INTRODUCTORY STATISTICS



Shafer and Zhang
LibreTexts



Introductory Statistics

Shafer and Zhang

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 06/17/2026

TABLE OF CONTENTS

Licensing

1: Introduction to Statistics

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

2: Descriptive Statistics

- 2.1: Three Popular Data Displays
- 2.2: Measures of Central Location - Three Kinds of Averages
- 2.3: Measures of Variability
- 2.4: Relative Position of Data
- 2.5: The Empirical Rule and Chebyshev's Theorem
- 2.E: Descriptive Statistics (Exercises)

3: Basic Concepts of Probability

- 3.1: Sample Spaces, Events, and Their Probabilities
- 3.2: Complements, Intersections, and Unions
- 3.3: Conditional Probability and Independent Events
- 3.E: Basic Concepts of Probability (Exercises)

4: Discrete Random Variables

- 4.1: Random Variables
- 4.2: Probability Distributions for Discrete Random Variables
- 4.3: The Binomial Distribution
- 4.E: Discrete Random Variables (Exercises)

5: Continuous Random Variables

- 5.1: Continuous Random Variables
- 5.2: The Standard Normal Distribution
- 5.3: Probability Computations for General Normal Random Variables
- 5.4: Areas of Tails of Distributions
- 5.E: Continuous Random Variables (Exercises)

6: Sampling Distributions

- 6.1: The Mean and Standard Deviation of the Sample Mean
- 6.2: The Sampling Distribution of the Sample Mean
- 6.3: The Sample Proportion
- 6.E: Sampling Distributions (Exercises)

7: Estimation

- 7.1: Large Sample Estimation of a Population Mean
- 7.2: Small Sample Estimation of a Population Mean

- 7.3: Large Sample Estimation of a Population Proportion
- 7.4: Sample Size Considerations
- 7.E: Estimation (Exercises)

8: Testing Hypotheses

- 8.1: The Elements of Hypothesis Testing
- 8.2: Large Sample Tests for a Population Mean
- 8.3: The Observed Significance of a Test
- 8.4: Small Sample Tests for a Population Mean
- 8.5: Large Sample Tests for a Population Proportion
- 8.E: Testing Hypotheses (Exercises)

9: Two-Sample Problems

- 9.1: Comparison of Two Population Means- Large, Independent Samples
- 9.2: Comparison of Two Population Means - Small, Independent Samples
- 9.3: Comparison of Two Population Means - Paired Samples
- 9.4: Comparison of Two Population Proportions
- 9.5: Sample Size Considerations
- 9.E: Two-Sample Problems (Exercises)

10: Correlation and Regression

- 10.1: Linear Relationships Between Variables
- 10.2: The Linear Correlation Coefficient
- 10.3: Modelling Linear Relationships with Randomness Present
- 10.4: The Least Squares Regression Line
- 10.5: Statistical Inferences About β_1
- 10.6: The Coefficient of Determination
- 10.7: Estimation and Prediction
- 10.8: A Complete Example
- 10.9: Formula List
- 10.E: Correlation and Regression (Exercises)

11: Chi-Square Tests and F-Tests

- 11.1: Chi-Square Tests for Independence
- 11.2: Chi-Square One-Sample Goodness-of-Fit Tests
- 11.3: F-tests for Equality of Two Variances
- 11.4: F-Tests in One-Way ANOVA
- 11.E: Chi-Square Tests and F-Tests (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

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.E: Introduction to Statistics (Exercises)

This page titled 1: Introduction to Statistics is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

1.1: Basic Definitions and Concepts

Learning Objectives

- 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.

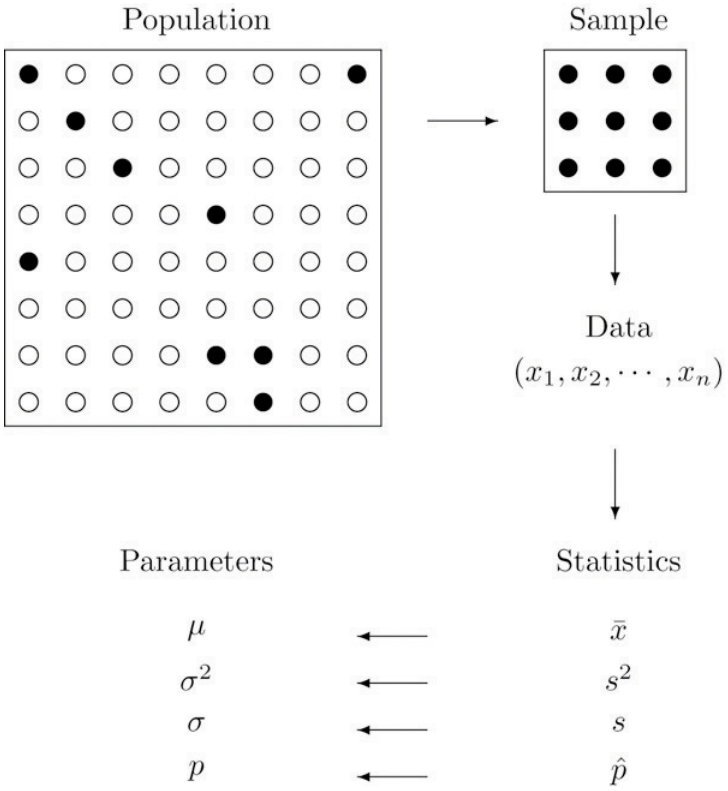


Figure 1.1.1: The Grand Picture of Statistics

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.

This page titled 1.1: Basic Definitions and Concepts is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

1.2: Overview

Learning Objectives

- 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.

This page titled 1.2: Overview is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Presentation of Data

Learning Objectives

- 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:

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

or in set notation as:

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

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

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.

This page titled 1.3: Presentation of Data is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

1.E: 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 $\text{January } 1, 1842$.
    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.

This page titled 1.E: Introduction to Statistics (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

2: Descriptive Statistics

Statistics naturally divides into two branches, descriptive statistics and inferential statistics. Our main interest is in inferential statistics to try to infer from the data what the population might thin or to evaluate the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Nevertheless, the starting point for dealing with a collection of data is to organize, display, and summarize it effectively. These are the objectives of descriptive statistics, the topic of this chapter.

2.1: Three Popular Data Displays
2.2: Measures of Central Location - Three Kinds of Averages
2.3: Measures of Variability
2.4: Relative Position of Data
2.5: The Empirical Rule and Chebyshev's Theorem
2.E: Descriptive Statistics (Exercises)

This page titled 2: Descriptive Statistics is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.1: Three Popular Data Displays

Learning Objectives

- To learn to interpret the meaning of three graphical representations of sets of data: stem and leaf diagrams, frequency histograms, and relative frequency histograms.

A well-known adage is that "a picture is worth a thousand words." This saying proves true when it comes to presenting statistical information in a data set. There are many effective ways to present data graphically. The three graphical tools that are introduced in this section are among the most commonly used and are relevant to the subsequent presentation of the material in this book.

Stem and Leaf Diagrams

Suppose 30 students in a statistics class took a test and made the following scores:

86	80	25	77	73	76	100	90	69	93
90	83	70	73	73	70	90	83	71	95
40	58	68	69	100	78	87	97	92	74

How did the class do on the test? A quick glance at the set of 30 numbers does not immediately give a clear answer. However the data set may be reorganized and rewritten to make relevant information more visible. One way to do so is to construct a stem and leaf diagram as shown in Figure $2.1.1$ The numbers in the tens place, from 2 through 9, and additionally the number 10, are the "stems," and are arranged in numerical order from top to bottom to the left of a vertical line. The number in the units place in each measurement is a "leaf," and is placed in a row to the right of the corresponding stem, the number in the tens place of that measurement. Thus the three leaves $9, 8, \text{and } 9$ in the row headed with the stem 6 correspond to the three exam scores in the $60s, 69$ (in the first row of data), 68 (in the third row), and 69 (also in the third row).

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

Figure $2.1.1$: *Stem and Leaf Diagram*

The display is made even more useful for some purposes by rearranging the leaves in numerical order, as shown in Figure $2.1.2$. Either way, with the data reorganized certain information of interest becomes apparent immediately. There are two perfect scores; three students made scores under 60; most students scored in the $70s, 80s \text{ and } 90s$; and the overall average is probably in the high $70s$ or low $80s$.

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

Figure $2.1.2$: *Ordered Stem and Leaf Diagram*

In this example the scores have a natural stem (the tens place) and leaf (the ones place). One could spread the diagram out by splitting each tens place number into lower and upper categories. For example, all the scores in the $80s$ may be represented on two separate stems, lower $80s$ and upper $80s$:

8 | 0 3 3

8	6 7

The definitions of stems and leaves are flexible in practice. The general purpose of a stem and leaf diagram is to provide a quick display of how the data are distributed across the range of their values; some improvisation could be necessary to obtain a diagram that best meets that goal.

Note that all of the original data can be recovered from the stem and leaf diagram. This will not be true in the next two types of graphical displays.

Frequency Histograms

The stem and leaf diagram is not practical for large data sets, so we need a different, purely graphical way to represent data. A frequency histogram is such a device. We will illustrate it using the same data set from the previous subsection. For the 30 scores on the exam, it is natural to group the scores on the standard ten-point scale, and count the number of scores in each group. Thus there are two $100s$, seven scores in the $90s$, six in the $80s$, and so on. We then construct the diagram shown in Figure $2.1.3$ by drawing for each group, or class, a vertical bar whose length is the number of observations in that group. In our example, the bar labeled 100 is 2 units long, the bar labeled 90 is 7 units long, and so on. While the individual data values are lost, we know the number in each class. This number is called the frequency of the class, hence the name frequency histogram.

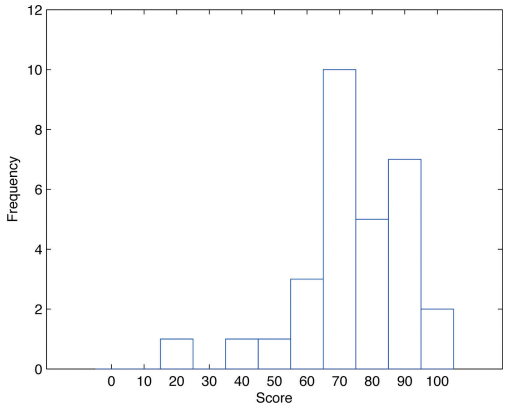


Figure $2.1.3$: *Frequency Histogram*

The same procedure can be applied to any collection of numerical data. Observations are grouped into several classes and the frequency (the number of observations) of each class is noted. These classes are arranged and indicated in order on the horizontal axis (called the x-axis), and for each group a vertical bar, whose length is the number of observations in that group, is drawn. The resulting display is a frequency histogram for the data. The similarity in Figure $2.1.1$ and Figure $2.1.3$ is apparent, particularly if you imagine turning the stem and leaf diagram on its side by rotating it a quarter turn counterclockwise.

Definition

In general, the definition of the classes in the frequency histogram is flexible. The general purpose of a frequency histogram is very much the same as that of a stem and leaf diagram, to provide a graphical display that gives a sense of data distribution across the range of values that appear.

We will not discuss the process of constructing a histogram from data since in actual practice it is done automatically with statistical software or even handheld calculators.

