheses.

H_0: _____

H_a: _____

Exercise 11.4.16

Enter expected values in Table. Round to two decimal places.

Calculate the following values:

Answer

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7777.57	3310.11	8248.02	10771.29	11383.01
11-20	6573.16	2797.52	6970.76	9103.29	9620.27
21-30	2863.02	1218.49	3036.20	3965.05	4190.23
31+	1616.25	687.87	1714.01	2238.37	2365.49

Exercise 11.4.17

df = _____

Exercise 11.4.18

χ^2test statistic = _____

Answer

10,301.8

Exercise 11.4.19

p-value = _____

Exercise 11.4.20

Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Answer

right

Exercise 11.4.21

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



Figure 11.4.3.

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

Exercise 11.4.22

$\alpha = 0.05$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

Answer

a. Reject the null hypothesis.
b. p-value $< \alpha$
c. There is sufficient evidence to conclude that smoking level is dependent on ethnic group.

Exercise 11.4.23

$\alpha = 0.05$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

Glossary

Contingency Table

a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

This page titled 13.4: Test of Independence is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.5: 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 χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of Freedom (df)

- $df = \text{number of columns} - 1$

Requirements

- All values in the table must be greater than or equal to five.

Common Uses

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Example 13.5.1

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 **13.5.1**. Do male and female college students have the same distribution of living arrangements?

***Table* 13.5.1:** *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: χ^2_3

Calculate the test statistic: $\chi^2 = 10.1287$ (calculator or computer)

Probability statement: $p\text{-value} = P(\chi^2 > 10.1287) = 0.0175$

Press the

```
MATRX
```

key and arrow over to

```
EDIT
```

. Press

```
1:[A]
```

. Press

```
2 ENTER 4 ENTER
```

. Enter the table values by row. Press

```
ENTER
```

after each. Press

```
2nd QUIT
```

. Press

```
STAT
```

and arrow over to

```
TESTS
```

. Arrow down to

```
C:χ2-TEST
```

. Press

```
ENTER
```

. You should see

```
Observed:[A] and Expected:[B]
```

. Arrow down to

```
Calculate
```

. Press

```
ENTER
```

. The test statistic is 10.1287 and the $p\text{-value} = 0.0175$. Do the procedure a second time but arrow down to

```
Draw
```

instead of

```
calculate
```

.

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. $p\text{-value} = 0.0175$. $\alpha > p\text{-value}$.

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.

Exercise 11.5.1

Do families and singles have the same distribution of cars? Use a level of significance of 0.05. Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in Table **13.5.2**. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

Table **13.5.1**

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Answer

With a $p\text{-value}$ of almost zero, we reject the null hypothesis. The data show that the distribution of cars is not the same for families and singles.

Example 11.5.2

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. Table shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Answer

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of Freedom (df):

df = number of columns $- 1 = 3 - 1 = 2$

Distribution for the test: χ^2_2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator or computer)

Probability statement: p-value $= P(\chi^2 > 3.2603) = 0.1959$

Press the MATRX key and arrow over to EDIT . Press 1:[A] . Press 2 ENTER 3 ENTER . Enter the table values by row. Press ENTER after each. Press 2nd QUIT . Press STAT and arrow over to TESTS . Arrow down to C:χ2-TEST . Press ENTER . You should see Observed:[A] and Expected:[B] . Arrow down to Calculate . Press ENTER . The test statistic is 3.2603 and the p-value = 0.1959. Do the procedure a second time but arrow down to Draw instead of calculate .

Compare α and the p-value: $\alpha = 0.05$ and the p-value $= 0.1959$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At a 5% level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

Exercise 11.5.2

Ivy League schools receive many applications, but only some can be accepted. At the schools listed in Table, two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the p-value, and draw a conclusion about the test of homogeneity.

Answer

H_0: The distribution of regular applications accepted is the same as the distribution of early applications accepted.

H_a: The distribution of regular applications accepted is not the same as the distribution of early applications accepted.

$df = 5$

χ^2test statistic $= 430.06$

This is a nonsymmetric chi-square curve with df = 5. The values 0, 5, and 430.06 are labeled on the horizontal axis. The value 5 coincides with the peak of the curve. A vertical upward line extends from 430.06 to the curve, and the region to the right of this line is shaded. The shaded area is equal to the p-value.

Figure 11.5.1.

Press the MATRX key and arrow over to EDIT . Press 1:[A] . Press 3 ENTER 3 ENTER . Enter the table values by row. Press ENTER after each. Press 2nd QUIT . Press STAT and arrow over to TESTS . Arrow down to C :χ2-TEST . Press ENTER . You should see Observed:[A] and Expected:[B] . Arrow down to Calculate . Press ENTER . The test statistic is 430.06 and the p-value $= 9.80E - 91$. Do the procedure a second time but arrow down to Draw instead of calculate .

References

1. Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).

2. “Energy use (kg of oil equivalent per capita).” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/...G.OE/countries (accessed May 24, 2013).
3. “Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/pubsinf...?pubid=2009030 (accessed May 24, 2013).
4. “Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES),” U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

Chapter 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

$\sum_{i,j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where: $O =$ observed values

$E =$ expected values

$i =$ number of rows in data contingency table

$j =$ number of columns in data contingency table

$df = (i-1)(j-1)$ Degrees of freedom

Exercise 11.5.3

A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

Answer

test for homogeneity

Exercise 11.5.4

What are the null and alternative hypotheses for Exercise?

Exercise 11.5.5

A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

Answer

test for homogeneity

Exercise 11.5.6

A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

Exercise 11.5.7

What condition must be met to use the test for homogeneity?

Answer

All values in the table must be greater than or equal to five.

Use the following information to answer the next five exercises: Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in Table.

	20–30	30–40	40–50	50–60
Private Practice	16	40	38	6
Hospital	8	44	59	39

Exercise 11.5.8

State the null and alternative hypotheses.

Exercise 11.5.9

$df =$ _____

Answer

3

Exercise 11.5.10

What is the test statistic?

Exercise 11.5.11

What is the p-value?

Answer

0.00005

Exercise 11.5.12

What can you conclude at the 5% significance level?

This page titled 13.5: Test for Homogeneity is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.6: Comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:

  H_0: The population fits the given distribution.

  H_a: The population does not fit the given distribution.

- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:

  H_0: The two variables (factors) are independent.

  H_a: The two variables (factors) are dependent.

- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:

  H_0: The two populations follow the same distribution.

  H_a: The two populations have different distributions.

Chapter Review

The goodness-of-fit test is typically used to determine if data fits a particular distribution. The test of independence makes use of a contingency table to determine the independence of two factors. The test for homogeneity determines whether two populations come from the same distribution, even if this distribution is unknown.

Exercise 11.6.1

Which test do you use to decide whether an observed distribution is the same as an expected distribution?

Answer

a goodness-of-fit test

Exercise 11.6.2

What is the null hypothesis for the type of test from Exercise?

Exercise 11.6.3

Which test would you use to decide whether two factors have a relationship?

Answer

a test for independence

Exercise 11.6.4

Which test would you use to decide if two populations have the same distribution?

Exercise 11.6.5

How are tests of independence similar to tests for homogeneity?

Answer

Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{i \cdot j} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

Exercise 11.6.6

How are tests of independence different from tests for homogeneity?

Bringing It Together

Exercise 11.6.7

a. Explain why a goodness-of-fit test and a test of independence are generally right-tailed tests.
b. If you did a left-tailed test, what would you be testing?

Answer

a. The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.
b. Testing to see if the data fits the distribution “too well” or is too perfect.

This page titled 13.6: Comparison of the Chi-Square Tests is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.7: 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:

$$\frac{(n-1)s^2}{\sigma^2} \tag{13.7.1}$$

where:

- $n =$ the total number of data
- $s^2 =$ sample variance
- $\sigma^2 =$ 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.** 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 11.7.1

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.

- $H_0 : \sigma^2 = 5^2$
- $H_a : \sigma^2 > 5^2$

Exercise 11.7.1

A SCUBA instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

Answer

$H_0 : \sigma^2 = 3^2$

$H_a : \sigma^2 > 3^2$

Example 11.7.2

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, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let $s =$ standard deviation for the waiting times.

- $H_0 : \sigma^2 = 7.2^2$

- $H_0 : \sigma^2 = 7.2^2$
- $H_a : \sigma^2 < 7.2^2$

The word "**less**" tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where:

- $n =$ the number of customers sampled
- \(df = n - 1 = 25 - 1 = 24\)

Calculate the test statistic:

$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph:

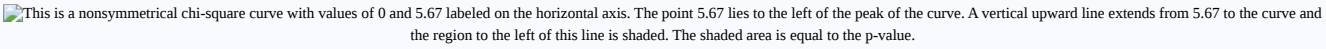

This is a nonsymmetrical chi-square curve with values of 0 and 5.67 labeled on the horizontal axis. The point 5.67 lies to the left of the peak of the curve. A vertical upward line extends from 5.67 to the curve and the region to the left of this line is shaded. The shaded area is equal to the p-value.

Figure 11.7.1.

Probability statement: $p\text{-value} = P(\chi^2 < 5.67) = 0.000042$

Compare α 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.

In `2nd DISTR`, use `7:χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example, `χ2cdf(-1E99,5.67,24)`. The $p\text{-value} = 0.000042$.

Exercise 11.7.2

The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August of 2012, the standard deviation of Internet speeds across Internet Service Providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, the test statistic, sketch the graph of the p-value, and draw a conclusion. Test at the 1% significance level.

Answer

$H_0 : \sigma^2 = 12.2^2$

$H_a : \sigma^2 > 12.2^2$

$df = 14$

$\text{chi}^2\text{test statistic} = 16.39$

Figure 11.7.2.

The p-value is 0.2902, so we decline to reject the null hypothesis. There is not enough evidence to suggest that the variance is greater than 12.2^2.

In `2nd DISTR`, use7: `χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. `χ2cdf(16.39,10^99,14)`. The $p\text{-value} = 0.2902$.

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.

Chapter 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

σ : population standard deviation

$df = n–1$ Degrees 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.

Exercise 11.7.3

What type of test should be used?

Answer

a test of a single variance

Exercise 11.7.4

State the null and alternative hypotheses.

Exercise 11.7.5

Is this a right-tailed, left-tailed, or two-tailed test?

Answer

a left-tailed test

Use the following information to answer the next three exercises: The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.

Exercise 11.7.6

What type of test should be used?

Exercise 11.7.5

State the null and alternative hypotheses.

Answer

$H_0 : \sigma^2 = 0.81^2$;

$H_a : \sigma^2 > 0.81^2$

Exercise 11.7.6

$df =$ _____

Use the following information to answer the next four exercises: The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

Exercise 11.7.7

What type of test should be used?

Answer

a test of a single variance

Exercise 11.7.8

What is the test statistic?

Exercise 11.7.9

What is the p-value?

Answer

0.0542

Exercise 11.7.10

What can you conclude at the 5% significance level?

This page titled 13.7: Test of a Single Variance is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet)

This page titled 13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.9: Lab 2: Chi-Square Test of Independence (Worksheet)

This page titled 13.9: Lab 2: Chi-Square Test of Independence (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

13.10: The Chi-Square Distribution (Exercises)

This page titled 13.10: The Chi-Square Distribution (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

14: 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.

14.1: Prelude to Linear Regression and Correlation

14.2: Linear Equations

Linear Equations (Exercises)

14.3: Scatter Plots

Scatter Plots (Exercises)

14.4: The Regression Equation

The Regression Equation (Exercise)

14.5: Testing the Significance of the Correlation Coefficient

Testing the Significance of the Correlation Coefficient (Exercises)

14.6: Prediction

Prediction (Exercises)

14.7: Outliers

Outliers (Exercises)

14.8: Regression - Distance from School (Worksheet)

14.9: Regression - Textbook Cost (Worksheet)

14.10: Regression - Fuel Efficiency (Worksheet)

14.11: Linear Regression and Correlation (Exercises)

Contributors

- Template:ContribOpenStax

This page titled 14: Linear Regression and Correlation is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

14.1: Prelude to Linear Regression and Correlation

CHAPTER OBJECTIVES

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

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it? In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.



Figure* 14.1.1*: *Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)*

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables. In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Contributors

- Template:ContribOpenStax

This page titled 14.1: Prelude to Linear Regression and Correlation is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

14.2: 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 14.2.1

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

Exercise 14.2.1

Is the following an example of a linear equation?

$$y = -0.125 - 3.5x$$

Answer

yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example 14.2.2

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



***Figure* 14.2.1.**

Exercise 14.2.2

Is the following an example of a linear equation? Why or why not?

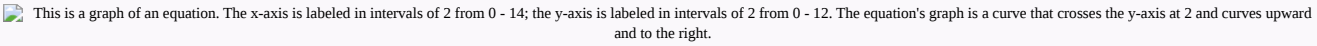


***Figure* 14.2.2.**

Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

Example 14.2.3

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Answer

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

Let $y =$ the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Exercise 14.2.3

Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer

$y = 50 + 20x$

Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, $b =$ slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

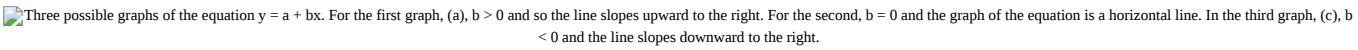


***Figure* 14.2.3.** Three possible graphs of $y = a + bx$ (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 14.2.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Exercise 14.2.4

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Answer

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is 25 ($a = 25$). At the start of a visit, Ethan charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Ethan earns $20 for each hour he works.

Summary

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change

between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

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. Graphically, the slope is represented by three line types in elementary statistics.

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

- Template:ContribOpenStax

This page titled 14.2: Linear Equations is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Linear Equations (Exercises)

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

Exercise 12.2.5

What are the dependent and independent variables?

Answer

dependent variable: fee amount; independent variable: time

Exercise 12.2.6

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise 12.2.7

Graph the equation from Exercise.

Answer

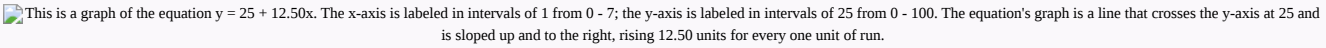

This is a graph of the equation y = 25 + 12.50x. The x-axis is labeled in intervals of 1 from 0 - 7; the y-axis is labeled in intervals of 25 from 0 - 100. The equation's graph is a line that crosses the y-axis at 25 and is sloped up and to the right, rising 12.50 units for every one unit of run.

Figure 12.2.4.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise 12.2.8

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise 12.2.9

Graph the equation from Exercise.