Relative Frequency Histograms

In our example of the exam scores in a statistics class, five students scored in the $80s$. The number 5 is the frequency of the group labeled "$80s$." Since there are 30 students in the entire statistics class, the proportion who scored in the $80s$ is $5/30$. The number $5/30$, which could also be expressed as $0.1\overline{6} \approx .1667$, or as 16.67%, is the relative frequency of the group labeled "$80s$." Every group (the $70s$, the $80s$, and so on) has a relative frequency. We can thus construct a diagram by drawing for each group, or class, a vertical bar whose length is the relative frequency of that group. For example, the bar for the $80s$ will have length $5/30$ unit, not 5 units. The diagram is a relative frequency histogram for the data, and is shown in Figure $2.1.4$. It is exactly the same as the frequency histogram except that the vertical axis in the relative frequency histogram is not frequency but relative frequency.

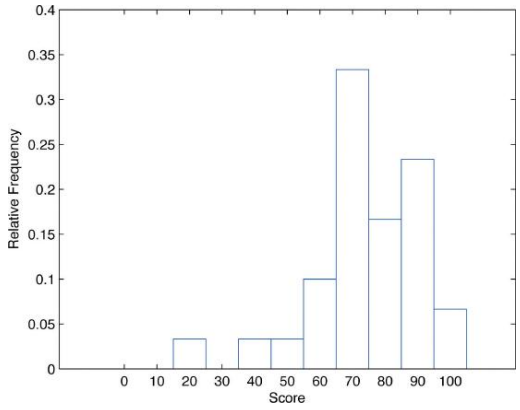
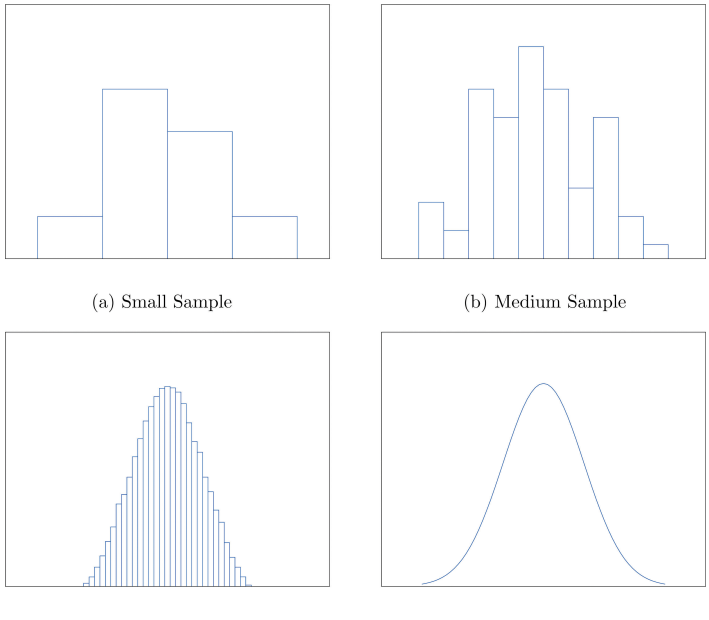


Figure 2.1.4: *Relative Frequency Histogram*

The same procedure can be applied to any collection of numerical data. Classes are selected, the relative frequency of each class is noted, the classes are arranged and indicated in order on the horizontal axis, and for each class a vertical bar, whose length is the relative frequency of the class, is drawn. The resulting display is a relative frequency histogram for the data. A key point is that now if each vertical bar has width 1 unit, then the total area of all the bars is 1 or 100%.

Although the histograms in Figure 2.1.3 and Figure 2.1.4 have the same appearance, the relative frequency histogram is more important for us, and it will be relative frequency histograms that will be used repeatedly to represent data in this text. To see why this is so, reflect on what it is that you are actually seeing in the diagrams that quickly and effectively communicates information to you about the data. It is the relative sizes of the bars. The bar labeled "70s" in either figure takes up 1/3 of the total area of all the bars, and although we may not think of this consciously, we perceive the proportion 1/3 in the figures, indicating that a third of the grades were in the 70s. The relative frequency histogram is important because the labeling on the vertical axis reflects what is important visually: the relative sizes of the bars.



(a) Small Sample (b) Medium Sample

(c) Large Sample (d) Very Large Sample

Figure 2.1.5: *Sample Size and Relative Frequency Histogram*

When the size n of a sample is small only a few classes can be used in constructing a relative frequency histogram. Such a histogram might look something like the one in panel (a) of Figure 2.1.5. If the sample size n were increased, then more classes could be used in constructing a relative frequency histogram and the vertical bars of the resulting histogram would be finer, as indicated in panel (b) of Figure 2.1.5. For a very large sample the relative frequency histogram would look very fine, like the one in (c) of Figure 2.1.5. If the sample size were to increase indefinitely then the corresponding relative frequency histogram would be so fine that it would look like a smooth curve, such as the one in panel (d) of Figure 2.1.5.

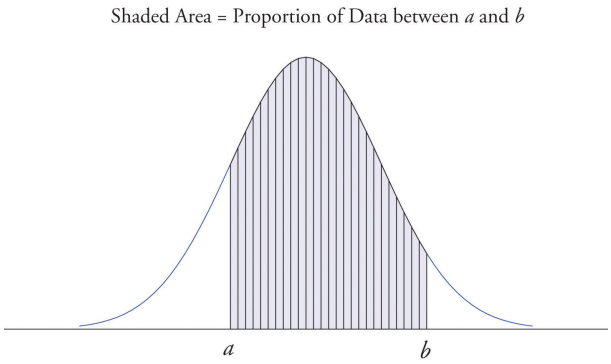


Figure 2.1.6: *A Very Fine Relative Frequency Histogram*

It is common in statistics to represent a population or a very large data set by a smooth curve. It is good to keep in mind that such a curve is actually just a very fine relative frequency histogram in which the exceedingly narrow vertical bars have disappeared. Because the area of each such vertical bar is the proportion of the data that lies in the interval of numbers over which that bar stands, this means that for any two numbers a and b, the proportion of the data that lies between the two numbers a and b is the area under the curve that is above the interval (a, b) in the horizontal axis. This is the area shown in Figure 2.1.6. In particular the total area under the curve is 1, or 100%.

Key Takeaway

- Graphical representations of large data sets provide a quick overview of the nature of the data.
- A population or a very large data set may be represented by a smooth curve. This curve is a very fine relative frequency histogram in which the exceedingly narrow vertical bars have been omitted.
- When a curve derived from a relative frequency histogram is used to describe a data set, the proportion of data with values between two numbers a and b is the area under the curve between a and b, as illustrated in Figure 2.1.6.

This page titled 2.1: Three Popular Data Displays is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Measures of Central Location - Three Kinds of Averages

Learning Objectives

- To learn the concept of the “center” of a data set.
- To learn the meaning of each of three measures of the center of a data set—the mean, the median, and the mode—and how to compute each one.

This section is be titled “three kinds of averages” because any kind of average could be used to answer the question "where is the center of the data?". We will see that the nature of the data set, as indicated by a relative frequency histogram, will determine what constitutes a good answer. Different shapes of the histogram call for different measures of central location.

The Mean

The first measure of central location is the usual “average” that is familiar to everyone: add up all the values, then divide by the number of values. Before writing a formula for the mean let us introduce some handy mathematical notation.

notations: $\sum$ "sum" and n "sample size"

The Greek letter $\sum$, pronounced "sigma", is a handy mathematical shorthand that stands for "add up all the values" or "sum". For example $\sum x$ means "add up all the values of x", and $\sum x^2$ means "add up all the values of x^2". In these expressions x usually stands for a value of the data, so $\sum x$ stands for "the sum of all the data values" and $\sum x^2$ means "the sum of the squares of all the data values".

$\mathbf{n}$ stands for the *sample size*, the number of data values. An example will help make this clear.

Example 2.2.1

Find n, $\sum x$, $\sum x^2$ and $\sum (x-1)^2$ for the data:

$$1,\ 3,\ 4$$

Solution

$$\begin{aligned} n &= 3 \quad \text{because there are three data values} \\ \sum x &= 1+3+4=8 \\ \sum x^2 &= 1^2+3^2+4^2=1+9+16=26 \\ \sum (x-1)^2 &= (1-1)^2+(3-1)^2+(4-1)^2=0^2+2^2+3^2=13 \end{aligned}$$

Using these handy notations it's easy to write a formula defining the mean $\bar{x}$ of a sample.

Definition: Sample Mean

The *sample mean* of a set of n sample data values is the number $\bar{x}$ defined by the formula

$$\bar{x} = \frac{\sum x}{n} \tag{2.2.1}$$

Example 2.2.2

Find the mean of the following sample data: $2,\ -1,\ 0,\ 2$

Solution

This is a application of Equation 2.2.1:

$$\bar{x} = \frac{\sum x}{n} = \frac{2+(-1)+0+2}{4} = \frac{3}{4} = 0.75$$

Example 2.2.3

A random sample of ten students is taken from the student body of a college and their GPAs are recorded as follows:

$$1.90, 3.00, 2.53, 3.71, 2.12, 1.76, 2.71, 1.39, 4.00, 3.33$$

Find the mean.

Solution

This is a application of Equation 2.2.1:

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33}{10} = \frac{26.45}{10} = 2.645$$

Example 2.2.4

A random sample of 19 women beyond child-bearing age gave the following data, where x is the number of children and f is the frequency, or the number of times it occurred in the data set.

x	0	1	2	3	4
f	3	6	6	3	1

Find the sample mean.

Solution

In this example the data are presented by means of a data frequency table, introduced in Chapter 1. Each number in the first line of the table is a number that appears in the data set; the number below it is how many times it occurs. Thus the value 0 is observed three times, that is, three of the measurements in the data set are 0, the value 1 is observed six times, and so on. In the context of the problem this means that three women in the sample have had no children, six have had exactly one child, and so on. The explicit list of all the observations in this data set is therefore:

$$0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4$$

The sample size can be read directly from the table, without first listing the entire data set, as the sum of the frequencies: $n = 3 + 6 + 6 + 3 + 1 = 19$. The sample mean can be computed directly from the table as well:

$$\bar{x} = \frac{\sum x}{n} = \frac{0 \times 3 + 1 \times 6 + 2 \times 6 + 3 \times 3 + 4 \times 1}{19} = \frac{31}{19} = 1.6316$$

In the examples above the data sets were described as samples. Therefore the means were sample means $\bar{x}$. If the data come from a census, so that there is a measurement for every element of the population, then the mean is calculated by exactly the same process of summing all the measurements and dividing by how many of them there are, but it is now the ***population mean*** and is denoted by μ, the lower case Greek letter mu.

Definition: Population Mean

The *population mean* of a set of N population data is the number μ defined by the formula:

$$\mu = \frac{\sum x}{N}.$$

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is $(5 + 17) / 2 = 11$, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the "center" of the data set $\{5, 17\}$. For larger data sets the mean can similarly be regarded as the "center" of the data.

The Median

To see why another concept of average is needed, consider the following situation. Suppose we are interested in the average yearly income of employees at a large corporation. We take a random sample of seven employees, obtaining the sample data (rounded to the nearest hundred dollars, and expressed in thousands of dollars).

$$24.8, 22.8, 24.6, 192.5, 25.2, 18.5, 23.7$$

The mean (rounded to one decimal place) is $\bar{x} = 47.4$, but the statement "the average income of employees at this corporation is $\$47,400$" is surely misleading. It is approximately twice what six of the seven employees in the sample make and is nowhere near what any of them makes. It is easy to see what went wrong: the presence of the one executive in the sample, whose salary is so large compared to everyone else's, caused the numerator in the formula for the sample mean to be far too large, pulling the mean far to the right of where we think that the average "ought" to be, namely around $\$24,000$ or $\$25,000$. The number 192.5 in our data set is called an outlier, a number that is far removed from most or all of the remaining measurements. Many times an outlier is the result of some sort of error, but not always, as is the case here. We would get a better measure of the "center" of the data if we were to arrange the data in numerical order:

$$18.5, 22.8, 23.7, 24.6, 24.8, 25.2, 192.5$$

then select the middle number in the list, in this case 24.6. The result is called the median of the data set, and has the property that roughly half of the measurements are larger than it is, and roughly half are smaller. In this sense it locates the center of the data. If there are an even number of measurements in the data set, then there will be two middle elements when all are lined up in order, so we take the mean of the middle two as the median. Thus we have the following definition.