Answer

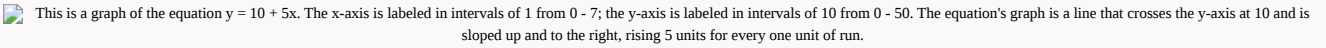

This is a graph of the equation y = 10 + 5x. The x-axis is labeled in intervals of 1 from 0 - 7; the y-axis is labeled in intervals of 10 from 0 - 50. The equation's graph is a line that crosses the y-axis at 10 and is sloped up and to the right, rising 5 units for every one unit of run.

Figure 12.2.5.

Exercise 12.2.10

Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise 12.2.11

Which of the following equations are linear?

a. $y = 6x + 8$
b. $y + 7 = 3x$
c. $y - x = 8x^2$
d. $4y = 8$

Answer

$y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

Exercise 12.2.12

Does the graph show a linear equation? Why or why not?

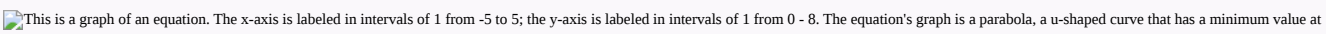

This is a graph of an equation. The x-axis is labeled in intervals of 1 from -5 to 5; the y-axis is labeled in intervals of 1 from 0 - 8. The equation's graph is a parabola, a u-shaped curve that has a minimum value at (0, 0).

Figure 12.2.6.

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Exercise 12.2.13

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Answer

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise 12.2.14

What are the independent and dependent variables?

Exercise 12.2.15

What is the *y*-intercept and what is the slope? Interpret them using complete sentences.

Answer

The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12,000x$.

Exercise 12.2.16

What are the independent and dependent variables?

Exercise 12.2.17

How many pounds of soil does the shoreline lose in a year?

Answer

12,000 pounds of soil

Exercise 12.2.18

What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15–1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise 12.2.19

What are the slope and *y*-intercept? Interpret their meaning.

Answer

The slope is -1.5 ($b = -1.5$). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

Exercise 12.2.19

If you owned this stock, would you want a positive or negative slope? Why?

Linear Equations (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.3: 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 14.3.1

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* 14.3.1:** *: 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

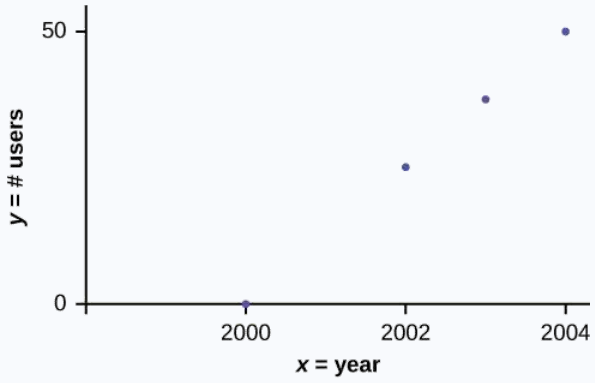


***Figure* 14.3.1:** Scatter plot showing the number of m-commerce users (in millions) by year.

To create a scatter plot

a. Enter your X data into list L1 and your Y data into list L2.
b. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
c. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
d. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
e. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
f. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
g. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Exercise 14.3.1

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

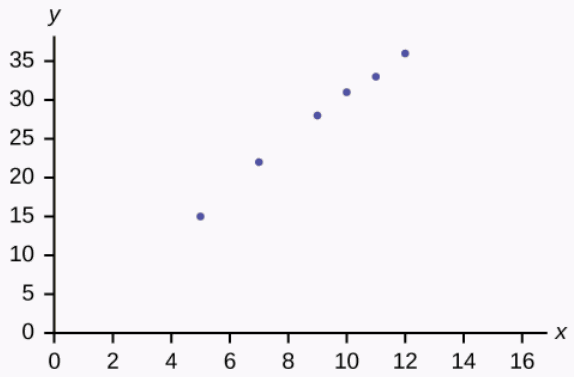


***Figure* 14.3.2**

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.

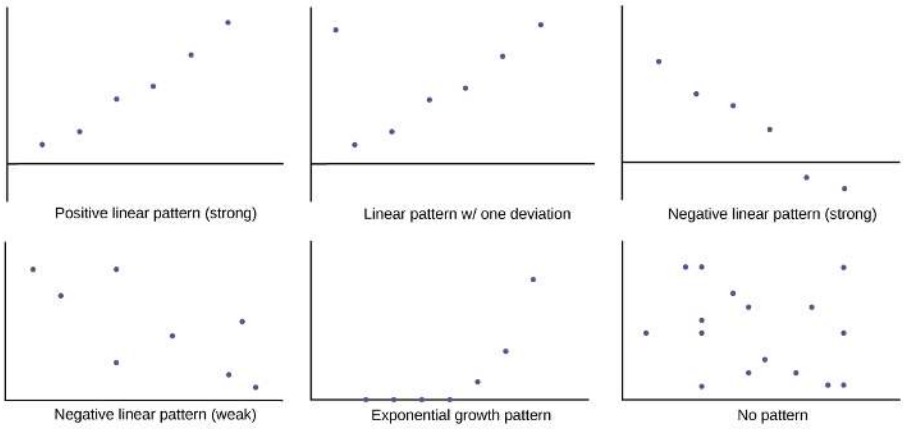


Figure* 14.3.3*:

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

- Template:ContribOpenStax

This page titled 14.3: Scatter Plots is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Scatter Plots (Exercises)

Exercise 1

Does the scatter plot appear linear? Strong or weak? Positive or negative?

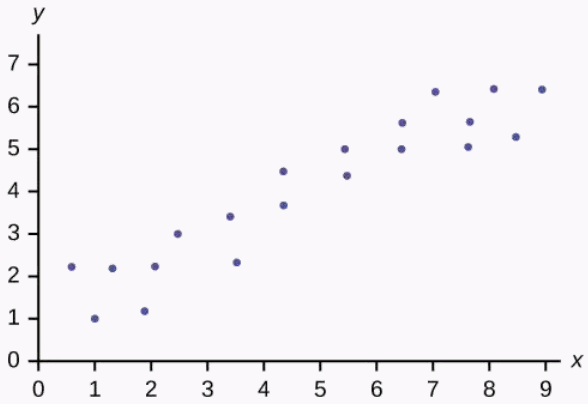


***Figure* 4**

Answer

The data appear to be linear with a strong, positive correlation.

Exercise 12.3.3

Does the scatter plot appear linear? Strong or weak? Positive or negative?

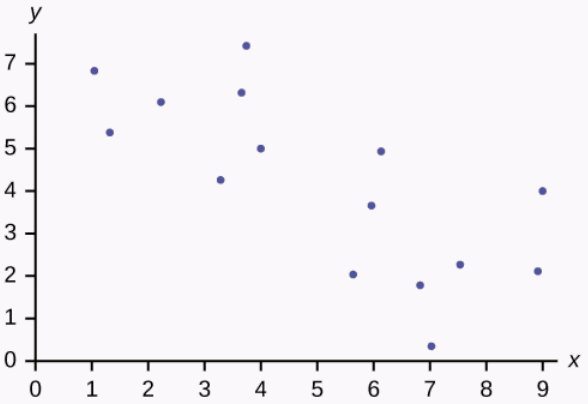


***Figure* 5**

Exercise 12.3.4

Does the scatter plot appear linear? Strong or weak? Positive or negative?

https://socialsci.libretexts.org/@go/page/19133

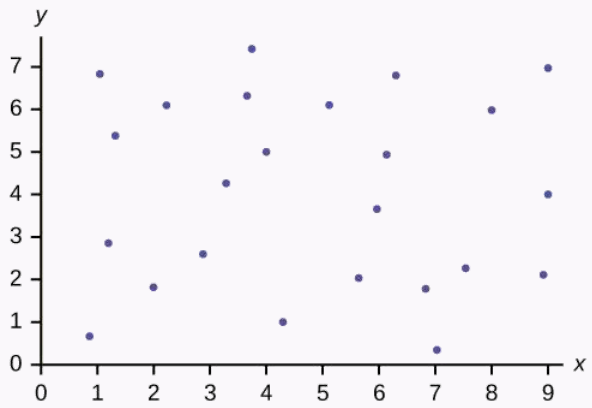


***Figure* 6**

Answer

The data appear to have no correlation.

Scatter Plots (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.4: The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a Line of Best Fit **or** Least-Squares Line.

COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 14.4.1

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

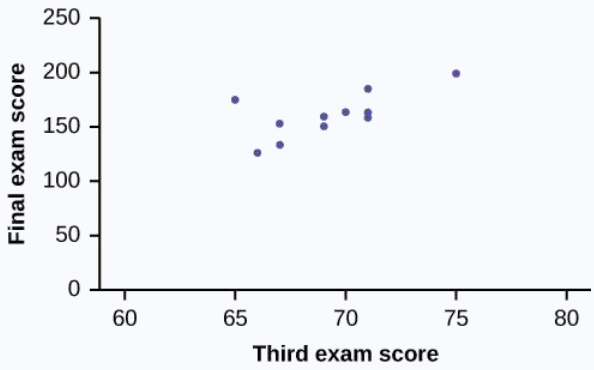


***Figure* 14.4.1**: Scatter plot showing the scores on the final exam based on scores from the third exam.

Exercise 14.4.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

X (depth in feet)	Y (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Answer

$\hat{y} = 127.24–1.11x$

At 110 feet, a diver could dive for only five minutes.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read "y **hat**" and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

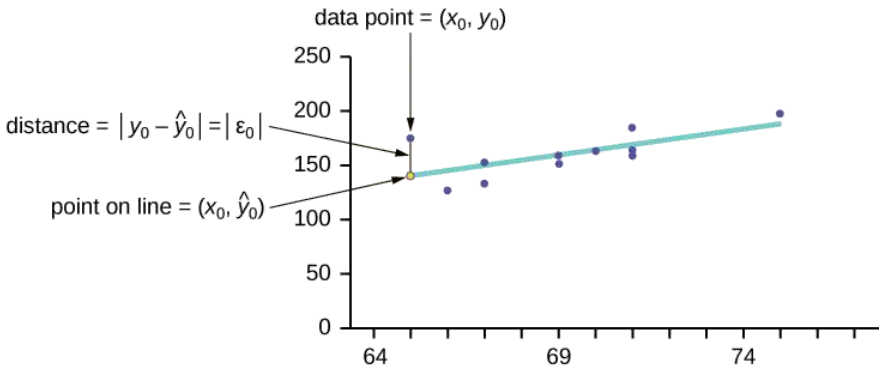


***Figure* 14.4.2**

The term $y_0–\hat{y}_0 = \varepsilon_0$ is called the **"error" or** residual. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram in Figure, $y_0–\hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

$\varepsilon =$ the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + ... + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2 \quad (14.4.1)$$

Equation14.4.1 is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx \quad (14.4.2)$$

where

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

The sample means of the x values and the x values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where $s_y =$ the standard deviation of the y values and $s_x =$ the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

Least Square Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line** .

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

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

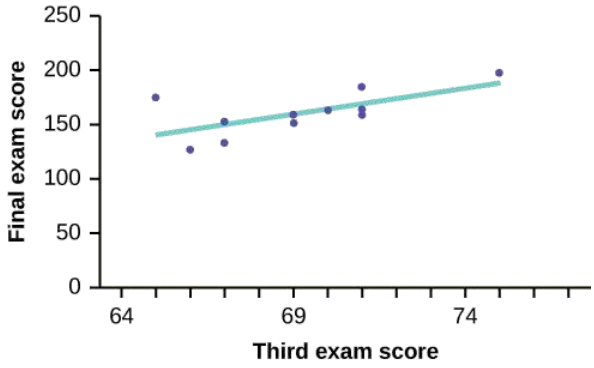


***Figure* 14.4.3**

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

$$\hat{y} = -173.51 + 4.83x \quad (14.4.3)$$

REMINDER

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.** You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

Understanding Slope

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the Linear Regression T Test: LinRegTTest

a. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x, y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
b. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
c. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
d. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER
e. Leave the line for "RegEq:" blank
f. Highlight Calculate and press ENTER.

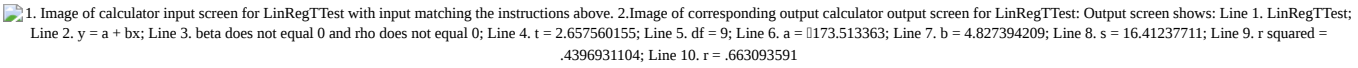

1. Image of calculator input screen for LinRegTTest with input matching the instructions above. 2.Image of corresponding output calculator output screen for LinRegTTest: Output screen shows: Line 1. LinRegTTest; Line 2. y = a + bx; Line 3. beta does not equal 0 and rho does not equal 0; Line 4. t = 2.657560155; Line 5. df = 9; Line 6. a = 173.513363; Line 7. b = 4.827394209; Line 8. s = 16.41237711; Line 9. r squared = .4396931104; Line 10. r = .663093591

***Figure* 14.4.4**

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and \(b = 4.8273\); the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$

The two items at the bottom are $r_2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On**, and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data

8. To graph the best-fit line, press the "$Y =$" key and type the equation $-173.5 + 4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

a. Make sure you have done the scatter plot. Check it on your screen.
b. Go to LinRegTTest and enter the lists.
c. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
d. Press $Y =$ (**you will see the regression equation**).
e. Press GRAPH. The line will be drawn."

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y. The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum y)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{14.4.4}$$

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."**

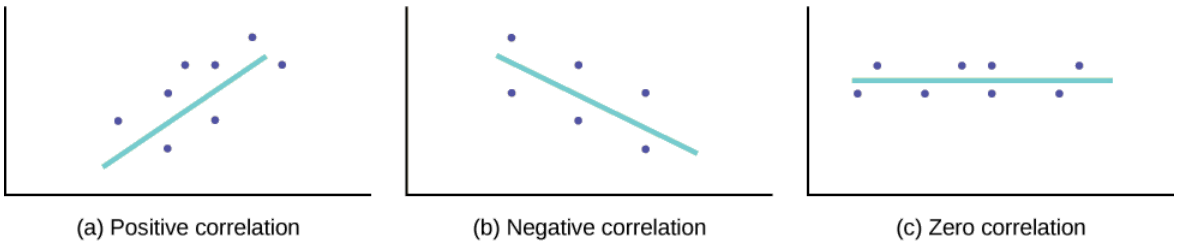


***Figure* 14.4.5:** (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called *the coefficient of determination* and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation $(1 - 0.44 = 0.56)$ in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Summary

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Glossary

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \tag{14.4.5}$$

where n is the number of data points. The coefficient cannot be more then 1 and less then –1. The closer the coefficient is to ± 1, the stronger the evidence of a significant linear relationship between x and y.