Definition: Sample Median

The *sample median* $\tilde{x}$ of a set of sample data for which there are an odd number of measurements is the middle measurement when the data are arranged in numerical order.

The sample median of a set of sample data for which there are an even number of measurements, is the mean of the two middle measurements when the data are arranged in numerical order.

Definition: Population Median

The *population median* is defined in the same way as the sample median except for the entire population.

The median is a value that divides the observations in a data set so that 50% of the data are on its left and the other 50% on its right. In accordance with Figure $2.2.7$, therefore, in the curve that represents the distribution of the data, a vertical line drawn at the median divides the area in two, area 0.5 (50% of the total area 1) to the left and area 0.5 (50% of the total area 1) to the right, as shown in Figure $2.2.1$. In our income example the median, $\$24,600$, clearly gave a much better measure of the middle of the data set than did the mean $\$47,400$. This is typical for situations in which the distribution is skewed. (Skewness and symmetry of distributions are discussed at the end of this subsection.)

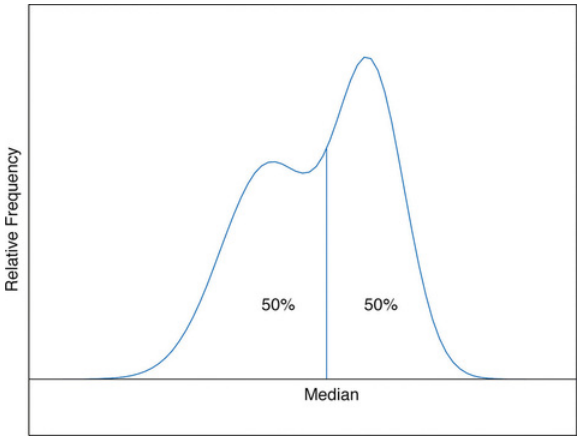


Figure 2.2.1: The Median

Example 2.2.5

Compute the sample median for the data from Example 2.2.2

Solution

The data in numerical order are $-1, 0, 2, 2$. The two middle measurements are 0 and 2, so $\tilde{x} = (0 + 2)/2 = 1$.

Example 2.2.6

Compute the sample median for the data from Example 2.2.3

Solution

The data in numerical order are

$$1.39, 1.76, 1.90, 2.12, 2.53, 2.71, 3.00, 3.33, 3.71, 4.00$$

The number of observations is ten, which is even, so there are two middle measurements, the fifth and sixth, which are 2.53 and 2.71. Therefore the median of these data is $\tilde{x} = (2.53 + 2.71)/2 = 2.62$.

Example 2.2.7

Compute the sample median for the data from Example 2.2.4

Solution

The data in numerical order are:

$$0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4$$

The number of observations is 19, which is odd, so there is one middle measurement, the tenth. Since the tenth measurement is 2, the median is $\tilde{x} = 2$.

In the last example it is important to note that we could have computed the median directly from the frequency table, without first explicitly listing all the observations in the data set. We already saw in Example 2.2.4 how to find the number of observations directly from the frequencies listed in the table $n = 3 + 6 + 6 + 3 + 1 = 19$. Thus the median is the tenth observation. The second line of the table in Example 2.2.4 shows that when the data are listed in order there will be three $0s$ followed by six $1s$, so the tenth observation, the median, is 2.

The relationship between the mean and the median for several common shapes of distributions is shown in Figure 2.2.2. The distributions in panels (a) and (b) are said to be *symmetric* because of the symmetry that they exhibit. The distributions in the remaining two panels are said to be *skewed*. In each distribution we have drawn a vertical line that divides the area under the curve in half, which in accordance with Figure 2.2.1 is located at the median. The following facts are true in general:

- When the distribution is symmetric, as in panels (a) and (b) of Figure 2.2.2, the mean and the median are equal.
- When the distribution is as shown in panel (c), it is said to be skewed right. The mean has been pulled to the right of the median by the long "right tail" of the distribution, the few relatively large data values.
- When the distribution is as shown in panel (d), it is said to be skewed left. The mean has been pulled to the left of the median by the long "left tail" of the distribution, the few relatively small data values.

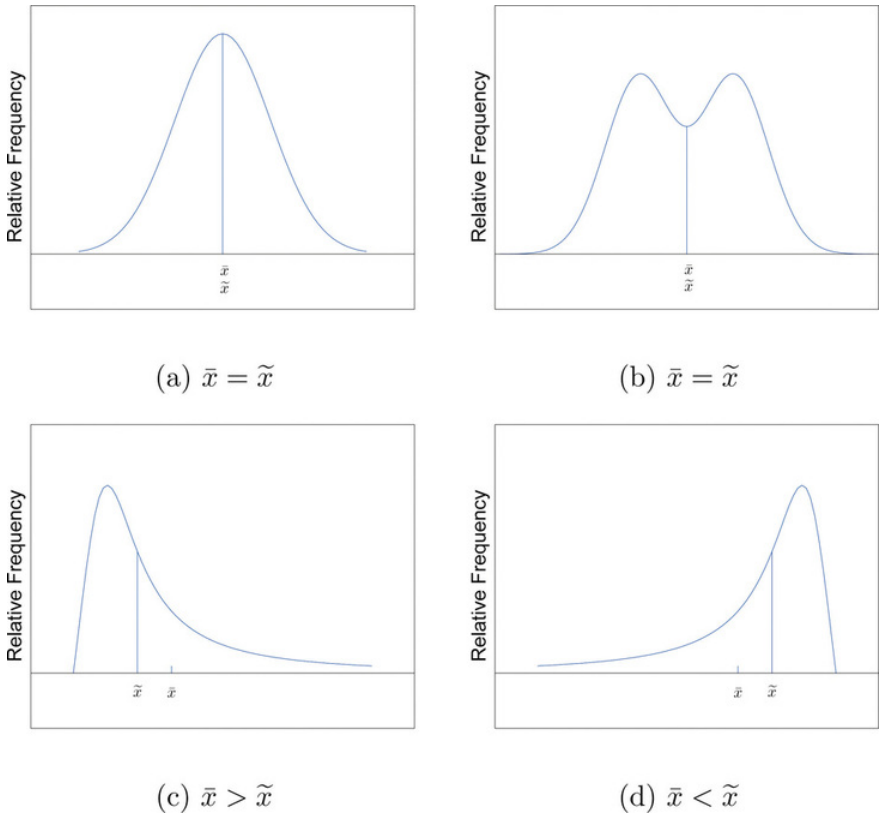


Figure 2.2.2: Skewness of Relative Frequency Histograms

The Mode

Perhaps you have heard a statement like "The average number of automobiles owned by households in the United States is 1.37," and have been amused at the thought of a fraction of an automobile sitting in a driveway. In such a context the following measure for central location might make more sense.

Definition: Sample Mode

The *sample mode* of a set of sample data is the most frequently occurring value.

On a relative frequency histogram, the highest point of the histogram corresponds to the mode of the data set. Figure 2.2.3 illustrates the mode.

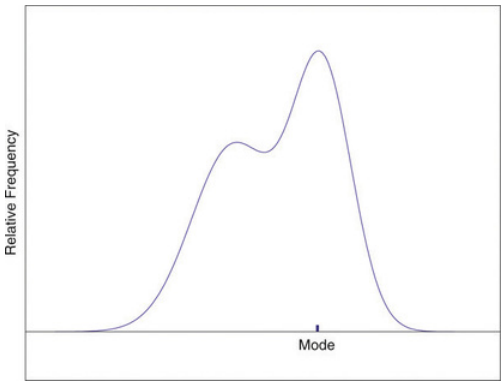


***Figure** 2.2.3**:** Mode*

For any data set there is always exactly one mean and exactly one median. This need not be true of the mode; several different values could occur with the highest frequency, as we will see. It could even happen that every value occurs with the same

frequency, in which case the concept of the mode does not make much sense.

Example 2.2.8

Find the mode of the following data set: $-1,\ 0,\ 2,\ 0$.

Solution

The value 0 is most frequently observed in the data set, so the mode is 0.

Example 2.2.9

Compute the sample mode for the data of Example 2.2.4

Solution

The two most frequently observed values in the data set are 1 and 2. Therefore mode is a set of two values: $\{1, 2\}$.

The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.

Key Takeaway

- The mean, the median, and the mode each answer the question "Where is the center of the data set?" The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

This page titled 2.2: Measures of Central Location - Three Kinds of Averages is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.3: Measures of Variability

Learning Objectives

- To learn the concept of the variability of a data set.
- To learn how to compute three measures of the variability of a data set: the range, the variance, and the standard deviation.

Look at the two data sets in Table 2.3.1 and the graphical representation of each, called a *dot plot*, in Figure 2.3.1.

Table 2.3.1: *Two Data Sets*

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

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

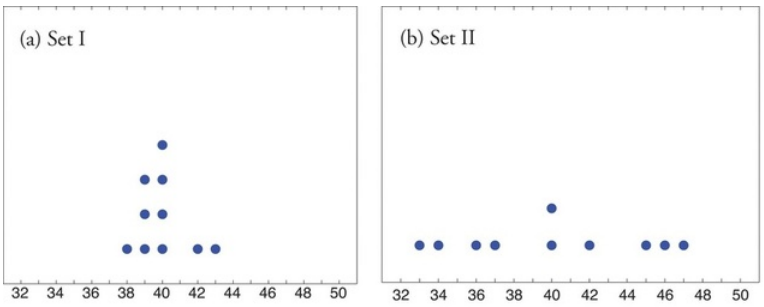


Figure 2.3.1: Dot Plots of Data Sets

The Range

First we discuss the simplest measure of variability.

Definition: range

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

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

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

Example 2.3.1: Identifyig the Range of a dataset

Find the range of each data set in Table 2.3.1.

Solution

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

The range is a measure of variability because it indicates the size of the interval over which the data points are distributed. A smaller range indicates less variability (less dispersion) among the data, whereas a larger range indicates the opposite.

The Variance and the Standard Deviation

The other two measures of variability that we will consider are more elaborate and also depend on whether the data set is just a sample drawn from a much larger population or is the whole population itself (that is, a census).

Definition: sample variance and sample Standard Deviation

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

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

which by algebra is equivalent to the formula

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

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

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

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

Example 2.3.2: Identifying the Variance and Standard Deviation of a Dataset

Find the sample variance and the sample standard deviation of Data Set II in Table 2.3.1.

Solution

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

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

Thus

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

so the variance is

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

and the standard deviation is

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

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

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

and

$$s = 20/9 \approx 1.49$$

Example 2.3.2

Find the sample variance and the sample standard deviation of the ten GPAs in "Example 2.2.3" in Section 2.2.

$$1.90 \;\; 3.00 \;\; 2.53 \;\; 3.71 \;\; 2.12 \;\; 1.76 \;\; 2.71 \;\; 1.39 \;\; 4.00 \;\; 3.33$$

Solution

Since

$$\sum x = 1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4.00 + 3.33 = 26.45$$

and

$$\sum x^2 = 1.902 + 3.002 + 2.532 + 3.712 + 2.122 + 1.762 + 2.712 + 1.392 + 4.002 + 3.332 = 76.7321$$

the shortcut formula gives

$$s^2 = \frac{\sum x^2 - (\sum x)^2}{n-1} = \frac{76.7321 - (26.45)^2/10}{10-1} = \frac{6.77185}{9} = 0.752427\overline{7}$$

and

$$s = \sqrt{0.752427\overline{7}} \approx 0.867$$

The sample variance has different units from the data. For example, if the units in the data set were inches, the new units would be inches squared, or square inches. It is thus primarily of theoretical importance and will not be considered further in this text, except in passing.