Contributors

- Template:ContribOpenStax

This page titled 14.4: The Regression Equation is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

The Regression Equation (Exercise)

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y	x	y
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Exercise 12.4.2

Draw a scatter plot of the data.

Exercise 12.4.3

Use regression to find the equation for the line of best fit.

Answer

$\hat{y} = 2.23 + 1.99x$

Exercise 12.4.4

Draw the line of best fit on the scatter plot.

Exercise 12.4.5

What is the slope of the line of best fit? What does it represent?

Answer

The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

Exercise 12.4.6

What is the y-intercept of the line of best fit? What does it represent?

Exercise 12.4.7

What does an r value of zero mean?

Answer

It means that there is no correlation between the data sets.

Exercise 12.4.8

When $n = 2$ and $r = 1$, are the data significant? Explain.

Exercise 12.4.9

When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

The Regression Equation (Exercise) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.5: Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together. We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

- 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 β or ρ, highlight "$\neq 0$"

The output screen shows the p-value on the line that reads "$p =$".

(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p-value. The following describes the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with $n - 2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.

An alternative way to calculate the p-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 14.5.1

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.

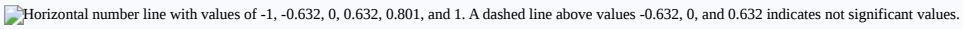


Figure 12.5.1. r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

Exercise 14.5.1

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 14.5.2

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

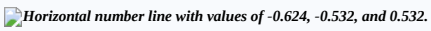


Figure 12.5.2. $r = -0.624 - 0.532$. Therefore, r is significant.

Exercise 14.5.2

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 14.5.3

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 12.5.3. $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

Exercise 14.5.3

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 14.5.4

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 14.5.4

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.



The left graph shows three sets of points. Each set falls in a vertical line. The points in each set are normally distributed along the line — they are densely packed in the middle and more spread out at the top and bottom. A downward sloping regression line passes through the mean of each set. The right graph shows the same regression line plotted. A vertical normal curve is shown for each line.

Figure 12.5.4. *The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.*

Summary

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.
- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis $H_0 : \rho =$ *hypothesized value*, use a linear regression t-test. The most common null hypothesis is $H_0 : \rho = 0$ which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx \tag{14.5.1}$$

where

$$a = y\text{-intercept} \tag{14.5.2}$$

$$b = \text{slope} \tag{14.5.3}$$

Standard deviation of the residuals:

$$s = \sqrt{\frac{SEE}{n-2}} \tag{14.5.4}$$

where

\(SSE = \text{sum of squared errors}\]

$$n = \text{the number of data points} \tag{14.5.5}$$

This page titled 14.5: Testing the Significance of the Correlation Coefficient is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Testing the Significance of the Correlation Coefficient (Exercises)

Exercise 12.5.5

When testing the significance of the correlation coefficient, what is the null hypothesis?

Exercise 12.5.6

When testing the significance of the correlation coefficient, what is the alternative hypothesis?

Answer

$H_a : \rho \neq 0$

Exercise 12.5.7

If the level of significance is 0.05 and the ***p*-value** is **0.04**, what conclusion can you draw?

Testing the Significance of the Correlation Coefficient (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.6: 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 14.6.1

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 14.6.1

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Answer

86.5

Summary

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

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 http://www.census.gov/compendia/stat...atalities.html
4. Data from the National Center for Health Statistics.

Contributors

- Template:ContribOpenStax

This page titled 14.6: Prediction is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Prediction (Exercises)

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$$\hat{y} = 101.32 + 2.48x \quad (1)$$

where $\hat{y}$ is in thousands of dollars.

Exercise 12.6.2

What would you predict the sales to be on day 60?

Answer

$250,120

Exercise 12.6.3

What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$$\hat{y} = 1350 - 1.2x \quad (2)$$

where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

Exercise 12.6.4

How many acres will be left to mow after 20 hours of work?

Answer

1,326 acres

Exercise 12.6.5

How many acres will be left to mow after 100 hours of work?

Exercise 12.6.7

How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

Answer

1,125 hours, or when $x = 1,125$

Table contains real data for the first two decades of AIDS reporting.

Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Exercise 12.6.8

Graph “year” versus “# AIDS cases diagnosed” (plot the scatter plot). Do not include pre-1981 data.

Exercise 12.6.9

Perform linear regression. What is the linear equation? Round to the nearest whole number.

Answer

Check student’s solution.

Exercise 12.6.10

Write the equations:

a. Linear equation: _____
b. $a =$ _____
c. $b =$ _____
d. $r =$ _____
e. $n =$ _____

Exercise 12.6.11

Solve.

a. When $x = 1985$, $\hat{y} =$ _____
b. When $x = 1990$, $\hat{y} =$_____
c. When $x = 1970$, $\hat{y} =$_____ Why doesn't this answer make sense?

Answer

a. When $x = 1985$, $\hat{y} = 25,52$
b. When $x = 1990$, $\hat{y} = 34,275$
c. When $x = 1970$, $\hat{y} = –725$ Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, where we are predicting AIDS cases diagnosed.

Exercise 12.6.11

Does the line seem to fit the data? Why or why not?

Exercise 12.6.12

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Also, the correlation $r = 0.4526$. If r is compared to the value in the 95% Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But the scatter plot indicates otherwise.

Exercise 12.6.13

Plot the two given points on the following graph. Then, connect the two points to form the regression line.

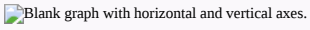


Figure 12.6.1.

Obtain the graph on your calculator or computer.

Exercise 12.6.14

Write the equation: $\hat{y} =$ _____

Answer

$\hat{y} = 3,448,225 + 1750x$

Exercise 12.6.15

Hand draw a smooth curve on the graph that shows the flow of the data.

Exercise 12.6.16

Does the line seem to fit the data? Why or why not?

Answer

There was an increase in AIDS cases diagnosed until 1993. From 1993 through 2002, the number of AIDS cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

Exercise 12.6.17

Do you think a linear fit is best? Why or why not?

Exercise 12.6.18

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Answer

Since there is no linear association between year and # of AIDS cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable “causes” the other variable.

Exercise 12.6.19

Graph “year” vs. “# AIDS cases diagnosed.” Do not include pre-1981. Label both axes with words. Scale both axes.

Exercise 12.6.20

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding to four decimal places:

Answer

We don’t know if the pre-1981 data was collected from a single year. So we don’t have an accurate *x* value for this figure.

Regression equation: Misplaced '#'

	Coefficients
Intercept	–3,448,225
X Variable 1	1,749.777

Exercise 12.6.21

Calculate the following:

a. $a =$ _____
b. $b =$ _____
c. correlation = _____
d. $n =$ _____

Prediction (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.7: 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 14.7.1

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.

Exercise 14.7.1

Identify the potential outlier in the scatter plot. The standard deviation of the residuals or errors is approximately 8.6.

Figure 12.7.2.

Answer:

The outlier appears to be at (6, 58). The expected y value on the line for the point (6, 58) is approximately 82. Fifty-eight is 24 units from 82. Twenty-four is more than two standard deviations ($2s = (2)(8.6) = 17.2$). So 82 is more than two standard deviations from 58, which makes $(6, 58)$ a potential outlier.

Numerical Identification of Outliers

In 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 14.7.2

Using this new line of best fit (based on the remaining ten data points in the third exam/final exam example), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Answer

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Exercise 14.7.2

The data points for the graph from the third exam/final exam example are as follows: (1, 5), (2, 7), (2, 6), (3, 9), (4, 12), (4, 13), (5, 18), (6, 19), (7, 12), and (7, 21). Remove the outlier and recalculate the line of best fit. Find the value of $\hat{y}$ when $x = 10$.

Answer

$\hat{y} = 1.04 + 2.96x; 30.64$

Example 14.7.3: 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.)

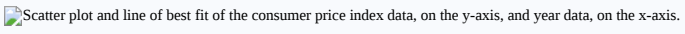


Figure 12.7.3.

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?)

Exercise 14.7.3

The following table shows economic development measured in per capita income PCINC.

Year	PCINC	Year	PCINC
1870	340	1920	1050
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

a. What are the independent and dependent variables?
b. Draw a scatter plot.
c. Use regression to find the line of best fit and the correlation coefficient.
d. Interpret the significance of the correlation coefficient.
e. Is there a linear relationship between the variables?
f. Find the coefficient of determination and interpret it.
g. What is the slope of the regression equation? What does it mean?
h. Use the line of best fit to estimate PCINC for 1900, for 2000.
i. Determine if there are any outliers.

Answer a

The independent variable (x) is the year and the dependent variable (y) is the per capita income.

Answer b

Figure 12.7.4.

Answer c

$\hat{y} = 18.61x–34574$; $r = 0.9732$

Answer d

At $df = 8$, the critical value is 0.632. The r value is significant because it is greater than the critical value.

Answer e

There does appear to be a linear relationship between the variables.

Answer f

The coefficient of determination is 0.947, which means that 94.7% of the variation in PCINC is explained by the variation in the years.

Answer g and h

The slope of the regression equation is 18.61, and it means that per capita income increases by $18.61 for each passing year. $\hat{y} = 785$ when the year is 1900, and $\hat{y} = 2,646$ when the year is 2000.

Answer i

There do not appear to be any outliers.

https://socialsci.libretexts.org/@go/page/19140

95% Critical Values of the Sample Correlation Coefficient Table

Degrees of Freedom: n	Critical Values: (+ and –)
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273

Degrees of Freedom: n–2	Critical Values: (+ and –)
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Summary

To determine if a point is an outlier, do one of the following:

1. Input the following equations into the TI 83, 83+,84, 84+:

$$y_1 = a + bx$$

$$y_2 = (2s)a + bx$$

$$y_3 = -(2s)a + bx$$

   where s is the standard deviation of the residuals

   If any point is above y_2 or below y_3 then the point is considered to be an outlier.
2. Use the residuals and compare their absolute values to $2s$ where s is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to $2s$, then the corresponding point is an outlier.

Note: The calculator function LinRegTTest (STATS TESTS LinRegTTest) calculates s.

References

1. Data from the House Ways and Means Committee, the Health and Human Services Department.
2. Data from Microsoft Bookshelf.
3. Data from the United States Department of Labor, the Bureau of Labor Statistics.
4. Data from the Physician's Handbook, 1990.
5. Data from the United States Department of Labor, the Bureau of Labor Statistics.

Glossary

Outlier

an observation that does not fit the rest of the data

Contributors

- Template:ContribOpenStax

This page titled 14.7: Outliers is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

Outliers (Exercises)

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

Figure 12.7.5.

Exercise 12.7.4

Do there appear to be any outliers?

Answer

Yes, there appears to be an outlier at $(6, 58)$.

Exercise 12.7.5

A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

Exercise 12.7.6

What effect did the potential outlier have on the line of best fit?

Answer

The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate is a predictor for the data.

Exercise 12.7.7

Are you more or less confident in the predictive ability of the new line of best fit?

Exercise 12.7.8

The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?

Answer

$s = 1.75$

Exercise 12.7.9

The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

Bring It Together

Exercise 12.7.10

The average number of people in a family that received welfare for various years is given in Table.

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

a. Using “year” as the independent variable and “welfare family size” as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
e. Based on the data in Table, is there a linear relationship between the year and the average number of people in a welfare family?
f. Using the least-squares line, estimate the welfare family sizes for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
g. Are there any outliers in the data?
h. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.7.11

The percent of female wage and salary workers who are paid hourly rates is given in Table for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

a. Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?

c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated percents for 1991 and 1988.
f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
g. Are there any outliers in the data?
h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a. Check student's solution.
b. yes
c. $\hat{y} = -266.8863 + 0.1656x$
d. 0.9448; Yes
e. $62.8233; 62.3265$
f. yes
g. yes; $(1987, 62.7)$
h. 72.5937; no
i. $slope = 0.1656$.

As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in Table.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Exercise 12.7.12

a. Using “size” as the independent variable and “cost” as the dependent variable, draw a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. If the laundry detergent were sold in a 40-ounce size, find the estimated cost.
f. If the laundry detergent were sold in a 90-ounce size, find the estimated cost.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the given data?
i. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would you cost? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.13

a. Complete Table for the cost per ounce of the different sizes.
b. Using “size” as the independent variable and “cost per ounce” as the dependent variable, draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?

d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
g. If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the the data?
j. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost per ounce? Why or why not?
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a.

Size (ounces)	Cost ($)	cents/oz
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

b. Check student's solution.
c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
d. $\hat{y} = 20.2368 - 0.0819x$
e. $r = -0.8086$
f. 40-oz: 16.96 cents/oz
g. 90-oz: 12.87 cents/oz
h. The relationship is not linear; the least squares line is not appropriate.
i. no outliers
j. No, you would be extrapolating. The 300-oz size is outside the range of x.
k. $slope = -0.08194$; for each additional ounce in size, the cost per ounce decreases by 0.082 cents.

Exercise 12.7.14

According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.

c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a next taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.15

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a. Size is x, the independent variable, price is y, the dependent variable.
b. Check student's solution.
c. The relationship does not appear to be linear.
d. $\hat{y} = -745.252 + 54.75569x$
e. $r = 0.8944$, yes it is significant
f. 32-inch: $1006.93, 50-inch: $1992.53
g. No, the relationship does not appear to be linear. However, r is significant.
h. yes, the 60-inch TV
i. For each additional inch, the price increases by $54.76

Exercise 12.7.16

Table shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise 12.7.17

State	# letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. What are the independent and dependent variables?
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?

d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in part f to plot the least-squares line on your graph from part b.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
k. Delete "Hawaii" and substitute "Alaska" for it. Alaska is the forty-ninth, state with an area of 656,424 square miles.
l. Calculate the new least-squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

Answer

a. Let rank be the independent variable and area be the dependent variable.
b. Check student's solution.
c. There appears to be a linear relationship, with one outlier.
d. $\hat{y}\ (\text{area}) = 24177.06 + 1010.478x$
e. $r = 0.50047$, r is not significant so there is no relationship between the variables.
f. Alabama: 46407.576 Colorado: 62575.224
g. Alabama estimate is closer than Colorado estimate.
h. If the outlier is removed, there is a linear relationship.
i. There is one outlier (Hawaii).
j. rank 51: 75711.4; no
k.

Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Alaska	6	1959	51	656,424
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

l. $\hat{y} = -87065.3 + 7828.532x$
m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.
n. yes, with the exception of Hawaii

Outliers (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

14.8: Regression - Distance from School (Worksheet)

This page titled 14.8: Regression - Distance from School (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

14.9: Regression - Textbook Cost (Worksheet)

This page titled 14.9: Regression - Textbook Cost (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

14.10: Regression - Fuel Efficiency (Worksheet)

This page titled 14.10: Regression - Fuel Efficiency (Worksheet) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

14.11: Linear Regression and Correlation (Exercises)

This page titled 14.11: Linear Regression and Correlation (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

15: 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

15.1: Prelude to F Distribution and One-Way ANOVA

15.2: One-Way ANOVA

15.3: The F Distribution and the F-Ratio

15.4: Facts About the F Distribution

15.5: Test of Two Variances

15.6: Lab: One-Way ANOVA

15.7: F Distribution and One-Way ANOVA (Exercises)

Contributors

Template:ContribOpenStax

This page titled 15: F Distribution and One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.1: 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

Template:ContribOpenStax

This page titled 15.1: Prelude to F Distribution and One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.2: 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).

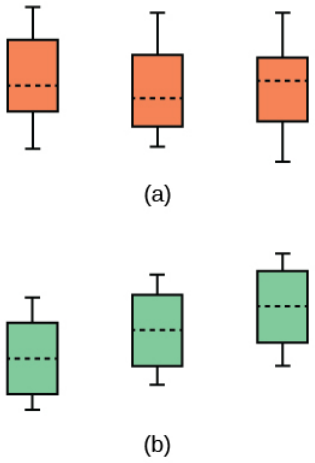


***Figure* 15.2.1:** *(a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.*

Chapter 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:

b. Each population from which a sample is taken is assumed to be normal.
c. 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-Way$ANOVA$

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

Template:ContribOpenStax

This page titled 15.2: One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.3: 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 σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
b. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- SS_{between} = **the sum of squares that represents the variation among the different samples**
- SS_{within} = **the sum of squares that represents the variation within samples that is due to chance.**

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in discussed previously.

MS means "mean square." MS_{between} is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the 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):

$$n = \sum n_j \tag{15.3.1}$$

- x = one value:

$$\sum x = \sum s_j \tag{15.3.2}$$

- Sum of squares of all values from every group combined:

$$\sum x^2 \tag{15.3.3}$$

- Between group variability:

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x^2\right)}{n} \tag{15.3.4}$$

- Total sum of squares:

$$\sum x^2 - \frac{(\sum x)^2}{n} \tag{15.3.5}$$

- Explained variation: sum of squares representing variation among the different samples:

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{(\sum s_j)^2}{n} \tag{15.3.6}$$

- Unexplained variation: sum of squares representing variation within samples due to chance:

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}} \tag{15.3.7}$$

- df's for different groups (df's for the numerator):

$$df = k - 1 \tag{15.3.8}$$

- Equation for errors within samples (df's for the denominator):

$$df_{\text{within}} = n - k \tag{15.3.9}$$

- Mean square (variance estimate) explained by the different groups:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} \tag{15.3.10}$$

- Mean square (variance estimate) that is due to chance (unexplained):

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} \tag{15.3.11}$$

MS_{between} and MS_{within} can be written as follows:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1} \tag{15.3.12}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{SS_{\text{within}}}{n-k} \tag{15.3.13}$$

The one-way $ANOVA$ test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{15.3.14}$$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}.Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2} \tag{15.3.15}$$

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)	$SS(\text{Factor})$	$k-1$	$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)}$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})}$
Error (Within)	$SS(\text{Error})$	$n-k$	$MS(\text{Error}) = \frac{SS(\text{Error})}{(n-k)}$	
Total	$SS(\text{Total})$	$n-1$		

Example 15.3.1

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	

$$s_1 = 16.5, s_2 = 15, s_3 = 15.7 \tag{15.3.16}$$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(\text{between}) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n} \tag{15.3.17}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} + \frac{(s_1 + s_2 + s_3)^2}{10} \tag{15.3.18}$$

where $n_1 = 4, n_2 = 3, n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$ so

$$SS(\text{between}) = \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} = \frac{(16.5 + 15 + 15.5)^2}{10} \tag{15.3.19}$$

$$= 2.2458 \tag{15.3.20}$$

$$S(\text{total}) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \tag{15.3.21}$$

$$= (5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2) \tag{15.3.22}$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10} \tag{15.3.23}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9 \tag{15.3.24}$$

$$= 23.1 \tag{15.3.25}$$

$$SS(\text{within}) = SS(\text{total}) - SS(\text{between}) \quad (15.3.26)$$
$$= 23.1 - 2.2458 \quad (15.3.27)$$
$$= 20.8542 \quad (15.3.28)$$

One-Way *ANOVA* Table: The formulas for $SS(\text{Total})$, $SS(\text{Factor}) = SS(\text{Between})$ and $SS(\text{Error}) = SS(\text{Within})$ as shown previously. The same information is provided by the TI calculator hypothesis test function *ANOVA* in STAT TESTS (syntax is $ANOVA(L1, L2, L3)$ where $L1, L2, L3$ have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	$SS(\text{Factor}) = SS(\text{Between}) = 2.2458$	$k - 1 = 3 \text{ groups} - 1 = 2$	$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)} = \frac{2.2458}{2} = 1.1229$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})} = \frac{1.1229}{2.9792} = 0.3769$
Error (Within)	$SS(\text{Error}) = SS(\text{Within}) = 20.8542$	$n–k = 10 \text{ total data} - 3 \text{ groups} = 7$	$MS(\text{Error})) = \frac{SS(\text{Error})}{(n-k)} = \frac{20.8542}{7} = 2.9792$	
Total	$SS(\text{Total}) = 2.2458 + 20.8542 = 23.1$	$n - 1 = 10 \text{ total data} - 1 = 9$		

Exercise 15.3.1

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	$5-1=4$	$\frac{36,648,561}{4}=9,162,140$	$\frac{9,162,140}{2,044,672.6}=4.4810$
Error (Within)	20,446,726	$15-5=10$	$\frac{20,446,726}{10}$ $=2,044,672.6$	
Total	57,095,287	$15-1=14$		

The one-way ***ANOVA*** *hypothesis test is always right-tailed* because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim Fdf(\text{num}), df(\text{denom})$

where $df(\text{num}) = df_{between} and df(\text{denom}) = df_{within}$

The mean for the F distribution is $\mu = \frac{df(\text{num})}{df(\text{denom})-1}$

References

1. Tomato Data, Marist College School of Science (unpublished student research)

Chapter 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

Template:ContribOpenStax

This page titled 15.3: The F Distribution and the F-Ratio is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.4: 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.

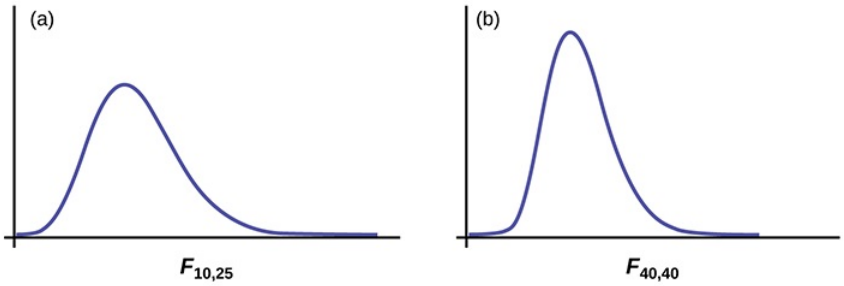


***Figure* 15.4.1**

Example 15.4.1

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	$5 - 1 = 4$	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	$15 - 5 = 10$	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	$15 - 1 = 14$		

Distribution for the test: $F_{4,10}$

$$df(\text{num}) = 5 - 1 = 4 \quad (15.4.1)$$

$$df(\text{denom}) = 15 - 5 = 10 \quad (15.4.2)$$

Test statistic: $F = 4.4810$

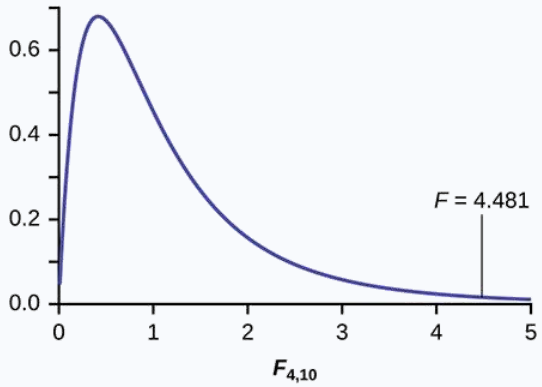


Figure 15.4.2

Probability Statement: $p\text{-value} = P(F > 4.481) = 0.0248$.

Compare α and the p-value: $\alpha = 0.05, p\text{-value} = 0.0248$

Make a decision: Since $\alpha > p\text{-value}$, we reject H_0.

Conclusion: At the 5% significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of mulches led to different mean yields.

To find these results on the calculator:

Press STAT. Press 1:EDIT. Put the data into the lists L_1, L_2, L_3, L_4, L_5.

Press STAT, and arrow over to TESTS, and arrow down to ***ANOVA***. Press ENTER, and then enter L_1, L_2, L_3, L_4, L_5). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the p-value of the test.

The calculator displays:

- $F = 4.4810$
- $p = 0.0248$ (p-value)

Factor

- $df = 4$
- $SS = 36648560.9$
- $MS = 9162140.23$

Error

- $df = 10$
- $SS = 20446726$
- $MS = 2044672.6$

Exercise 15.4.1

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:

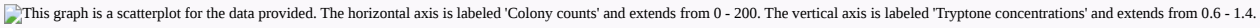


***Figure* 15.4.3**

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 15.4.1), 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 15.4.1

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	10,233	$5 - 1 = 4$	$\frac{10,233}{4} = 2,558.25$	$\frac{2,558.25}{4,194.9} = 0.6099$
Error (Within)	41,949	$15 - 5 = 10$		
Total	52,182	$15 - 1 = 14$	$\frac{41,949}{10} = 4,194.9$	

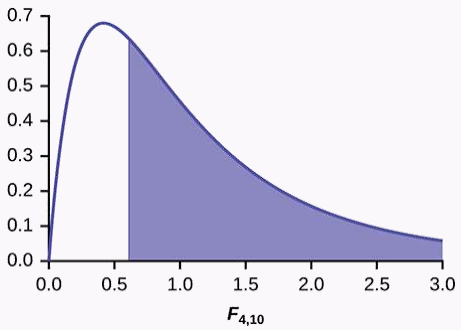


***Figure* 15.4.2**

Distribution for the test: $F_{4,10}$

Probability Statement: $p\text{-value} = P(F > 0.6099) = 0.6649$.

Compare α and the p-value: $\alpha = 0.05, p\text{-value} = 0.669, \alpha > p\text{-value}$

Make a decision: Since $\alpha > p\text{-value}$, we do not reject H_0.

Conclusion: At the 5% significance level, there is insufficient evidence from these data that different levels of tryptone will cause a significant difference in the mean number of bacterial colonies formed.

Example 15.4.2

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in Table.

***Figure* 15.4.1:** *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:

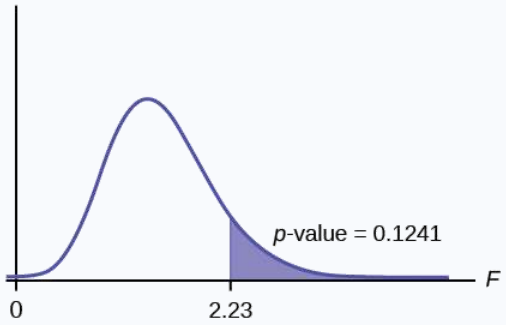


***Figure* 15.4.5**

Probability statement: $p\text{-value} = P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$

$p\text{-value} = 0.1241$

$\alpha < p\text{-value}$

Make a decision: Since $\alpha < p\text{-value}$, you cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Put the data into lists L_1, L_2, L_3, and L_4. Press `STAT` and arrow over to `TESTS`. Arrow down to `F:ANOVA`. Press `ENTER` and Enter (`L1,L2,L3,L4`).

The calculator displays the F statistic, the p-value and the values for the one-way $ANOVA$ table:

$F = 2.2303$

$p = 0.1241$ (p-value)

Factor

$df = 3$

$SS = 2.88732$

$MS = 0.96244$

Error

$df = 1$

$SS = 6.9044$

$MS = 0.431525$

Exercise 15.4.2

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 15.4.3

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 15.4.3.

Table 15.4.3

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

$$F' = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2}. \tag{15.4.3}$$

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 15.4.3

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 15.4.3.

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.

Chapter 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

- Template:ContribOpenStax

This page titled 15.4: Facts About the F Distribution is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.5: 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 σ_1^2 and σ_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:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} \tag{15.5.1}$$

F has the distribution

$$F \sim F(n_1 - 1, n_2 - 1) \tag{15.5.2}$$

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

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}. \tag{15.5.3}$$

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

$$F = \frac{(s_1)^2}{(s_2)^2} \tag{15.5.4}$$

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 15.5.1

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:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(s_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818 \tag{15.5.5}$$

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.

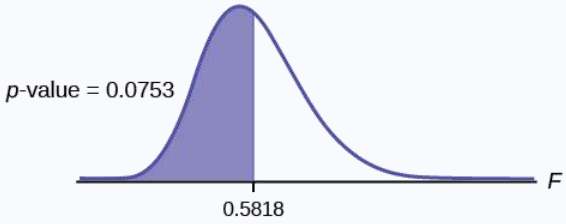


Figure 15.5.1

Probability statement: $p\text{-value} = P(F < 0.5818) = 0.0753$

Compare α 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 15.5.1

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

$H_0 : \sigma_{\mathrm{T1}}^2 = \sigma_{\mathrm{B2}}^2$ and $H_0 : \sigma_{\mathrm{T1}}^2 \neq \sigma_{\mathrm{B2}}^2$ (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.

Chapter 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

- Template:ContribOpenStax

This page titled 15.5: Test of Two Variances is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.6: Lab: One-Way ANOVA

This page titled 15.6: Lab: One-Way ANOVA is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

15.7: F Distribution and One-Way ANOVA (Exercises)

This page titled 15.7: F Distribution and One-Way ANOVA (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by OpenStax.

CHAPTER OVERVIEW

16: Advanced Graphs

Template:TopHeblLane

Topic hierarchy

16.1: Q-Q Plots

16.2: Contour Plots

16.3: 3D Plots

16.4: Statistical Literacy

16.5: Advanced Graphs (Exercises)

Contributors

- Template:ContribHeblLane

This page titled 16: Advanced Graphs is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

16.1: Q-Q Plots

Skills to Develop

- State what $q-q$ plots are used for.
- Describe the shape of a $q-q$ plot when the distributional assumption is met.
- Be able to create a normal $q-q$ plot.

The quantile-quantile or $q-q$ plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the $q-q$ plot will fall approximately on a straight line.