If the data set comprises the whole population, then the *population* standard deviation, denoted σ (the lower case Greek letter sigma), and its square, the *population* variance σ^2, are defined as follows.

Definitions: The population variance σ^2 and population standard deviation σ

The variability of a set of N population data is measured by the population variance

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \tag{2.3.1}$$

and its square root, the population standard deviation

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \tag{2.3.2}$$

where μ is the population mean as defined above.

Note that the denominator in the fraction is the full number of observations, not that number reduced by one, as is the case with the sample standard deviation. Since most data sets are samples, we will always work with the sample standard deviation and variance.

Finally, in many real-life situations the most important statistical issues have to do with comparing the means and standard deviations of two data sets. Figure 2.3.2 illustrates how a difference in one or both of the sample mean and the sample standard deviation are reflected in the appearance of the data set as shown by the curves derived from the relative frequency histograms built using the data.

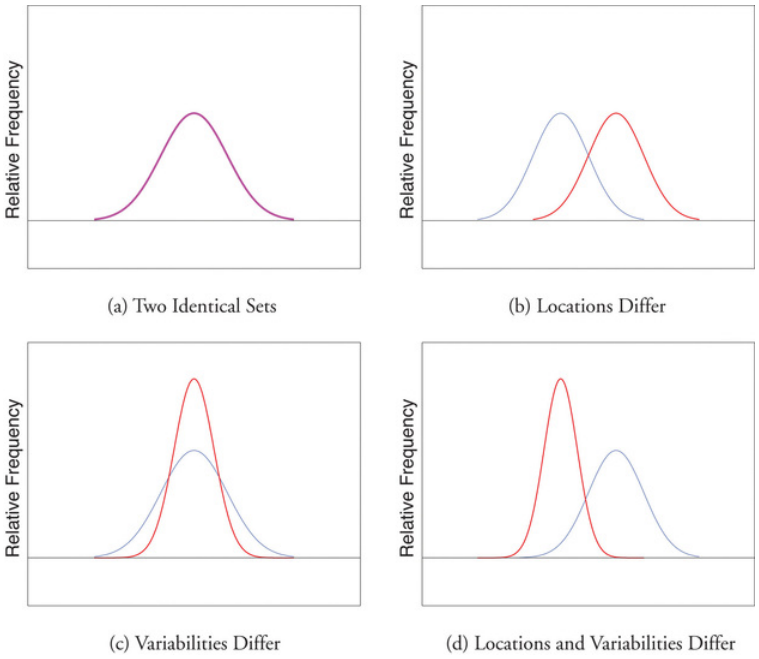


Figure 2.3.2: Difference between Two Data Sets

Key Takeaway

The range, the standard deviation, and the variance each give a quantitative answer to the question “How variable are the data?”

This page titled 2.3: Measures of Variability is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.4: Relative Position of Data

Learning Objectives

- To learn the concept of the relative position of an element of a data set.
- To learn the meaning of each of two measures, the percentile rank and the z-score, of the relative position of a measurement and how to compute each one.
- To learn the meaning of the three quartiles associated to a data set and how to compute them.
- To learn the meaning of the five-number summary of a data set, how to construct the box plot associated to it, and how to interpret the box plot.

When you take an exam, what is often as important as your actual score on the exam is the way your score compares to other students' performance. If you made a 70 but the average score (whether the mean, median, or mode) was 85, you did relatively poorly. If you made a 70 but the average score was only 55 then you did relatively well. In general, the significance of one observed value in a data set strongly depends on how that value compares to the other observed values in a data set. Therefore we wish to attach to each observed value a number that measures its relative position.

Percentiles and Quartiles

Anyone who has taken a national standardized test is familiar with the idea of being given both a score on the exam and a "percentile ranking" of that score. You may be told that your score was 625 and that it is the 85^{th} percentile. The first number tells how you actually did on the exam; the second says that 85% of the scores on the exam were less than or equal to your score, 625.

Definition: percentile of data

Given an observed value x in a data set, x is the P^{th} percentile of the data if $P\%$ of the data are less than or equal to x. The number P is the percentile rank of x.

Example 2.4.1

What percentile is the value 1.39 in the data set of ten GPAs considered in a previous Example? What percentile is the value 3.33?

Solution

The data, written in increasing order, are

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

The only data value that is less than or equal to 1.39 is 1.39 itself. Since 1 out of ten, or $1/10 = 10\%$, of the data points are less than or equal to 1.39, 1.39 is the 10^{th} percentile. Eight data values are less than or equal to 3.33. Since 8 out of ten, or $8 / 10 = .80 = 80\%$ of the data values are less than or equal to 3.33, the value 3.33 is the 80^{th} percentile of the data.

The Pth percentile cuts the data set in two so that approximately $P\%$ of the data lie below it and $(100 - P)\%$ of the data lie above it. In particular, the three percentiles that cut the data into fourths, as shown in Figure 2.4.1, are called the quartiles of a data set. The quartiles are the three numbers Q_1, Q_2, Q_3 that divide the data set approximately into fourths. The following simple computational definition of the three quartiles works well in practice.

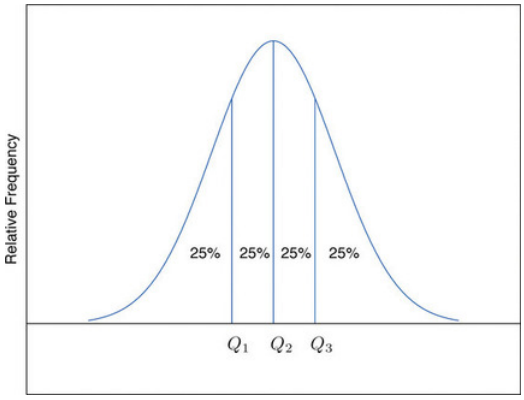


Figure 2.4.1: Data Division by Quartiles

Definition: quartile

For any data set:

1. The second quartile Q_2 of the data set is its median.
2. Define two subsets:
   1. the lower set: all observations that are strictly less than Q_2
   2. the upper set: all observations that are strictly greater than Q_2
3. The first quartile Q_1 of the data set is the median of the lower set.
4. The third quartile Q_3 of the data set is the median of the upper set.

Example 2.4.2

Find the quartiles of the data set of GPAs of discussed in a previous Example.

Solution

As in the previous example we first list the data in numerical order:

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 4.00$$

This data set has $n = 10$ observations. Since 10 is an even number, the median is the mean of the two middle observations:

$$\tilde{x} = (2.53 + 2.71) \,/\, 2 = 2.62.$$

Thus the second quartile is $Q_2 = 2.62$. The lower and upper subsets are

- Lower: $L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$,
- Upper: $U = \{2.71, 3.00, 3.33, 3.71, 4.00\}$.

Each has an odd number of elements, so the median of each is its middle observation. Thus the first quartile is $Q_1 = 1.90$, the median of L, and the third quartile is $Q_3 = 3.33$, the median of U.

Example 2.4.3

Adjoin the observation 3.88 to the data set of the previous example and find the quartiles of the new set of data.

Solution

As in the previous example we first list the data in numerical order:

$$1.39 \quad 1.76 \quad 1.90 \quad 2.12 \quad 2.53 \quad 2.71 \quad 3.00 \quad 3.33 \quad 3.71 \quad 3.88, 4.00$$

This data set has 11 observations. The second quartile is its median, the middle value 2.71. Thus $Q_2 = 2.71$. The lower and upper subsets are now

- Lower: $L = \{1.39, 1.76, 1.90, 2.12, 2.53\}$
- Upper: $U = \{3.00, 3.33, 3.71, 3.88, 4.00\}$.

The lower set L has median, the middle value, 1.90, so $Q_1 = 1.90$. The upper set has median 3.71, so $Q_3 = 3.71$.

In addition to the three quartiles, the two extreme values, the minimum x_{min} and the maximum x_{max} are also useful in describing the entire data set. Together these five numbers are called the five-number summary of a data set,

$$\{ X_{min}, Q_1, Q_2, Q_3, X_{max} \}$$

The five-number summary is used to construct a box plot, as in Figure 2.4.2. Each of the five numbers is represented by a vertical line segment, a box is formed using the line segments at Q_1 and Q_3 as its two vertical sides, and two horizontal line segments are extended from the vertical segments marking Q_1 and Q_3 to the adjacent extreme values. (The two horizontal line segments are referred to as “whiskers,” and the diagram is sometimes called a "box and whisker plot.") We caution the reader that there are other types of box plots that differ somewhat from the ones we are constructing, although all are based on the three quartiles.

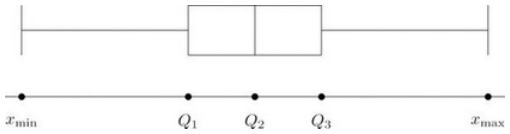


Figure 2.4.2: The Box Plot

Note that the distance from $Q1$ to $Q3$ is the length of the interval over which the middle half of the data range. Thus it has the following special name.

Definition: interquartile range

The interquartile range IQR is the quantity

$$IQR = Q3 - Q1$$

Example 2.4.4

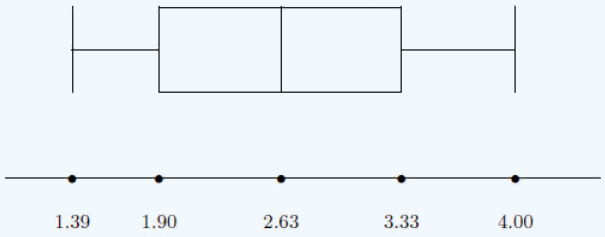
Construct a box plot and find the IQR for the data in Example 2.4.3.

Solution

From our work in Example 2.4.1, we know that the five-number summary is

- $X_{min} = 1.39$
- $Q1 = 1.90$
- $Q2 = 2.62$
- $Q3 = 3.33$
- $X_{max} = 4.00$

The box plot is:



The interquartile range is: $IQR = 3.33 - 1.90 = 1.43$.

z-Scores

Another way to locate a particular observation x in a data set is to compute its distance from the mean in units of standard deviation. The z-score indicates how many standard deviations an individual observation x is from the center of the data set, its mean. It is used on distributions that have been *standardized*, which allows us to better understand its properties. If z is negative then x is below average. If z is 0 then x is equal to the average. If z is positive then x is above the average

Definition: z-score

The z-score of an observation x is the number z given by the computational formula

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

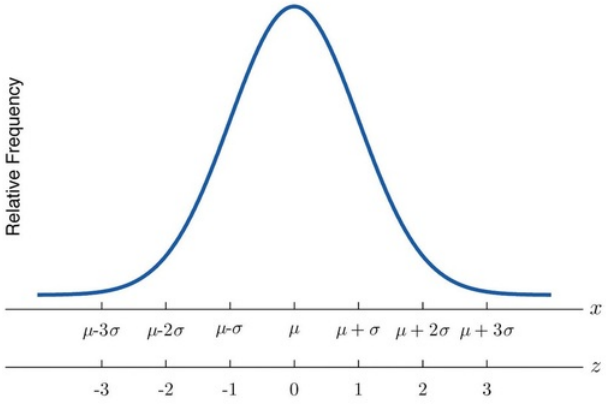


Figure 2.4.3***:*** *x*-Scale versus *z*-Score

Example 2.4.5

Suppose the mean and standard deviation of the GPA's of all currently registered students at a college are $\mu = 2.70$ and $\sigma = 0.50$. The z-scores of the GPA's of two students, Antonio and Beatrice, are $z = -0.62$ and $z = 1.28$, respectively. What are their GPAs?

Solution

Using the second formula right after the definition of z-scores we compute the GPA's as

- Antonio: $x = \mu + z\sigma = 2.70 + (-0.62)(0.50) = 2.39$
- Beatrice: $x = \mu + z\sigma = 2.70 + (1.28)(0.50) = 3.34$

Key Takeaways

- The percentile rank and z-score of a measurement indicate its relative position with regard to the other measurements in a data set.
- The three quartiles divide a data set into fourths.
- The five-number summary and its associated box plot summarize the location and distribution of the data.

This page titled 2.4: Relative Position of Data is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.5: The Empirical Rule and Chebyshev's Theorem

Learning Objectives

- To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the *Empirical Rule* and *Chebyshev's Theorem*.
- To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

You probably have a good intuitive grasp of what the average of a data set says about that data set. In this section we begin to learn what the standard deviation has to tell us about the nature of the data set.

The Empirical Rule

We start by examining a specific set of data. Table 2.5.1 shows the heights in inches of 100 randomly selected adult men. A relative frequency histogram for the data is shown in Figure 2.5.1. The mean and standard deviation of the data are, rounded to two decimal places, $\bar{x} = 69.92$ and $\sigma = 1.70$.

Table 2.5.1: Heights of Men

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $69.92 - 1.70 = 68.22$ and $69.92 + 1.70 = 71.62$ inches, there are 69 of them. If we count the number of observations that are within two standard deviations of the mean, that is, that are between $69.92 - 2(1.70) = 66.52$ and $69.92 + 2(1.70) = 73.32$ inches, there are 95 of them. All of the measurements are within three standard deviations of the mean, that is, between $69.92 - 3(1.70) = 64.822$ and $69.92 + 3(1.70) = 75.02$ inches. These tallies are not coincidences, but are in agreement with the following result that has been found to be widely applicable.

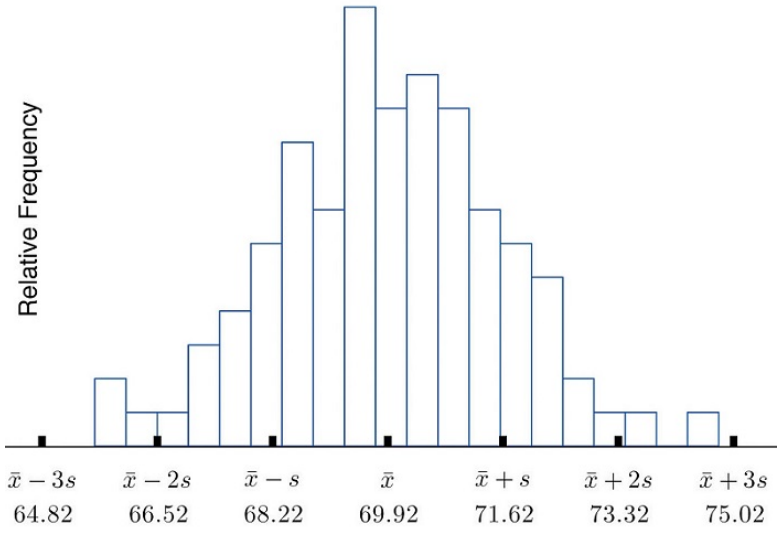


Figure 2.5.1: Heights of Adult Men

The Empirical Rule

Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations; if a data set has an approximately bell-shaped relative frequency histogram, then (Figure 2.5.2)

- approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations; and
- approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

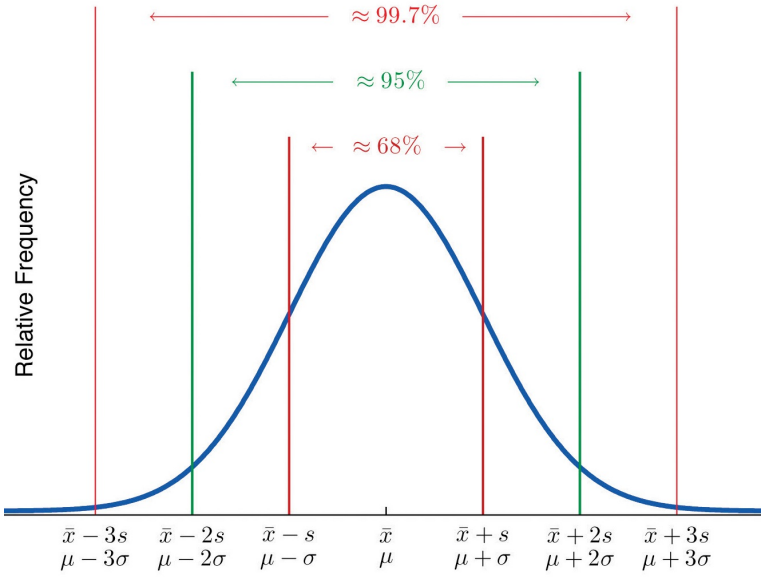


Figure 2.5.2: The Empirical Rule

Two key points in regard to the Empirical Rule are that the data distribution must be approximately bell-shaped and that the percentages are only approximately true. The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule. We see this with the example of the heights of the men: the Empirical Rule suggested 68 observations between 68.22 and 71.62 inches, but we counted 69.

Example 2.5.1

Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

1. About what proportion of all such men are between 68.2 and 71 inches tall?
2. What interval centered on the mean should contain about 95% of all such men?

Solution

A sketch of the distribution of heights is given in Figure 2.5.3.

1. Since the interval from 68.2 to 71.0 has endpoints $\bar{x} - s$ and $\bar{x} + s$, by the Empirical Rule about 68% of all 18-year-old males should have heights in this range.
2. By the Empirical Rule the shortest such interval has endpoints $\bar{x} - 2s$ and $\bar{x} + 2s$. Since

$$\bar{x} - 2s = 69.6 - 2(1.4) = 66.8$$

and

$$\bar{x} + 2s = 69.6 + 2(1.4) = 72.4$$

the interval in question is the interval from 66.8 inches to 72.4 inches.

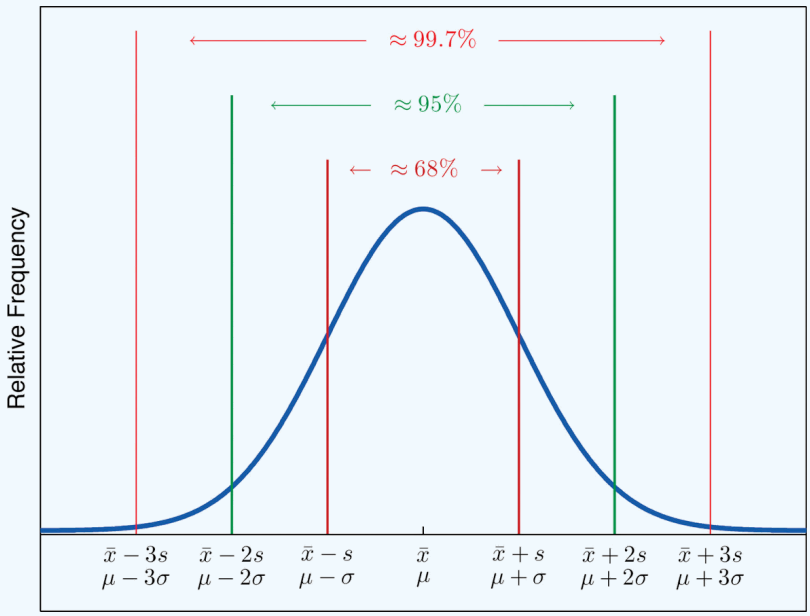


Figure 2.5.3: Distribution of Heights

Example 2.5.2

Scores on IQ tests have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution

A sketch of the IQ distribution is given in Figure 2.5.3. The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.

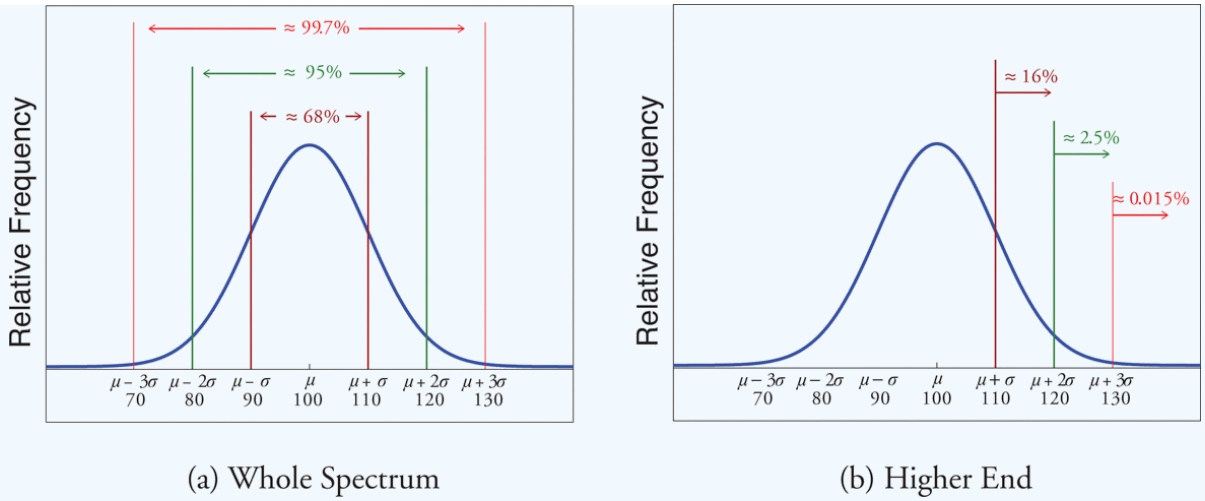


Figure 2.5.3: Distribution of IQ Scores.

1. Since 68% of the IQ scores lie *within* the interval from 90 to 110, it must be the case that 32% lie *outside* that interval. By symmetry approximately half of that 32%, or 16% of all IQ scores, will lie above 110. If 16% lie above 110, then 84% lie below. We conclude that the IQ score 110 is the 84^{th} percentile.
2. The same analysis applies to the score 120. Since approximately 95% of all IQ scores lie within the interval form 80 to 120, only 5% lie outside it, and half of them, or 2.5% of all scores, are above 120. The IQ score 120 is thus higher than 97.5% of all IQ scores, and is quite a high score.
3. By a similar argument, only $15/100$ of 1% of all adults, or about one or two in every thousand, would have an IQ score above 130. This fact makes the score 130 extremely high.

Chebyshev's Theorem

The Empirical Rule does not apply to all data sets, only to those that are bell-shaped, and even then is stated in terms of approximations. A result that applies to every data set is known as Chebyshev's Theorem.

Chebyshev's Theorem

For any numerical data set,

- at least $3/4$ of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations;
- at least $8/9$ of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations;
- at least $1 - 1/k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than 1.

Figure 2.5.4 gives a visual illustration of Chebyshev's Theorem.