Before delving into the details of $q-q$ plots, we first describe two related graphical methods for assessing distributional assumptions: the histogram and the cumulative distribution function (CDF). As will be seen, $q-q$ plots are more general than these alternatives.

Assessing Distributional Assumptions

As an example, consider data measured from a physical device such as the spinner depicted in Figure 16.1.1. The red arrow is spun around the center, and when the arrow stops spinning, the number between 0 and 1 is recorded. Can we determine if the spinner is fair?

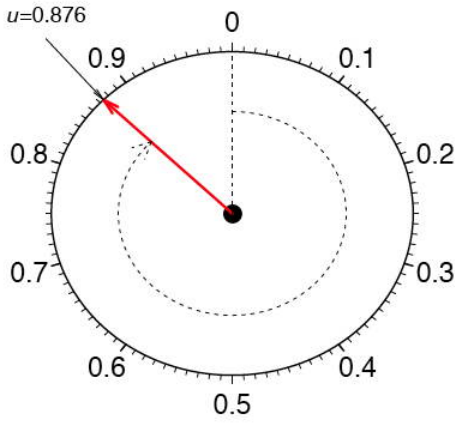


Figure 16.1.1***:*** *A physical device that gives samples from a uniform distribution*

If the spinner is fair, then these numbers should follow a uniform distribution. To investigate whether the spinner is fair, spin the arrow n times, and record the measurements by $\mu_1, \mu_2, ..., \mu_n$. In this example, we collect $n = 100$ samples. The histogram provides a useful visualization of these data. In Figure 16.1.2, we display three different histograms on a probability scale. The histogram should be flat for a uniform sample, but the visual perception varies depending on whether the histogram has 10, 5, or 3 bins. The last histogram looks flat, but the other two histograms are not obviously flat. It is not clear which histogram we should base our conclusion on.

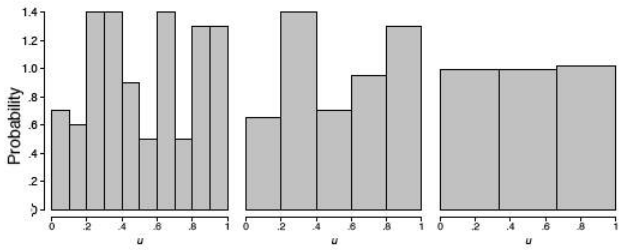


Figure 16.1.2***:****Three histograms of a sample of* 100 *uniform points*

Alternatively, we might use the cumulative distribution function (CDF), which is denoted by $F(\mu)$. The CDF gives the probability that the spinner gives a value less than or equal to μ, that is, the probability that the red arrow lands in the interval $[0, \mu]$. By simple arithmetic, $F(\mu) = \mu$, which is the diagonal straight line $y = x$. The CDF based upon the sample data is called the empirical CDF (ECDF), is denoted by $\widehat{F}_n(u)$, and is defined to be the fraction of the data less than or equal to μ; that is,

$$\widehat{F}_n(u) = \frac{\#u_i \leq u}{n} \tag{16.1.1}$$

In general, the ECDF takes on a ragged staircase appearance.

For the spinner sample analyzed in Figure 16.1.2, we computed the ECDF and CDF, which are displayed in Figure 16.1.3. In the left frame, the ECDF appears close to the line $y = x$, shown in the middle frame. In the right frame, we overlay these two curves and verify that they are indeed quite close to each other. Observe that we do not need to specify the number of bins as with the histogram.

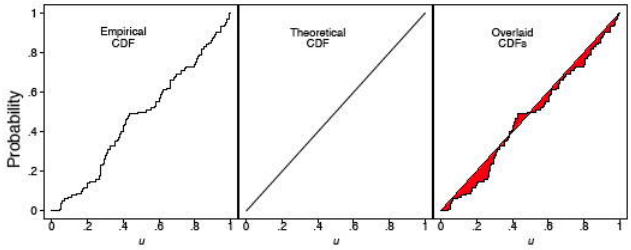


Figure* 16.1.3*: *The empirical and theoretical cumulative distribution functions of a sample of* 100 *uniform points*

q-q plot for uniform data

The $q - q$ plot for uniform data is very similar to the empirical CDF graphic, except with the axes reversed. The $q - q$ plot provides a visual comparison of the sample quantiles to the corresponding theoretical quantiles. In general, if the points in a $q - q$ plot depart from a straight line, then the assumed distribution is called into question.

Here we define the qth quantile of a batch of n numbers as a number ξ_q such that a fraction q x n of the sample is less than ξ_q, while a fraction $(1 - q) \times n$ of the sample is greater than ξ_q. The best known quantile is the median, $\xi_{0.5}$, which is located in the middle of the sample.

Consider a small sample of 5 numbers from the spinner:

$$\mu_1 = 0.41,\ \mu_2 = 0.24,\ \mu_3 = 0.59,\ \mu_4 = 0.03,\ \mu_5 = 0.67 \tag{16.1.2}$$

Based upon our description of the spinner, we expect a uniform distribution to model these data. If the sample data were "perfect," then on average there would be an observation in the middle of each of the 5 intervals: 0 to 0.2, 0.2 to 0.4, 0.4 to 0.6, and so on. Table 16.1.1 shows the 5 data points (sorted in ascending order) and the theoretically expected value of each based on the assumption that the distribution is uniform (the middle of the interval).

***Table* 16.1.1:** *Computing the Expected Quantile Values*

Data (μ)	Rank (*i*)	Middle of the *i*th Interval
0.03	1	0.1
0.24	2	0.3
0.41	3	0.5
0.59	4	0.7
0.67	5	0.9

The theoretical and empirical CDFs are shown in Figure 16.1.4 and the $q - q$ plot is shown in the left frame of Figure 16.1.5.

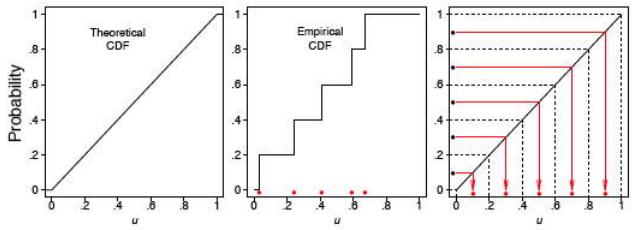


Figure* 16.1.4*: *The theoretical and empirical CDFs of a small sample of* 5 *uniform points, together with the expected values of the* 5 *points (red dots in the right frame)*

In general, we consider the full set of sample quantiles to be the sorted data values

$$\mu_{(1)} < \mu_{(2)} < \mu_{(3)} < \ldots < \mu_{(n-1)} < \mu_{(n)}, \tag{16.1.3}$$

where the parentheses in the subscript indicate the data have been ordered. Roughly speaking, we expect the first ordered value to be in the middle of the interval $(0, 1/n)$, the second to be in the middle of the interval $(1/n, 2/n)$, and the last to be in the middle of the interval $(\frac{n-1}{n}, 1)$. Thus, we take as the theoretical quantile the value

$$\xi_q = q \approx \frac{i - 0.5}{n} \tag{16.1.4}$$

where q corresponds to the i^{th} ordered sample value. We subtract the quantity 0.5 so that we are exactly in the middle of the interval $(\frac{i-1}{n}, \frac{i}{n})$. These ideas are depicted in the right frame of Figure 16.1.4 for our small sample of size $n = 5$.

We are now prepared to define the $q - q$ plot precisely. First, we compute the n expected values of the data, which we pair with the n data points sorted in ascending order. For the uniform density, the $q - q$ plot is composed of the n ordered pairs

$$(\tfrac{i-0.5}{n}, u_i), \ \ for\ i = 1, 2, \cdots, n \tag{16.1.5}$$

This definition is slightly different from the ECDF, which includes the points $(u_i, \frac{i}{n})$. In the left frame of Figure 16.1.5, we display the $q - q$ plot of the 5 points in Table 16.1.1. In the right two frames of Figure 16.1.5, we display the $q - q$ plot of the same batch of numbers used in Figure 16.1.2. In the final frame, we add the diagonal line $y = x$ as a point of reference.

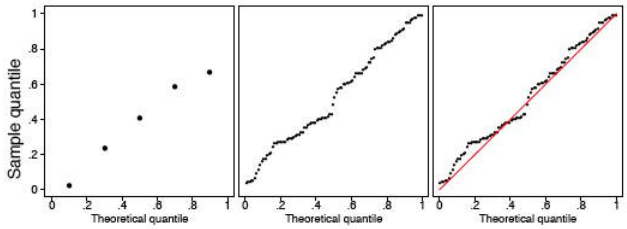


Figure* 16.1.5*: *(Left)* $q - q$ *plot of the* 5 *uniform points. (Right)* $q - q$ *plot of a sample of* 100 *uniform points*

The sample size should be taken into account when judging how close the $q - q$ plot is to the straight line. We show two other uniform samples of size $n = 10$ and $n = 1000$ in Figure 16.1.6. Observe that the $q - q$ plot when $n = 1000$ is almost identical to the line $y = x$, while such is not the case when the sample size is only $n = 10$.

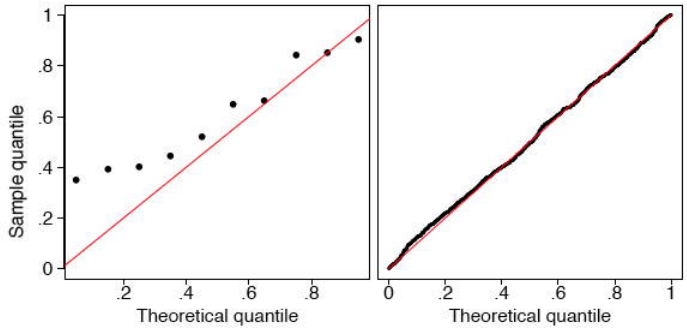


Figure* 16.1.6*: *$q-q$ plots of a sample of* 10 *and* 1000 *uniform points*

In Figure 16.1.7, we show the $q-q$ plots of two random samples that are not uniform. In both examples, the sample quantiles match the theoretical quantiles only at the median and at the extremes. Both samples seem to be symmetric around the median. But the data in the left frame are closer to the median than would be expected if the data were uniform. The data in the right frame are further from the median than would be expected if the data were uniform.

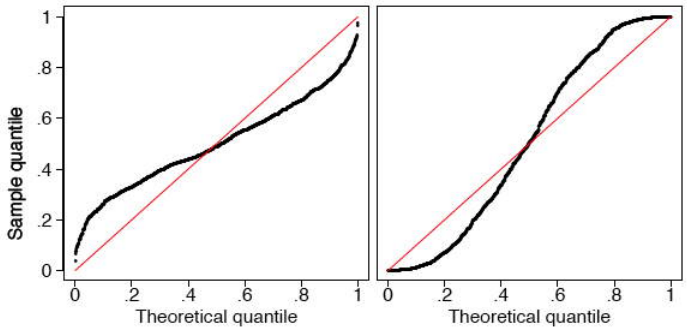


Figure 16.1.7. *$q-q$ plots of two samples of size* 1000 *that are not uniform.*

In fact, the data were generated in the *R* language from beta distributions with parameters $a=b=3$ on the left and $a=b=0.4$ on the right. In Figure 16.1.8 we display histograms of these two data sets, which serve to clarify the true shapes of the densities. These are clearly non-uniform.

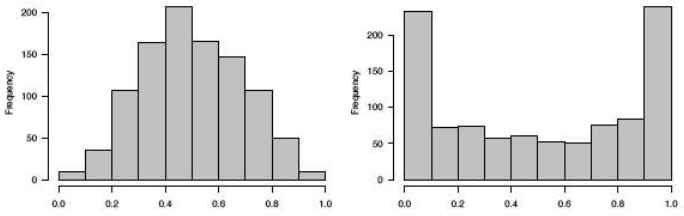


Figure* 16.1.8*: *Histograms of the two non-uniform data sets*

q-q plot for normal data

The definition of the $q-q$ plot may be extended to any continuous density. The $q-q$ plot will be close to a straight line if the assumed density is correct. Because the cumulative distribution function of the uniform density was a straight line, the $q-q$ plot was very easy to construct. For data that are not uniform, the theoretical quantiles must be computed in a different manner.

Let $z_1, z_2, ..., z_n$ denote a random sample from a normal distribution with mean $\mu=0$ and standard deviation $\sigma=1$. Let the ordered values be denoted by

$$z_{(1)} < z_{(2)} < z_{(3)} < ... < z_{(n-1)} < z_{(n)} \tag{16.1.6}$$

These n ordered values will play the role of the sample quantiles.

Let us consider a sample of 5 values from a distribution to see how they compare with what would be expected for a normal distribution. The 5 values in ascending order are shown in the first column of Table 16.1.2.

Table **16.1.2**: *Computing the expected quantile values for normal data.*

Data (z)	Rank (*i*)	Middle of the *i*th Interval	Normal(*z*)
-1.96	1	0.1	-1.28
-0.78	2	0.3	-0.52
0.31	3	0.5	0.00
1.15	4	0.7	0.52
1.62	5	0.9	1.28

Just as in the case of the uniform distribution, we have 5 intervals. However, with a normal distribution the theoretical quantile is not the middle of the interval but rather the inverse of the normal distribution for the middle of the interval. Taking the first interval as an example, we want to know the z value such that 0.1 of the area in the normal distribution is below z. This can be computed using the Inverse Normal Calculator as shown in Figure 16.1.9. Simply set the "Shaded Area" field to the middle of the interval (0.1) and click on the "Below" button. The result is -1.28. Therefore, 10% of the distribution is below a z value of -1.28.

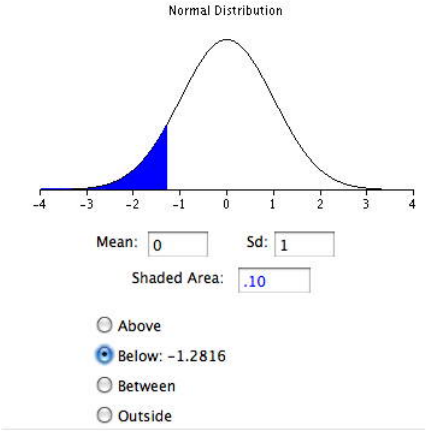


Figure* 16.1.9*: *Example of the Inverse Normal Calculator for finding a value of the expected quantile from a normal distribution*

The $q-q$ plot for the data in Table 16.1.2 is shown in the left frame of Figure 16.1.11.