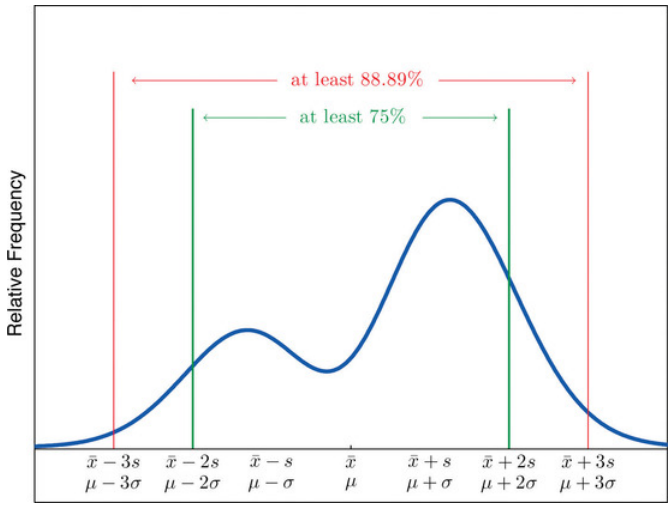


Figure 2.5.4: Chebyshev's Theorem

It is important to pay careful attention to the words **"at least"** at the beginning of each of the three parts of Chebyshev's Theorem. The theorem gives the *minimum* proportion of the data which must lie within a given number of standard deviations of the mean; the true proportions found within the indicated regions could be greater than what the theorem guarantees.

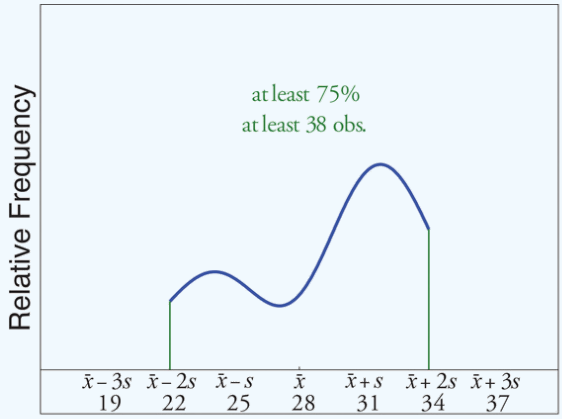
Example 2.5.3

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval $(22, 34)$? What can be said about the number of observations that lie outside that interval?

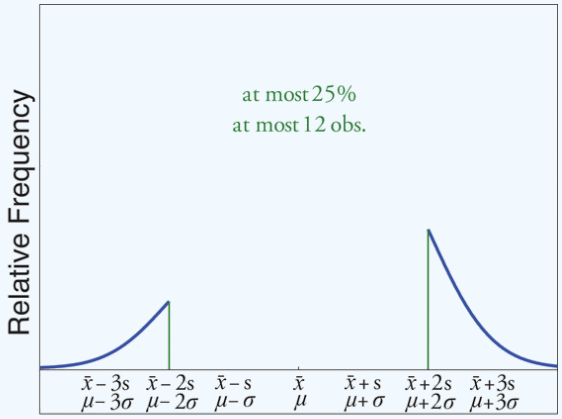
Solution

The interval $(22, 34)$ is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least $3/4$ of the data are within this interval. Since $3/4$ of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval $(22, 34)$.

If at least $3/4$ of the observations are in the interval, then at most $1/4$ of them are outside it. Since $1/4$ of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x $(22, 34)$.



(a) Within $\bar{x} \pm 2s$



(b) Outside $\bar{x} \pm 2$

Example 2.5.4

The number of vehicles passing through a busy intersection between $8:00\ a.m.$ and $10:00\ a.m.$ was observed and recorded on every weekday morning of the last year. The data set contains $n = 251$ numbers. The sample mean is $\bar{x} = 725$ and the sample standard deviation is $s = 25$. Identify which of the following statements *must* be true.

1. On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was between 675 and 775.
4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was either less than 675 or greater than 775.
5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.
6. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from $8:00\ a.m.$ to $10:00\ a.m.$ was less than 675.

Solution

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.
2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $\bar{x} - 2s, \bar{x} + 2s = (675, 775)$. It must be correct.
3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct.
4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.
5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval $(675, 775)$ are less than 75. Thus statement (5) might not be correct.
6. Statement (4) is definitely correct and statement (4) implies statement (6): even if every measurement that is outside the interval $(675, 775)$ is less than 675 (which is conceivable, since symmetry is not known to hold), even so at most 25% of all observations are less than 675. Thus statement (6) must definitely be correct.

Key Takeaway

- The Empirical Rule is an approximation that applies only to data sets with a bell-shaped relative frequency histogram. It estimates the proportion of the measurements that lie within one, two, and three standard deviations of the mean.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

This page titled 2.5: The Empirical Rule and Chebyshev's Theorem is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

2.E: Descriptive Statistics (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

2.1: Three popular data displays

Basic

1. Describe one difference between a frequency histogram and a relative frequency histogram.
2. Describe one advantage of a stem and leaf diagram over a frequency histogram.
3. Construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for the following data set. For the histograms use classes $51-60$, $61-70$, and so on.

Illegal pream-token (6)

4. Construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for the following data set. For the histograms use classes $6.0-6.9$, $7.0-7.9$, and so on.

Illegal pream-token (8)

5. A data set contains $n=10$ observations. The values x and their frequencies f are summarized in the following data frequency table.

x	-1	0	1	2
f	3	4	2	1

(2.E.1)

Construct a frequency histogram and a relative frequency histogram for the data set.

6. A data set contains the $n=20$ observations The values x and their frequencies f are summarized in the following data frequency table.

x	-1	0	1	2
f	3	a	2	1

(2.E.2)

The frequency of the value 0 is missing. Find a and then sketch a frequency histogram and a relative frequency histogram for the data set.

7. A data set has the following frequency distribution table:

x	1	2	3	4
f	3	a	2	1

(2.E.3)

The number a is unknown. Can you construct a frequency histogram? If so, construct it. If not, say why not.

8. A table of some of the relative frequencies computed from a data set is

x	1	2	3	4
f/n	0.3	p	0.2	0.1

(2.E.4)

The number p is yet to be computed. Finish the table and construct the relative frequency histogram for the data set.

Applications

9. The IQ scores of ten students randomly selected from an elementary school are given.

Illegal pream-token (1)

Grouping the measures in the $80s$, the $90s$, and so on, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.

10. The IQ scores of ten students randomly selected from an elementary school for academically gifted students are given.

Illegal pream-token (1)

Grouping the measures by their common hundreds and tens digits, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram.

11. During a one-day blood drive 300 people donated blood at a mobile donation center. The blood types of these 300 donors are summarized in the table.

Blood Type	*O*	*A*	*B*	*AB*
Frequency	136	120	32	12

(2.E.5)

Construct a relative frequency histogram for the data set.

12. In a particular kitchen appliance store an electric automatic rice cooker is a popular item. The weekly sales for the last 20weeks are shown.

Illegal pream-token (2)

Construct a relative frequency histogram with classes $6-10$, $11-15$, and $16-20$.

Additional Exercises

13. Random samples, each of size $n=10$, were taken of the lengths in centimeters of three kinds of commercial fish, with the following results:

Sample 1 :	108	100	99	125	87	105	107	105	119	118
Sample 2 :	133	140	152	142	137	145	160	138	139	138
Sample 3 :	82	60	83	82	82	74	79	82	80	80

(2.E.6)

Grouping the measures by their common hundreds and tens digits, construct a stem and leaf diagram, a frequency histogram, and a relative frequency histogram for each of the samples. Compare the histograms and describe any patterns they exhibit.

14. During a one-day blood drive 300 people donated blood at a mobile donation center. The blood types of these 300 donors are summarized below.

Blood Type	*O*	*A*	*B*	*AB*
Frequency	136	120	32	12

(2.E.7)

Identify the blood type that has the highest relative frequency for these 300 people. Can you conclude that the blood type you identified is also most common for all people in the population at large? Explain.

15. In a particular kitchen appliance store, the weekly sales of an electric automatic rice cooker for the last 20 weeks are as follows.

Illegal pream-token (2)

In retail sales, too large an inventory ties up capital, while too small an inventory costs lost sales and customer satisfaction. Using the relative frequency histogram for these data, find approximately how many rice cookers must be in stock at the beginning of each week if

a. the store is not to run out of stock by the end of a week for more than 15% of the weeks; and

b. the store is not to run out of stock by the end of a week for more than 5% of the weeks.

16. In retail sales, too large an inventory ties up capital, while too small an inventory costs lost sales and customer satisfaction. Using the relative frequency histogram for these data, find approximately how many rice cookers must be in stock at the beginning of each week if the store is not to run out of stock by the end of a week for more than 15% of the weeks; and the store is not to run out of stock by the end of a week for more than 5% of the weeks.

Answers

1. The vertical scale on one is the frequencies and on the other is the relative frequencies.
2.

5	3
6	8 9
7	0 0 0 5 6 7
8	0 2 3 5 5 5 8
9	2 3 6
10	0

(2.E.8)

3.
4.

5. Noting that $n = 10$ the relative frequency table is:

x	-1	0	1	2
f / n	0.3	0.4	0.2	0.1

(2.E.9)

6.
7. Since n is unknown, a is unknown, so the histogram cannot be constructed.
8.

8	7
9	9
10	0 5 5 7 8
11	8 9
12	5

(2.E.10)

9. Frequency and relative frequency histograms are similarly generated.
10.
11. Noting $n = 300$, the relative frequency table is therefore:

$Blood\,Type$	O	A	B	AB
f / n	0.4533	0.4	0.1067	0.04

(2.E.11)

A relative frequency histogram is then generated.

12.
13. The stem and leaf diagrams listed for Samples $1,\ 2,\ \text{and}\ 3$ in that order:

6	
7	
8	7
9	9
10	0 5 5 7 8
11	8 9
12	5
13	
14	
15	
16	

(2.E.12)

6	
7	
8	
9	
10	
11	
12	
13	3 7 8 8 9
14	0 2 5
15	2
16	0

(2.E.13)

6	0
7	4 9
8	0 0 2 2 2 2 3
9	
10	
11	
12	
13	
14	

(2.E.14)

15
16

The frequency tables are given below in the same order:

$Length$	$80 \sim 89$	$90 \sim 99$	$100 \sim 109$
f	1	1	5

(2.E.15)

$Length$	$110 \sim 119$	$120 \sim 129$
f	2	1

(2.E.16)

$Length$	$130 \sim 139$	$140 \sim 149$	$150 \sim 159$
f	5	3	1

(2.E.17)

$Length$	$160 \sim 169$
f	1

(2.E.18)

$Length$	$60 \sim 69$	$70 \sim 79$	$80 \sim 89$
f	1	2	7

(2.E.19)

The relative frequency tables are also given below in the same order:

$Length$	$80 \sim 89$	$90 \sim 99$	$100 \sim 109$
$f\ /\ n$	0.1	0.1	0.5

(2.E.20)

$Length$	$110 \sim 119$	$120 \sim 129$
$f\ /\ n$	0.2	0.1

(2.E.21)

$Length$	$130 \sim 139$	$140 \sim 149$	$150 \sim 159$
$f\ /\ n$	0.5	0.3	0.1

(2.E.22)

$Length$	$160 \sim 169$
$f\ /\ n$	0.1

(2.E.23)

$Length$	$60 \sim 69$	$70 \sim 79$	$80 \sim 89$
$f\ /\ n$	0.1	0.2	0.7

(2.E.24)

14.

15. a. 19
    b. 20

16.

2.2: Measures of Central Location

Basic

1. For the sample data set $\{1, 2, 6\}$ find
    1. $\sum x$
    2. $\sum x^2$
    3. $\sum (x - 3)$
    4. $\sum (x - 3)^2$
2. For the sample data set $\{-1, 0, 1, 4\}$ find
    1. $\sum x$
    2. $\sum x^2$
    3. $\sum (x - 1)$
    4. $\sum (x - 1)^2$
3. Find the mean, the median, and the mode for the sample

1 2 3 4

(2.E.25)

https://stats.libretexts.org/@go/page/1327

4. Find the mean, the median, and the mode for the sample

$$3\ 3\ 4\ 4 \tag{2.E.26}$$

5. Find the mean, the median, and the mode for the sample

$$2\ 1\ 2\ 7 \tag{2.E.27}$$

6. Find the mean, the median, and the mode for the sample

$$-1\ 0\ 1\ 4\ 1\ 1 \tag{2.E.28}$$

7. Find the mean, the median, and the mode for the sample data represented by the table

x	1	2	7
f	1	2	1

(2.E.29)

8. Find the mean, the median, and the mode for the sample data represented by the table

x	-1	0	1	4
f	1	1	3	1

(2.E.30)

9. Create a sample data set of size $n = 3$ for which the mean $\bar{x}$ is greater than the median $\tilde{x}$.
10. Create a sample data set of size $n = 3$ for which the mean $\bar{x}$ is less than the median $\tilde{x}$.
11. Create a sample data set of size $n = 4$ for which the mean $\bar{x}$, the median $\tilde{x}$, and the mode are all identical.
12. Create a sample data set of size $n = 4$ for which the median $\tilde{x}$ and the mode are identical but the mean $\bar{x}$ is different.

Applications

13. Find the mean and the median for the LDL cholesterol level in a sample of ten heart patients.

$$\begin{array}{ccccc} 132 & 162 & 133 & 145 & 148 \\ 139 & 147 & 160 & 150 & 153 \end{array} \tag{2.E.31}$$

14. Find the mean and the median, for the LDL cholesterol level in a sample of ten heart patients on a special diet.

$$\begin{array}{ccccc} 127 & 152 & 138 & 110 & 152 \\ 113 & 131 & 148 & 135 & 158 \end{array} \tag{2.E.32}$$

15. Find the mean, the median, and the mode for the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

(2.E.33)

16. The number of passengers in each of 120 randomly observed vehicles during morning rush hour was recorded, with the following results.

x	1	2	3	4	5
f	84	29	3	3	1

(2.E.34)

Find the mean, the median, and the mode of this data set.

17. Twenty-five $1 - lb$ boxes of $16d$ nails were randomly selected and the number of nails in each box was counted, with the following results.

x	47	48	49	50	51
f	1	3	18	2	1

(2.E.35)

Find the mean, the median, and the mode of this data set.

Additional Exercises

18. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 500 days, four mice have died but the fifth one survives. The recorded survival times for the five mice are

$$493 \quad 421 \quad 222 \quad 378 \quad 500^{*} \tag{2.E.36}$$

where 500^* indicates that the fifth mouse survived for at least 500 days but the survival time (i.e., the exact value of the observation) is unknown.

a. Can you find the sample mean for the data set? If so, find it. If not, why not?
b. Can you find the sample median for the data set? If so, find it. If not, why not?

19. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 450 days, three mice have died, and one of the remaining mice is sacrificed for analysis. By the end of the observational period, the last remaining mouse still survives. The recorded survival times for the five mice are

$$222 \quad 421 \quad 378 \quad 450^* \quad 500^* \tag{2.E.37}$$

where * indicates that the mouse survived for at least the given number of days but the exact value of the observation is unknown.

a. Can you find the sample mean for the data set? If so, find it. If not, explain why not.
b. Can you find the sample median for the data set? If so, find it. If not, explain why not.

20. A player keeps track of all the rolls of a pair of dice when playing a board game and obtains the following data.

x	2	3	4	5	6	7
f	10	29	40	56	68	77

(2.E.38)

x	8	9	10	11	12
f	67	55	39	28	11

(2.E.39)

Find the mean, the median, and the mode.

21. Cordelia records her daily commute time to work each day, to the nearest minute, for two months, and obtains the following data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

(2.E.40)

a. Based on the frequencies, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

22. An ordered stem and leaf diagram gives the scores of 71 students on an exam.

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

(2.E.41)

a. Based on the shape of the display, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

23. A man tosses a coin repeatedly until it lands heads and records the number of tosses required. (For example, if it lands heads on the first toss he records a 1; if it lands tails on the first two tosses and heads on the third he records a 3.) The data are shown.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

(2.E.42)

a. Find the mean of the data.
b. Find the median of the data.

24. a. Construct a data set consisting of ten numbers, all but one of which is above average, where the average is the mean.
b. Is it possible to construct a data set as in part (a) when the average is the median? Explain.

25. Show that no matter what kind of average is used (mean, median, or mode) it is impossible for all members of a data set to be above average.
26. a. Twenty sacks of grain weigh a total of $1,003\ lb$. What is the mean weight per sack?
    b. Can the median weight per sack be calculated based on the information given? If not, construct two data sets with the same total but different medians.
27. Begin with the following set of data, call it Data Set I.

$$5 \quad -2 \quad 6 \quad 14 \quad -3 \quad 0 \quad 1 \quad 4 \quad 3 \quad 2 \quad 5 \tag{2.E.43}$$

    a. Compute the mean, median, and mode.
    b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the mean, median, and mode of Data Set II.
    c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the mean, median, and mode of Data Set III.
    d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

Large Data Set Exercises

Note: For Large Data Set Exercises below, all of the data sets associated with these questions are missing, but the questions themselves are included here for reference.

28. Large Data Set 1 lists the SAT scores and GPAs of $1,000$ students.
    a. Compute the mean and median of the $1,000$ SAT scores.
    b. Compute the mean and median of the $1,000$ GPAs.
29. Large Data Set 1 lists the SAT scores of $1,000$ students.
    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.
    c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.
30. Large Data Set 1 lists the GPAs of $1,000$ students.
    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
    b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.
    c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample mean $\bar{x}$ and compare it to μ.
31. Large Data Sets 7, $7A$, and $7B$ list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.
    a. Compute the mean and median survival time for all mice, without regard to gender.
    b. Compute the mean and median survival time for the 65 male mice (separately recorded in Large Data Set 7A).
    c. Compute the mean and median survival time for the 75 female mice (separately recorded in Large Data Set 7B).

Answers

1. a. 9
   b. 41
   c. 0
   d. 14
2.
3. $\bar{x} = 2.5,\ \tilde{x} = 2.5,\ \text{mode} = \{1, 2, 3, 4\}$
4.
5. $\bar{x} = 3,\ \tilde{x} = 2,\ \text{mode} = 2$
6.

7. $\bar{x} = 3,\ \tilde{x} = 2,\ \text{mode} = 2$
8.
9. $\{0, 0, 3\}$
10.
11. $\{0, 1, 1, 2\}$
12.
13. $\bar{x} = 146.9,\ \tilde{x} = 147.5$
14.
15. $\bar{x} = 2.6,\ \tilde{x} = 2,\ \text{mode} = 2$
16.
17. $\bar{x} = 48.96,\ \tilde{x} = 49,\ \text{mode} = 49$
18.
19. a. No, the survival times of the fourth and fifth mice are unknown.
    b. Yes, $\bar{x} = 421$.
20.
21. $\bar{x} = 28.55,\ \tilde{x} = 28,\ \text{mode} = 28$
22.
23. $\bar{x} = 2.05,\ \tilde{x} = 2,\ \text{mode} = 1$
24.
25. **Mean**: $nx_{min} \leq \sum x$ so dividing by n yields $x_{min} \leq \bar{x}$, so the minimum value is not above average. **Median**: the middle measurement, or average of the two middle measurements, $\tilde{x}$, is at least as large as x_{min}, so the minimum value is not above average. **Mode**: the mode is one of the measurements, and is not greater than itself
26.
27. a. $\bar{x} = 3.18,\ \tilde{x} = 3,\ \text{mode} = 5$
    b. $\bar{x} = 6.18,\ \tilde{x} = 6,\ \text{mode} = 8$
    c. $\bar{x} = -2.81,\ \tilde{x} = -3,\ \text{mode} = -1$
    d. If a number is added to every measurement in a data set, then the mean, median, and mode all change by that number.
28.
29. a. $\mu = 1528.74$
    b. $\bar{x} = 1502.8$
    c. $\bar{x} = 1532.2$
30.
31. a. $\bar{x} = 553.4286,\ \tilde{x} = 552.5$
    b. $\bar{x} = 665.9692,\ \tilde{x} = 667$
    c. $\bar{x} = 455.8933,\ \tilde{x} = 448$

2.3 Measures of Variability

Basic

1. Find the range, the variance, and the standard deviation for the following sample.

$$1\ 2\ 3\ 4 \tag{2.E.44}$$

2. Find the range, the variance, and the standard deviation for the following sample.

$$2\ -3\ 6\ 0\ 3\ 1 \tag{2.E.45}$$

3. Find the range, the variance, and the standard deviation for the following sample.

$$2\ 1\ 2\ 7 \tag{2.E.46}$$

4. Find the range, the variance, and the standard deviation for the following sample.

$$-1\ 0\ 1\ 4\ 1\ 1 \tag{2.E.47}$$

5. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	1	2	7
f	1	2	1

(2.E.48)

6. Find the range, the variance, and the standard deviation for the sample represented by the data frequency table.

x	-1	0	1	4
f	1	1	3	1

(2.E.49)

Applications

7. Find the range, the variance, and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

$$\begin{array}{ccccc} 132 & 162 & 133 & 145 & 148 \\ 139 & 147 & 160 & 150 & 153 \end{array} \tag{2.E.50}$$

8. Find the range, the variance and the standard deviation for the sample of ten IQ scores randomly selected from a school for academically gifted students.

$$\begin{array}{ccccc} 142 & 152 & 138 & 145 & 148 \\ 139 & 147 & 155 & 150 & 153 \end{array} \tag{2.E.51}$$

Additional Exercises

9. Consider the data set represented by the table

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

(2.E.52)

   a. Use the frequency table to find that $\sum x = 1256$ and $\sum x^2 = 35,926$.
   b. Use the information in part (a) to compute the sample mean and the sample standard deviation.

10. Find the sample standard deviation for the data

x	1	2	3	4	5
f	384	208	98	56	28

(2.E.53)

x	6	7	8	9	10
f	12	8	2	3	1

(2.E.54)

11. A random sample of 49 invoices for repairs at an automotive body shop is taken. The data are arrayed in the stem and leaf diagram shown. (Stems are thousands of dollars, leaves are hundreds, so that for example the largest observation is $3,800$.)

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

(2.E.55)

   For these data, $\sum x = 101$, $\sum x^2 = 244,830,000$.

   a. Compute the mean, median, and mode.
   b. Compute the range.
   c. Compute the sample standard deviation.

12. What must be true of a data set if its standard deviation is 0?

13. A data set consisting of 25 measurements has standard deviation 0. One of the measurements has value 17. What are the other 24 measurements?