In general, what should we take as the corresponding theoretical quantiles? Let the cumulative distribution function of the normal density be denoted by $\Phi(z)$. In the previous example, $\Phi(-1.28) = 0.10$ and $\Phi(0.00) = 0.50$. Using the quantile notation, if ξ_q is the q^{th} quantile of a normal distribution, then

$$\Phi(\xi_q) = q \tag{16.1.7}$$

That is, the probability a normal sample is less than ξ_q is in fact just q.

Consider the first ordered value, z_1. What might we expect the value of $\Phi(z_1)$ to be? Intuitively, we expect this probability to take on a value in the interval $(0, 1/n)$. Likewise, we expect $\Phi(z_2)$ to take on a value in the interval $(1/n, 2/n)$. Continuing, we expect $\Phi(z_n)$ to fall in the interval $((n-1)/n, 1)$. Thus, the theoretical quantile we desire is defined by the inverse (not reciprocal) of the normal CDF. In particular, the theoretical quantile corresponding to the empirical quantile z_i should be

$$\xi_q \approx \frac{i-0.5}{n} \tag{16.1.8}$$

for $i = 1, 2, \ldots, n$.

The empirical CDF and theoretical quantile construction for the small sample given in Table 16.1.2 are displayed in Figure 16.1.10. For the larger sample of size 100, the first few expected quantiles are -2.576, -2.170, and -1.960.

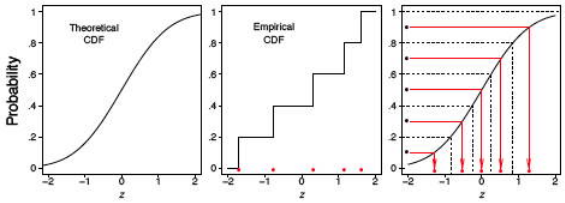


***Figure* 16.1.10:** *The empirical CDF of a small sample of* 5 *normal points, together with the expected values of the* 5 *points (red dots in the right frame)*

In the left frame of Figure 16.1.11, we display the $q-q$ plot of the small normal sample given in Table 16.1.2. The remaining frames in Figure 16.1.11 display the $q-q$ plots of normal random samples of size $n=100$ and $n=1000$. As the sample size increases, the points in the $q-q$ plots lie closer to the line $y=x$.

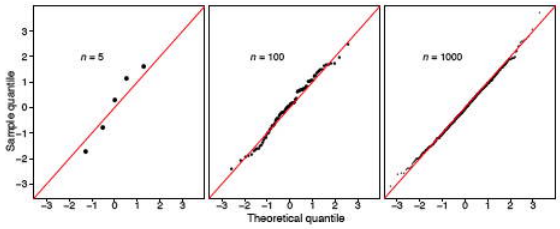


***Figure* 16.1.11:** *$q-q$ plots of normal data*

As before, a normal $q-q$ plot can indicate departures from normality. The two most common examples are skewed data and data with heavy tails (large kurtosis). In Figure 16.1.12, we show normal $q-q$ plots for a chi-squared (skewed) data set and a Student's-t (kurtotic) data set, both of size $n=1000$. The data were first standardized. The red line is again $y=x$. Notice, in particular, that the data from the t distribution follow the normal curve fairly closely until the last dozen or so points on each extreme.

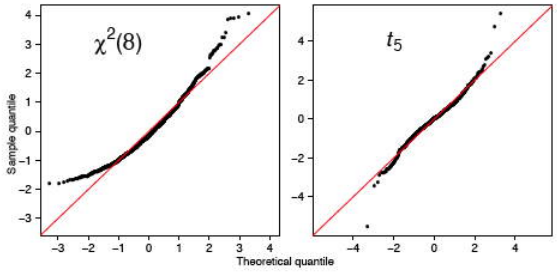


***Figure* 16.1.12:** *$q-q$ plots for standardized non-normal data ($n=1000$)*

q-q plots for normal data with general mean and scale

Our previous discussion of $q-q$ plots for normal data all assumed that our data were standardized. One approach to constructing $q-q$ plots is to first standardize the data and then proceed as described previously. An alternative is to construct the plot directly from raw data.

In this section, we present a general approach for data that are not standardized. Why did we standardize the data in Figure 16.1.12? The $q-q$ plot is comprised of the n points

$$\left(\Phi^{-1}\left(\frac{i-5}{n}\right), z_i\right) \text{ for } i=1,2,\ldots,n \tag{16.1.9}$$

If the original data $\{z_i\}$ are normal, but have an arbitrary mean μ and standard deviation σ, then the line $y=x$ will not match the expected theoretical quantiles. Clearly, the linear transformation

$$\mu + \sigma\xi_q \tag{16.1.10}$$

would provide the q^{th} theoretical quantile on the transformed scale. In practice, with a new data set $\{x_1, x_2, \ldots, x_n\}$,

the normal $q-q$ plot would consist of the n points

$$\left(\Phi^{-1}\left(\frac{i-5}{n}\right), x_i\right) \text{ for } i = 1, 2, \ldots, n \tag{16.1.11}$$

Instead of plotting the line $y = x$ as a reference line, the line

$$y = M + s \cdot x \tag{16.1.12}$$

should be composed, where M and s are the sample moments (mean and standard deviation) corresponding to the theoretical moments μ and σ. Alternatively, if the data are standardized, then the line $y = x$ would be appropriate, since now the sample mean would be 0 and the sample standard deviation would be 1.

Example 16.1.1: SAT Case Study

The SAT case study followed the academic achievements of 105 college students majoring in computer science. The first variable is their verbal SAT score and the second is their grade point average (GPA) at the university level. Before we compute inferential statistics using these variables, we should check if their distributions are normal. In Figure 16.1.13, we display the $q-q$ plots of the verbal SAT and university GPA variables.

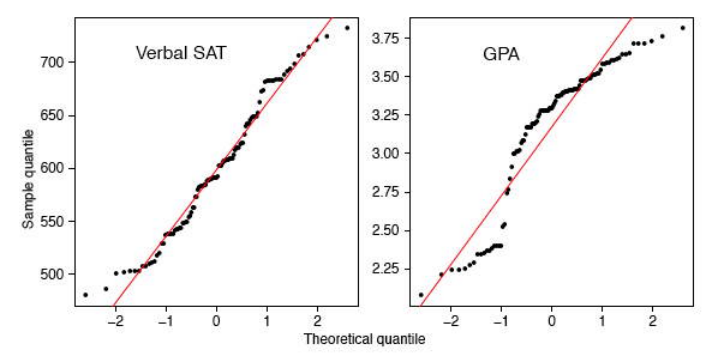


***Figure* 16.1.13:** *$q-q$ plots for the student data ($n = 105$).*

The verbal SAT seems to follow a normal distribution reasonably well, except in the extreme tails. However, the university GPA variable is highly non-normal. Compare the GPA $q-q$ plot to the simulation in the right frame of Figure 16.1.7. These figures are very similar, except for the region where $x \approx -1$. To follow these ideas, we computed histograms of the variables and their scatter diagram in Figure 16.1.14. These figures tell quite a different story. The university GPA is bimodal, with about 20% of the students falling into a separate cluster with a grade of C. The scatter diagram is quite unusual. While the students in this cluster all have below average verbal SAT scores, there are as many students with low SAT scores whose GPAs were quite respectable. We might speculate as to the cause(s): different distractions, different study habits, but it would only be speculation. But observe that the raw correlation between verbal SAT and GPA is a rather high 0.65, but when we exclude the cluster, the correlation for the remaining 86 students falls a little to 0.59.

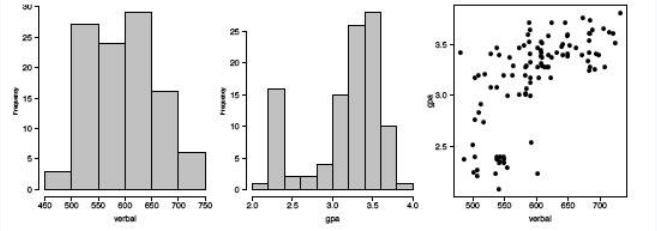


***Figure* 16.1.14:** *Histograms and scatter diagram of the verbal SAT and GPA variables for the 105 students.*

Discussion

Parametric modeling usually involves making assumptions about the shape of data, or the shape of residuals from a regression fit. Verifying such assumptions can take many forms, but an exploration of the shape using histograms and $q-q$ plots is very effective. The $q-q$ plot does not have any design parameters such as the number of bins for a histogram.

In an advanced treatment, the $q-q$ plot can be used to formally test the null hypothesis that the data are normal. This is done by computing the correlation coefficient of the n points in the $q-q$ plot. Depending upon n, the null hypothesis is rejected if the correlation coefficient is less than a threshold. The threshold is already quite close to 0.95 for modest sample sizes.

We have seen that the $q-q$ plot for uniform data is very closely related to the empirical cumulative distribution function. For general density functions, the so-called probability integral transform takes a random variable X and maps it to the interval $(0,1)$ through the CDF of X itself, that is,

$$Y = F_X(X) \tag{16.1.13}$$

which has been shown to be a uniform density. This explains why the $q-q$ plot on standardized data is always close to the line $y = x$ when the model is correct.

Finally, scientists have used special graph paper for years to make relationships linear (straight lines). The most common example used to be semi-log paper, on which points following the formula $y = ae^{bx}$ appear linear. This follows of course since $log(y) = log(a) + bx$, which is the equation for a straight line. The $q-q$ plots may be thought of as being "probability graph paper" that makes a plot of the ordered data values into a straight line. Every density has its own special probability graph paper.

Contributor

- Template:ContribHeblLane

This page titled 16.1: Q-Q Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

16.2: Contour Plots

Skills to Develop

- Describe a contour plot
- Interpret a contour plot

Contour plots portray data for three variables in two dimensions. The plot contains a number of contour lines. Each contour line is shown in an $X-Y$ plot and has a constant value on a third variable. Consider the Figure 16.2.1 that contains data on the fat, non-sugar carbohydrates, and calories present in a variety of breakfast cereals. Each line shows the carbohydrate and fat levels for cereals with the same number of calories. Note that the number of calories is not determined exactly by the fat and non-sugar carbohydrates since cereals also differ in sugar and protein.

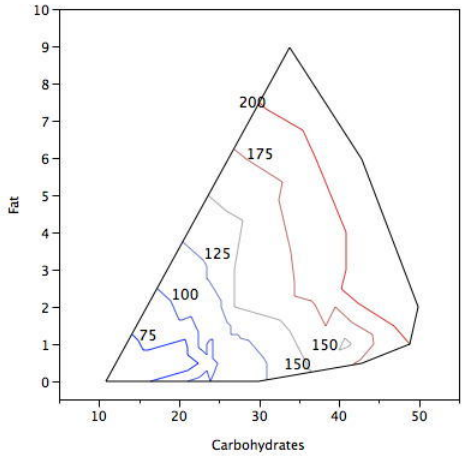


***Figure* 16.2.1:** *A contour plot showing calories as a function of fat and carbohydrates*

An alternative way to draw the plot is shown in Figure 16.2.2. The areas with the same number of calories are shaded.

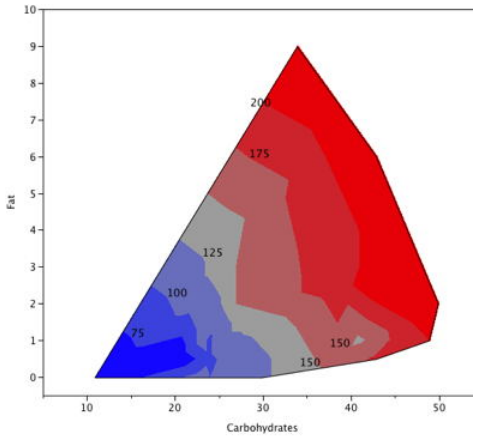


***Figure* 16.2.2:** *A contour plot showing calories as a function of fat and carbohydrates with areas shaded. An area represents values less than or equal to the label to the right of the area.*

Contour plots are used in many disciplines. For example, in cartography, contour lines can indicate areas of equal elevation, whereas in meteorological maps, contour lines can show equal temperature or equal barometric pressure.

The following links explain the drawing and interpretation of countour maps.

Application in Geology
Application in Meteorology

Contributor

- Template:ContribHeblLane

This page titled 16.2: Contour Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

16.3: 3D Plots

Skills to Develop

- Describe a $3D$ Plot
- Give an example of the value of a $3D$ plot

Just as two-dimensional scatter plots show the data in two dimensions, $3D$ plots show data in three dimensions. Figure 16.3.1 shows a $3D$ scatter plot of the fat, non-sugar carbohydrates, and calories from a variety of cereal types.

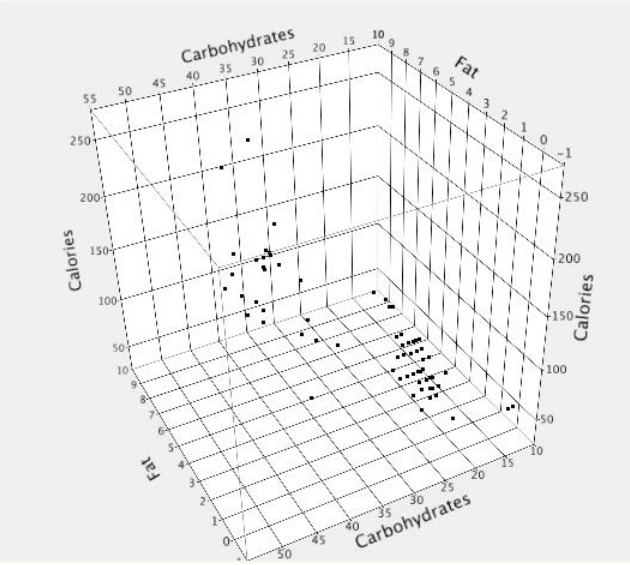


***Figure* 16.3.1:** *A $3D$ scatter plot showing fat, non-sugar carbohydrates, and calories from a variety of cereal types*

Many statistical packages allow you to rotate the axes interactively to view the data from a different vantage point. Figure 16.3.2 is an example.

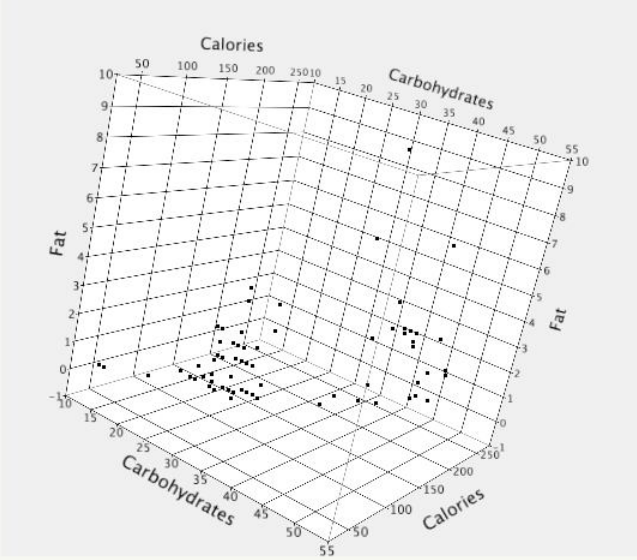


***Figure* 16.3.2:** *An alternative $3D$ scatter plot showing fat, non-sugar carbohydrates, and calories*

A fourth dimension can be represented as long as it is represented as a nominal variable. Figure 16.3.3 represents the different manufacturers by using different colors.

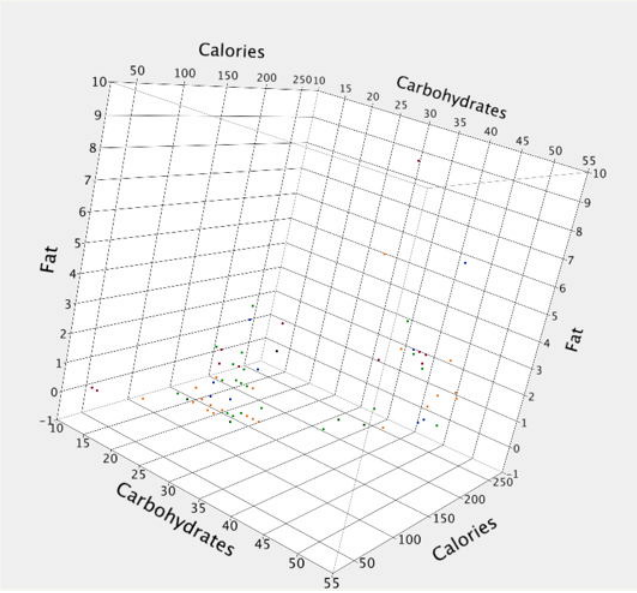


***Figure* 16.3.3:** *The different manufacturers are color coded*

Interactively rotating $3D$ plots can sometimes reveal aspects of the data not otherwise apparent. Figure 16.3.4 shows data from a pseudo random number generator. Figure 16.3.4 does not show anything systematic and the random number generator appears to generate data with properties similar to those of true random numbers.

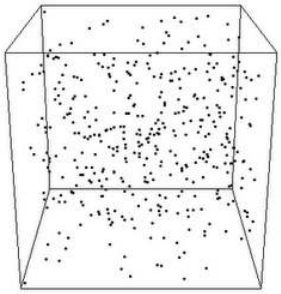


***Figure* 16.3.4:** *A* $3D$ *scatter plot showing* 400 *values of* X, Y, *and* Z *from a pseudo random number generator*

Figure 16.3.5 shows a different perspective on these data. Clearly they were not generated by a random process.

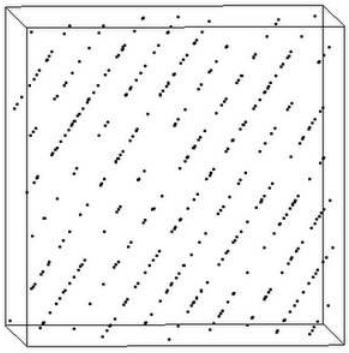


***Figure* 16.3.5:** *A different perspective on the* $3D$ *scatter plot showing* 400 *values of* X, Y, *and* Z *from a pseudo random number generator.*

Figures **16.3.4** and **16.3.5** are reproduced with permission from R snippets by Bogumil Kaminski.

Contributor

- Template:ContribHeblLane

This page titled 16.3: 3D Plots is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

16.4: Statistical Literacy

Skills to Develop

- Reading a Weather Map

This web page portrays altitudes in the United States.

Example 16.4.1: What do you think?

What part of the state of Texas (North, South, East, or West) contains the highest elevation?

Solution

West Texas.

Contributor

- Template:ContribHeblLane

This page titled 16.4: Statistical Literacy is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

16.5: Advanced Graphs (Exercises)

General Questions

Q1

What are $Q - Q$ plots useful for?

Q2

For the following data, plot the theoretically expected z score as a function of the actual z score (a $Q - Q$ plot).

0	1.2
0	1.2
0	1.2
0	1.2
0	1.3
0	1.3
0.1	1.3
0.1	1.3
0.1	1.3
0.1	1.4
0.1	1.4
0.2	1.5
0.2	1.6
0.3	1.7
0.3	1.7
0.4	1.7
0.5	1.8
0.6	1.8
0.6	1.9
0.6	1.9
0.6	2.0
0.6	2.0
0.6	2.0
0.6	2.1
0.6	2.1
0.6	2.1
0.7	2.1
0.7	2.1
0.8	2.1
0.8	2.1
0.8	2.3
0.8	2.5

0.8	2.7
0.9	3.0
1.0	4.2
1.0	5.0
1.1	5.7
1.1	12.4
1.2	15.2
1.2	

Q3

For the data in problem 2, describe how the data differ from a normal distribution.

Questions from Case studies

Q4

For the "SAT and College GPA" case study data, create a contour plot looking at University GPA as a function of Math SAT and High School GPA. Naturally, you should use a computer to do this.

Q5

For the "SAT and College GPA" case study data, create a $3D$ plot using the variables University GPA, Math SAT, and High School GPA. Naturally, you should use a computer to do this.

Contributor

- Template:ContribHeblLane

This page titled 16.5: Advanced Graphs (Exercises) is shared under a Public Domain license and was authored, remixed, and/or curated by David Lane.

Index

A

Absolute Risk Reduction (ARR)
9.2: Proportions
Adding probabilities
5.4: Two Basic Rules of Probability
ANOVA
15.2: One-Way ANOVA

B

bar graph
3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
4.8: Bar Charts
binomial probability distribution
10.4: A Population Proportion
blinding
2.5: Experimental Design and Ethics
box plots
3.5: Box Plots
4.6: Box Plots

C

Central Limit Theorem
7.2: The Sampling Distribution of the Sample Mean
cluster sampling
2.3: Data, Sampling, and Variation in Data and Sampling
coefficient of determination
14.4: The Regression Equation
Cohen's Standards
12.2: Two Population Means with Unknown Standard Deviations
Comparing two population means
12.3: Two Population Means with Known Standard Deviations
Comparing Two Population Proportions
12.4: Comparing Two Independent Population Proportions
complement
5.2: Terminology
5.3: Independent and Mutually Exclusive Events
conditional probability
5.2: Terminology
Confidence interval
11.1: Prelude to Hypothesis Testing
contingency table
5.5: Contingency Tables
13.4: Test of Independence
continuous data
2.3: Data, Sampling, and Variation in Data and Sampling
contour plots
16.2: Contour Plots
control group
2.5: Experimental Design and Ethics
cumulative distribution function
16.1: Q-Q Plots
cumulative relative frequency
2.4: Frequency, Frequency Tables, and Levels of Measurement

D

data frequency table
1.3: Presentation of Data
decision
11.5: Rare Events, the Sample, Decision and Conclusion
degrees of freedom
12.2: Two Population Means with Unknown Standard Deviations
direction of a relationship between the variables
14.3: Scatter Plots
discrete data
2.3: Data, Sampling, and Variation in Data and Sampling
Distribution for the differences
12.4: Comparing Two Independent Population Proportions
dot plot
4.10: Dot Plots

E

empirical CDF
16.1: Q-Q Plots
Equal variance
14.5: Testing the Significance of the Correlation Coefficient
ethics
2.5: Experimental Design and Ethics
event
5.2: Terminology
experimental unit
2.5: Experimental Design and Ethics
explanatory variable
2.5: Experimental Design and Ethics
extrapolation
14.6: Prediction

F

frequency
2.4: Frequency, Frequency Tables, and Levels of Measurement
Frequency Polygons
3.3: Histograms, Frequency Polygons, and Time Series Graphs
4.5: Frequency Polygons
frequency table
2.4: Frequency, Frequency Tables, and Levels of Measurement

G

goodness of fit
13.3: Goodness-of-Fit Test

H

histogram
16.1: Q-Q Plots
histograms
3.3: Histograms, Frequency Polygons, and Time Series Graphs
4.4: Histograms
homogeneity
13.5: Test for Homogeneity
hypothesis testing
11.1: Prelude to Hypothesis Testing
11.2: Null and Alternative Hypotheses
11.4: Distribution Needed for Hypothesis Testing
11.6: Additional Information and Full Hypothesis Test Examples

I

independent events
5.3: Independent and Mutually Exclusive Events
5.4: Two Basic Rules of Probability
13.4: Test of Independence
inferential statistics
10.1: Prelude to Confidence Intervals
Institutional Review Board
2.5: Experimental Design and Ethics
interpolation
14.6: Prediction

L

level of measurement
2.4: Frequency, Frequency Tables, and Levels of Measurement
line graph
3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
4.9: Line Graphs
linear correlation coefficient
14.4: The Regression Equation
14.5: Testing the Significance of the Correlation Coefficient
linear equations
14.2: Linear Equations
LINEAR REGRESSION MODEL
14.4: The Regression Equation
lurking variable
2.5: Experimental Design and Ethics

M

margin of error
10.2: A Single Population Mean using the Normal Distribution
matched samples
12.5: Matched or Paired Samples
mean
3.7: Skewness and the Mean, Median, and Mode
mean of the sample proportion
7.3: The Sample Proportion
median
3.4: Measures of the Location of the Data
3.6: Measures of the Center of the Data
3.7: Skewness and the Mean, Median, and Mode
mode
3.6: Measures of the Center of the Data
3.7: Skewness and the Mean, Median, and Mode

Multiplying probabilities
5.4: Two Basic Rules of Probability
mutually exclusive
5.3: Independent and Mutually Exclusive Events
5.4: Two Basic Rules of Probability

N

Normal distribution
6.3: Using the Normal Distribution
6.4: Normal Distribution - Lap Times (Worksheet)
6.5: Normal Distribution - Pinkie Length (Worksheet)
NOTESTURGES' RULE
4.4: Histograms

O

OpenStax
2.1: Introduction
2.2: Definitions of Statistics, Probability, and Key Terms
2.3: Data, Sampling, and Variation in Data and Sampling
2.4: Frequency, Frequency Tables, and Levels of Measurement
2.5: Experimental Design and Ethics
2.6: Data Collection Experiment (Worksheet)
2.7: Sampling Experiment (Worksheet)
3: Descriptive Statistics
3.1: Prelude to Descriptive Statistics
3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
3.3: Histograms, Frequency Polygons, and Time Series Graphs
3.6: Measures of the Center of the Data
3.7: Skewness and the Mean, Median, and Mode
3.8: Measures of the Spread of the Data
3.9: Descriptive Statistics (Worksheet)
5: Probability Topics
5.1: Introduction
5.2: Terminology
5.3: Independent and Mutually Exclusive Events
5.4: Two Basic Rules of Probability
5.5: Contingency Tables
5.6: Tree and Venn Diagrams
5.7: Probability Topics (Worksheet)
6: The Normal Distribution
6.1: Prelude to The Normal Distribution
6.2: The Standard Normal Distribution
6.3: Using the Normal Distribution
6.4: Normal Distribution - Lap Times (Worksheet)
6.5: Normal Distribution - Pinkie Length (Worksheet)
10: Confidence Intervals
10.1: Prelude to Confidence Intervals
10.2: A Single Population Mean using the Normal Distribution
10.3: A Single Population Mean using the Student t-Distribution
10.4: A Population Proportion
10.5: Confidence Interval - Home Costs (Worksheet)
10.6: Confidence Interval -Place of Birth (Worksheet)
10.7: Confidence Interval -Women's Heights (Worksheet)
11: Hypothesis Testing with One Sample
11.1: Prelude to Hypothesis Testing
11.2: Null and Alternative Hypotheses
11.3: Outcomes and the Type I and Type II Errors
11.4: Distribution Needed for Hypothesis Testing
11.5: Rare Events, the Sample, Decision and Conclusion
11.6: Additional Information and Full Hypothesis Test Examples
11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet)
12: Hypothesis Testing with Two Samples
12.1: Prelude to Hypothesis Testing with Two Samples
12.2: Two Population Means with Unknown Standard Deviations
12.3: Two Population Means with Known Standard Deviations
12.4: Comparing Two Independent Population Proportions
12.5: Matched or Paired Samples
12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet)
13: The Chi-Square Distribution
13.1: Prelude to The Chi-Square Distribution
13.2: Facts About the Chi-Square Distribution
13.3: Goodness-of-Fit Test
13.4: Test of Independence
13.5: Test for Homogeneity
13.6: Comparison of the Chi-Square Tests
13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet)
13.9: Lab 2: Chi-Square Test of Independence (Worksheet)
14: Linear Regression and Correlation
14.1: Prelude to Linear Regression and Correlation
14.2: Linear Equations
14.3: Scatter Plots
14.4: The Regression Equation
14.5: Testing the Significance of the Correlation Coefficient
14.6: Prediction
14.7: Outliers
14.8: Regression - Distance from School (Worksheet)
14.9: Regression - Textbook Cost (Worksheet)
14.10: Regression - Fuel Efficiency (Worksheet)
15: F Distribution and One-Way ANOVA
15.1: Prelude to F Distribution and One-Way ANOVA
15.2: One-Way ANOVA
15.3: The F Distribution and the F-Ratio
15.4: Facts About the F Distribution
15.5: Test of Two Variances
15.6: Lab: One-Way ANOVA
outcome
5.2: Terminology
outliers
3.4: Measures of the Location of the Data
14.7: Outliers

P

Paired Samples
12.5: Matched or Paired Samples
parameter
2.2: Definitions of Statistics, Probability, and Key Terms
Pareto chart
2.3: Data, Sampling, and Variation in Data and Sampling
placebo
2.5: Experimental Design and Ethics
Pooled Proportion
12.4: Comparing Two Independent Population Proportions
pooled variance
15.3: The F Distribution and the F-Ratio
population
2.2: Definitions of Statistics, Probability, and Key Terms
population mean
3.6: Measures of the Center of the Data
Population Standard Deviation
3.8: Measures of the Spread of the Data
power of the test
8.1: Introduction to Power
8.5: Factors Affecting Power
11.3: Outcomes and the Type I and Type II Errors
11.6: Additional Information and Full Hypothesis Test Examples
prediction
14.6: Prediction
probability
2.2: Definitions of Statistics, Probability, and Key Terms

probability distribution function
6.3: Using the Normal Distribution
Proportion of Variance
9.4: Proportion of Variance Explained

Q

qualitative data
2.3: Data, Sampling, and Variation in Data and Sampling
quantitative data
2.3: Data, Sampling, and Variation in Data and Sampling
quartiles
3.4: Measures of the Location of the Data

R

random assignment
2.5: Experimental Design and Ethics
rare events
11.5: Rare Events, the Sample, Decision and Conclusion
response variable
2.5: Experimental Design and Ethics
rounding
2.4: Frequency, Frequency Tables, and Levels of Measurement

S

sample mean
3.6: Measures of the Center of the Data
7.1: The Mean and Standard Deviation of the Sample Mean
sample proportion
7.3: The Sample Proportion
sample space
5.2: Terminology
sample Standard Deviation
3.8: Measures of the Spread of the Data
7.1: The Mean and Standard Deviation of the Sample Mean
sampling
2: Sampling and Data
Sampling Bias
2.3: Data, Sampling, and Variation in Data and Sampling
sampling distribution
7.3: The Sample Proportion
sampling error
2.3: Data, Sampling, and Variation in Data and Sampling
sampling with replacement
2.3: Data, Sampling, and Variation in Data and Sampling
5.3: Independent and Mutually Exclusive Events
5.6: Tree and Venn Diagrams
sampling without replacement
2.3: Data, Sampling, and Variation in Data and Sampling
5.3: Independent and Mutually Exclusive Events
5.6: Tree and Venn Diagrams
scatter plot
14.3: Scatter Plots
set notation
1.3: Presentation of Data
significance level
11.5: Rare Events, the Sample, Decision and Conclusion
Skewed
3.5: Box Plots
3.7: Skewness and the Mean, Median, and Mode
slope
14.2: Linear Equations
standard deviation
3.8: Measures of the Spread of the Data
standard deviation of the sample proportion
7.3: The Sample Proportion
standard error
12.2: Two Population Means with Unknown Standard Deviations
standard normal distribution
6.1: Prelude to The Normal Distribution
6.2: The Standard Normal Distribution
statistic
2.2: Definitions of Statistics, Probability, and Key Terms
stemplot
3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
strength of a relationship between the variables
14.3: Scatter Plots
Sturges' rule
4.4: Histograms

T

test for homogeneity
13.5: Test for Homogeneity
test statistic
12.5: Matched or Paired Samples
The alternative hypothesis
11.2: Null and Alternative Hypotheses
The AND Event
5.2: Terminology
The null hypothesis
11.2: Null and Alternative Hypotheses
The Or Event
5.2: Terminology
The OR of Two Events
5.3: Independent and Mutually Exclusive Events
Time Series Graphs
3.3: Histograms, Frequency Polygons, and Time Series Graphs
treatments
2.5: Experimental Design and Ethics
tree diagram
5.6: Tree and Venn Diagrams
type I error
11.3: Outcomes and the Type I and Type II Errors
type II error
11.3: Outcomes and the Type I and Type II Errors

V

Variable
2.2: Definitions of Statistics, Probability, and Key Terms
variation due to error or unexplained variation
15.3: The F Distribution and the F-Ratio
variation due to treatment or explained variation
15.3: The F Distribution and the F-Ratio
Venn diagram
5.6: Tree and Venn Diagrams

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Social Science Statistics

Webpages: 161

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 66.5% (107 pages)
- Public Domain: 21.1% (34 pages)
- Undeclared: 6.2% (10 pages)
- CC BY-NC-SA 4.0: 6.2% (10 pages)

By Page

- Social Science Statistics - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistics - *CC BY-NC-SA 4.0*
    - 1.1: Basic Definitions and Concepts - *CC BY-NC-SA 4.0*
    - 1.2: Overview - *CC BY-NC-SA 4.0*
    - 1.3: Presentation of Data - *CC BY-NC-SA 4.0*
    - 1.4: Introduction to Statistics (Exercises) - *CC BY-NC-SA 4.0*
  - 2: Sampling and Data - *CC BY 4.0*
    - 2.1: Introduction - *CC BY 4.0*
    - 2.2: Definitions of Statistics, Probability, and Key Terms - *CC BY 4.0*
    - 2.3: Data, Sampling, and Variation in Data and Sampling - *CC BY 4.0*
    - 2.4: Frequency, Frequency Tables, and Levels of Measurement - *CC BY 4.0*
    - 2.5: Experimental Design and Ethics - *CC BY 4.0*
    - 2.6: Data Collection Experiment (Worksheet) - *CC BY 4.0*
    - 2.7: Sampling Experiment (Worksheet) - *CC BY 4.0*
    - 2.8: Sampling and Data (Exercises) - *CC BY 4.0*
  - 3: Descriptive Statistics - *CC BY 4.0*
    - 3.1: Prelude to Descriptive Statistics - *CC BY 4.0*
    - 3.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs - *CC BY 4.0*
    - 3.3: Histograms, Frequency Polygons, and Time Series Graphs - *CC BY 4.0*
    - 3.4: Measures of the Location of the Data - *CC BY 4.0*
      - Measures of the Location of the Data (Exercises) - *CC BY 4.0*
    - 3.5: Box Plots - *CC BY 4.0*
    - 3.6: Measures of the Center of the Data - *CC BY 4.0*
    - 3.7: Skewness and the Mean, Median, and Mode - *CC BY 4.0*
    - 3.8: Measures of the Spread of the Data - *CC BY 4.0*
    - 3.9: Descriptive Statistics (Worksheet) - *CC BY 4.0*
    - 3.10: Descriptive Statistics (Exercises) - *CC BY 4.0*
  - 4: Graphing Distributions - *Public Domain*
    - 4.1: Graphing Qualitative Variables - *Public Domain*
    - 4.2: Quantitative Variables - *Public Domain*
    - 4.3: Stem and Leaf Displays - *Public Domain*
    - 4.4: Histograms - *Public Domain*
    - 4.5: Frequency Polygons - *Public Domain*
    - 4.6: Box Plots - *Public Domain*
    - 4.7: Box Plot Demo - *Public Domain*
    - 4.8: Bar Charts - *Public Domain*
    - 4.9: Line Graphs - *Public Domain*
    - 4.10: Dot Plots - *Public Domain*
    - 4.11: Statistical Literacy - *Public Domain*
    - 4.12: Graphing Distributions (Exercises) - *Public Domain*
  - 5: Probability Topics - *CC BY 4.0*
    - 5.1: Introduction - *CC BY 4.0*
    - 5.2: Terminology - *CC BY 4.0*
    - 5.3: Independent and Mutually Exclusive Events - *CC BY 4.0*
    - 5.4: Two Basic Rules of Probability - *CC BY 4.0*
    - 5.5: Contingency Tables - *CC BY 4.0*
    - 5.6: Tree and Venn Diagrams - *CC BY 4.0*

- 5.7: Probability Topics (Worksheet) - *CC BY 4.0*
- 5.8: Probability Topics (Exericses) - *CC BY 4.0*

- 6: The Normal Distribution - *CC BY 4.0*
  - 6.1: Prelude to The Normal Distribution - *CC BY 4.0*
  - 6.2: The Standard Normal Distribution - *CC BY 4.0*
    - The Standard Normal Distribution (Exercises) - *CC BY 4.0*
  - 6.3: Using the Normal Distribution - *CC BY 4.0*
  - 6.4: Normal Distribution - Lap Times (Worksheet) - *CC BY 4.0*
  - 6.5: Normal Distribution - Pinkie Length (Worksheet) - *CC BY 4.0*
  - 6.6: The Normal Distribution (Exercises) - *CC BY 4.0*
- 7: Sampling Distributions - *CC BY-NC-SA 4.0*
  - 7.1: The Mean and Standard Deviation of the Sample Mean - *CC BY-NC-SA 4.0*
  - 7.2: The Sampling Distribution of the Sample Mean - *CC BY-NC-SA 4.0*
  - 7.3: The Sample Proportion - *CC BY-NC-SA 4.0*
  - 7.4: Sampling Distributions (Exercises) - *CC BY-NC-SA 4.0*
- 8: Power - *Public Domain*
  - 8.1: Introduction to Power - *Public Domain*
  - 8.2: Example Calculations - *Public Domain*
  - 8.3: Power Demo - *Public Domain*
  - 8.4: Power Demo II - *Public Domain*
  - 8.5: Factors Affecting Power - *Public Domain*
  - 8.6: Statistical Literacy - *Public Domain*
  - 8.7: Power (Exercises) - *Public Domain*
- 9: Effect Size - *Public Domain*
  - 9.1: Prelude to Effect Size - *Public Domain*
  - 9.2: Proportions - *Public Domain*
  - 9.3: Difference Between Two Means - *Public Domain*
  - 9.4: Proportion of Variance Explained - *Public Domain*
  - 9.5: Statistical Literacy - *Public Domain*
  - 9.6: Effect Size (Exercises) - *Public Domain*
- 10: Confidence Intervals - *CC BY 4.0*
  - 10.1: Prelude to Confidence Intervals - *CC BY 4.0*
  - 10.2: A Single Population Mean using the Normal Distribution - *CC BY 4.0*
    - A Single Population Mean using the Normal Distribution (Exercises) - *CC BY 4.0*
  - 10.3: A Single Population Mean using the Student t-Distribution - *CC BY 4.0*
  - 10.4: A Population Proportion - *CC BY 4.0*
  - 10.5: Confidence Interval - Home Costs (Worksheet) - *CC BY 4.0*
  - 10.6: Confidence Interval -Place of Birth (Worksheet) - *CC BY 4.0*
  - 10.7: Confidence Interval -Women's Heights (Worksheet) - *CC BY 4.0*
  - 10.8: Confidence Intervals (Exercises) - *CC BY 4.0*
  - 10.9: Confidence Intervals (Summary) - *CC BY 4.0*
- 11: Hypothesis Testing with One Sample - *CC BY 4.0*
  - 11.1: Prelude to Hypothesis Testing - *CC BY 4.0*
  - 11.2: Null and Alternative Hypotheses - *CC BY 4.0*
    - Null and Alternative Hypotheses (Exercises) - *CC BY 4.0*
  - 11.3: Outcomes and the Type I and Type II Errors - *CC BY 4.0*
    - Outcomes and the Type I and Type II Errors (Exercises) - *CC BY 4.0*
  - 11.4: Distribution Needed for Hypothesis Testing - *CC BY 4.0*
    - Distribution Needed for Hypothesis Testing (Exercises) - *CC BY 4.0*
  - 11.5: Rare Events, the Sample, Decision and Conclusion - *CC BY 4.0*
    - Rare Events, the Sample, Decision and Conclusion (Exercises) - *CC BY 4.0*
  - 11.6: Additional Information and Full Hypothesis Test Examples - *CC BY 4.0*
  - 11.7: Hypothesis Testing of a Single Mean and Single Proportion (Worksheet) - *CC BY 4.0*
  - 11.8: Hypothesis Testing with One Sample (Exercises) - *CC BY 4.0*
- 12: Hypothesis Testing with Two Samples - *CC BY 4.0*
  - 12.1: Prelude to Hypothesis Testing with Two Samples - *CC BY 4.0*
  - 12.2: Two Population Means with Unknown Standard Deviations - *CC BY 4.0*
  - 12.3: Two Population Means with Known Standard Deviations - *CC BY 4.0*
  - 12.4: Comparing Two Independent Population Proportions - *CC BY 4.0*
  - 12.5: Matched or Paired Samples - *CC BY 4.0*
  - 12.6: Hypothesis Testing for Two Means and Two Proportions (Worksheet) - *CC BY 4.0*
  - 12.7: Hypothesis Testing with Two Samples (Exercises) - *CC BY 4.0*
- 13: The Chi-Square Distribution - *CC BY 4.0*

- 13.1: Prelude to The Chi-Square Distribution - *CC BY 4.0*
- 13.2: Facts About the Chi-Square Distribution - *CC BY 4.0*
- 13.3: Goodness-of-Fit Test - *CC BY 4.0*
- 13.4: Test of Independence - *CC BY 4.0*
- 13.5: Test for Homogeneity - *CC BY 4.0*
- 13.6: Comparison of the Chi-Square Tests - *CC BY 4.0*
- 13.7: Test of a Single Variance - *CC BY 4.0*
- 13.8: Lab 1: Chi-Square Goodness-of-Fit (Worksheet) - *CC BY 4.0*
- 13.9: Lab 2: Chi-Square Test of Independence (Worksheet) - *CC BY 4.0*
- 13.10: The Chi-Square Distribution (Exercises) - *CC BY 4.0*

14: Linear Regression and Correlation - *CC BY 4.0*

- 14.1: Prelude to Linear Regression and Correlation - *CC BY 4.0*
- 14.2: Linear Equations - *CC BY 4.0*
  - Linear Equations (Exercises) - *CC BY 4.0*
- 14.3: Scatter Plots - *CC BY 4.0*
  - Scatter Plots (Exercises) - *CC BY 4.0*
- 14.4: The Regression Equation - *CC BY 4.0*
  - The Regression Equation (Exercise) - *CC BY 4.0*
- 14.5: Testing the Significance of the Correlation Coefficient - *CC BY 4.0*
  - Testing the Significance of the Correlation Coefficient (Exercises) - *CC BY 4.0*
- 14.6: Prediction - *CC BY 4.0*
  - Prediction (Exercises) - *CC BY 4.0*
- 14.7: Outliers - *CC BY 4.0*
  - Outliers (Exercises) - *CC BY 4.0*
- 14.8: Regression - Distance from School (Worksheet) - *CC BY 4.0*
- 14.9: Regression - Textbook Cost (Worksheet) - *CC BY 4.0*
- 14.10: Regression - Fuel Efficiency (Worksheet) - *CC BY 4.0*
- 14.11: Linear Regression and Correlation (Exercises) - *CC BY 4.0*

15: F Distribution and One-Way ANOVA - *CC BY 4.0*

- 15.1: Prelude to F Distribution and One-Way ANOVA - *CC BY 4.0*
- 15.2: One-Way ANOVA - *CC BY 4.0*
- 15.3: The F Distribution and the F-Ratio - *CC BY 4.0*
- 15.4: Facts About the F Distribution - *CC BY 4.0*
- 15.5: Test of Two Variances - *CC BY 4.0*
- 15.6: Lab: One-Way ANOVA - *CC BY 4.0*
- 15.7: F Distribution and One-Way ANOVA (Exercises) - *CC BY 4.0*

16: Advanced Graphs - *Public Domain*

- 16.1: Q-Q Plots - *Public Domain*
- 16.2: Contour Plots - *Public Domain*
- 16.3: 3D Plots - *Public Domain*
- 16.4: Statistical Literacy - *Public Domain*
- 16.5: Advanced Graphs (Exercises) - *Public Domain*

Back Matter - *Undeclared*

- Index - *Undeclared*
- Glossary - *Undeclared*
- Detailed Licensing - *Undeclared*