14. Create a sample data set of size $n = 3$ for which the range is 0 and the sample mean is 2.

15. Create a sample data set of size $n = 3$ for which the sample variance is 0 and the sample mean is 1.
16. The sample $\{-1, 0, 1\}$ has mean $\bar{x} = 0$ and standard deviation $\bar{x} = 0$. Create a sample data set of size $n = 3$ for which $\bar{x} = 0$ and s is greater than 1.
17. The sample $\{-1, 0, 1\}$ has mean $\bar{x} = 0$ and standard deviation $\bar{x} = 0$. Create a sample data set of size $n = 3$ for which $\bar{x} = 0$ and the standard deviation s is less than 1.
18. Begin with the following set of data, call it Data Set I.

$$5 \quad -2 \; 6 \; 1 \; 4 \quad -3 \; 0 \; 1 \; 4 \; 3 \; 2 \; 5 \tag{2.E.56}$$

    a. Compute the sample standard deviation of Data Set I.
    b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the sample standard deviation of Data Set II.
    c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the sample standard deviation of Data Set III.
    d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

Large Data Set Exercises

Note: For Large Data Set Exercises below, all of the data sets associated with these questions are missing, but the questions themselves are included here for reference.

19. Large Data Set 1 lists the SAT scores and GPAs of $1,000$ students.

    a. Compute the range and sample standard deviation of the $1,000$ SAT scores.
    b. Compute the range and sample standard deviation of the $1,000$ GPAs.

20. Large Data Set 1 lists the SAT scores of $1,000$ students.

    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population range and population standard deviation σ.
    b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.
    c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

21. Large Data Set 1 lists the GPAs of $1,000$ students.

    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population range and population standard deviation σ.
    b. Regard the first 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.
    c. Regard the next 25 observations as a random sample drawn from this population. Compute the sample range and sample standard deviation s and compare them to the population range and σ.

22. Large Data Set 7, 7A, and 7B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.

    a. Compute the range and sample standard deviation of survival time for all mice, without regard to gender.
    b. Compute the range and sample standard deviation of survival time for the 65 male mice (separately recorded in Large Data Set 7A).
    c. Compute the range and sample standard deviation of survival time for the 75 female mice (separately recorded in Large Data Set 7B). Do you see a difference in the results for male and female mice? Does it appear to be significant?

Answers

1. $R = 3,\ s^2 = 1.7,\ s = 1.3.$
2.
3. $R = 6,\ s^2 = 7.\overline{3},\ s = 2.7.$
4.

5. $R = 6,\ s^2 = 7.3,\ s = 2.7.$
6.
7. $R = 30,\ s^2 = 103.2,\ s = 10.2.$
8.
9. $\bar{x} = 28.55,\ s = 1.3.$
10.
11. a. $\bar{x} = 2063,\ \tilde{x} = 2000,\ \text{mode} = 2000.$
    b. $R = 3400.$
    c. $s = 869.$
12.
13. All are 17.
14.
15. $\{1, 1, 1\}$
16.
17. One example is $\{-.5, 0, .5\}$.
18.
19. a. $R = 1350$ and $s = 212.5455$
    b. $R = 4.00$ and $s = 0.7407$
20.
21. a. $R = 4.00$ and $\sigma = 0.740375$
    b. $R = 3.04$ and $s = 0.808045$
    c. $R = 2.49$ and $s = 0.657843$

2.4 Relative Position of Data

Basic

1. Consider the data set

$$\begin{array}{ccccc} 69 & 92 & 68 & 77 & 80 \\ 93 & 75 & 76 & 82 & 100 \\ 70 & 85 & 88 & 85 & 96 \\ 53 & 70 & 70 & 82 & 85 \end{array} \tag{2.E.57}$$

   a. Find the percentile rank of 82.
   b. Find the percentile rank of 68.

2. Consider the data set

$$\begin{array}{ccccc} 8.5 & 8.2 & 7.0 & 7.0 & 4.9 \\ 9.6 & 8.5 & 8.8 & 8.5 & 8.7 \\ 6.5 & 8.2 & 7.6 & 1.5 & 9.3 \\ 8.0 & 7.7 & 2.9 & 9.2 & 6.9 \end{array} \tag{2.E.58}$$

   a. Find the percentile rank of 6.5.
   b. Find the percentile rank of 7.7.

3. Consider the data set represented by the ordered stem and leaf diagram

$$\begin{array}{c|cccccccccccccccccc} 10 & 0 & 0 \\ 9 & 1 & 1 & 1 & 1 & 2 & 3 \\ 8 & 0 & 1 & 1 & 2 & 2 & 3 & 4 & 5 & 7 & 8 & 8 & 9 \\ 7 & 0 & 0 & 0 & 1 & 1 & 2 & 4 & 4 & 5 & 6 & 6 & 6 & 7 & 7 & 7 & 8 & 8 & 9 \\ 6 & 0 & 1 & 2 & 2 & 2 & 3 & 4 & 4 & 5 & 7 & 7 & 7 & 7 & 8 & 8 \\ 5 & 0 & 2 & 3 & 3 & 4 & 4 & 6 & 7 & 7 & 8 & 9 \\ 4 & 2 & 5 & 6 & 8 & 8 \\ 3 & 9 & 9 \end{array} \tag{2.E.59}$$

a. Find the percentile rank of the grade 75.
b. Find the percentile rank of the grade 57.

4. Is the 90^{th} percentile of a data set always equal to 90%? Why or why not?
5. The 29^{th} percentile in a large data set is 5.
   a. Approximately what percentage of the observations are less than 5?
   b. Approximately what percentage of the observations are greater than 5?
6. The 54^{th} percentile in a large data set is 98.6.
   a. Approximately what percentage of the observations are less than 98.6?
   b. Approximately what percentage of the observations are greater than 98.6?
7. In a large data set the 29^{th} percentile is 5 and the 79^{th} percentile is 10. Approximately what percentage of observations lie between 5 and 10?
8. In a large data set the 40^{th} percentile is 125 and the 82^{nd} percentile is 158. Approximately what percentage of observations lie between 125 and 158?
9. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the stem and leaf diagram in Figure 2.1.2 "Ordered Stem and Leaf Diagram".
10. Find the five-number summary and the IQR and sketch the box plot for the sample explicitly displayed in "Example 2.2.7" in Section 2.2.
11. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the data frequency table

$$\begin{array}{c|ccccc} x & 1 & 2 & 5 & 8 & 9 \\ \hline f & 5 & 2 & 3 & 6 & 4 \end{array} \tag{2.E.60}$$

12. Find the five-number summary and the IQR and sketch the box plot for the sample represented by the data frequency table

$$\begin{array}{c|cccccccccc} x & -5 & -3 & -2 & -1 & 0 & 1 & 3 & 4 & 5 \\ \hline f & 2 & 1 & 3 & 2 & 4 & 1 & 1 & 2 & 1 \end{array} \tag{2.E.61}$$

13. Find the z-score of each measurement in the following sample data set.

$$-5 \;\; 6 \;\; 2 \;\; -1 \;\; 0 \tag{2.E.62}$$

14. Find the z-score of each measurement in the following sample data set.

$$1.6 \;\; 5.2 \;\; 2.8 \;\; 3.7 \;\; 4.0 \tag{2.E.63}$$

15. The sample with data frequency table

$$\begin{array}{c|ccc} x & 1 & 2 & 7 \\ \hline f & 1 & 2 & 1 \end{array} \tag{2.E.64}$$

has mean $\bar{x} = 3$ and standard deviation $s \approx 2.71$. Find the z-score for every value in the sample.

16. The sample with data frequency table

$$\begin{array}{c|cccc} x & -1 & 0 & 1 & 4 \\ \hline f & 1 & 1 & 3 & 1 \end{array} \tag{2.E.65}$$

has mean $\bar{x} = 1$ and standard deviation $s \approx 1.67$. Find the z-score for every value in the sample.

17. For the population

$$0 \;\; 0 \;\; 2 \;\; 2 \tag{2.E.66}$$

compute each of the following.
   a. The population mean μ.
   b. The population variance σ^2.
   c. The population standard deviation σ.
   d. The z-score for every value in the population data set.

18. For the population

$$0.5 \;\; 2.1 \;\; 4.4 \;\; 1.0 \tag{2.E.67}$$

compute each of the following.

a. The population mean μ.
b. The population variance σ^2.
c. The population standard deviation σ.
d. The z-score for every value in the population data set.

19. A measurement x in a sample with mean $\bar{x} = 10$ and standard deviation $s = 3$ has z-score $z = 2$. Find x.
20. A measurement x in a sample with mean $\bar{x} = 10$ and standard deviation $s = 3$ has z-score $z = -1$. Find x.
21. A measurement x in a population with mean $\mu = 2.3$ and standard deviation $\sigma = 1.3$ has z-score $z = 2$. Find x.
22. A measurement x in a sample with mean $\mu = 2.3$ and standard deviation $\sigma = 1.3$ has z-score $z = -1.2$. Find x.

Applications

23. The weekly sales for the last 20 weeks in a kitchen appliance store for an electric automatic rice cooker are

$$\begin{array}{ccccc} 20 & 15 & 14 & 14 & 18 \\ 15 & 19 & 12 & 13 & 9 \\ 15 & 17 & 16 & 16 & 18 \\ 19 & 15 & 15 & 16 & 15 \end{array} \tag{2.E.68}$$

a. Find the percentile rank of 15.
b. If the sample accurately reflects the population, then what percentage of weeks would an inventory of 15 rice cookers be adequate?

24. The table shows the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

(2.E.69)

a. Find the percentile rank of 2.
b. If the sample accurately reflects the population, then what percentage of households have at most two vehicles?

25. For two months Cordelia records her daily commute time to work each day to the nearest minute and obtains the following data:

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

(2.E.70)

Cordelia is supposed to be at work at $8:00\ a.m.$ but refuses to leave her house before $7:30\ a.m.$.

a. Find the percentile rank of 30, the time she has to get to work.
b. Assuming that the sample accurately reflects the population of *all* of Cordelia's commute times, use your answer to part (a) to predict the proportion of the work days she is late for work.

26. The mean score on a standardized grammar exam is 49.6; the standard deviation is 1.35. Dromio is told that the z-score of his exam score is -1.19.

a. Is Dromio's score above average or below average?
b. What was Dromio's actual score on the exam?

27. A random sample of 49 invoices for repairs at an automotive body shop is taken. The data are arrayed in the stem and leaf diagram shown. (Stems are thousands of dollars, leaves are hundreds, so that for example the largest observation is $3,800$.)

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

(2.E.71)

For these data, $\sum x = 101,100$, $\sum x^2 = 244,830,000$.

a. Find the z-score of the repair that cost $\$1,100$.

b. Find the z-score of the repairs that cost $\$2,700$.

28. The stem and leaf diagram shows the time in seconds that callers to a telephone-order center were on hold before their call was taken.

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

(2.E.72)

a. Find the quartiles.
b. Give the five-number summary of the data.
c. Find the range and the IQR.

Additional Exercises

29. Consider the data set represented by the ordered stem and leaf diagram

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

(2.E.73)

a. Find the three quartiles.
b. Give the five-number summary of the data.
c. Find the range and the IQR.

30. For the following stem and leaf diagram the units on the stems are thousands and the units on the leaves are hundreds, so that for example the largest observation is $3,800$.

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

(2.E.74)

a. Find the percentile rank of 800.
b. Find the percentile rank of $3,200$.

31. Find the five-number summary for the following sample data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

(2.E.75)

32. Find the five-number summary for the following sample data.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

(2.E.76)

33. For the following stem and leaf diagram the units on the stems are thousands and the units on the leaves are hundreds, so that for example the largest observation is $3,800$.

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

(2.E.77)

a. Find the three quartiles.
b. Find the IQR.
c. Give the five-number summary of the data.

34. Determine whether the following statement is true. "In any data set, if an observation x_1 is greater than another observation x_2, then the z-score of x_1 is greater than the z-score of x_2."
35. Emilia and Ferdinand took the same freshman chemistry course, Emilia in the fall, Ferdinand in the spring. Emilia made an 83 on the common final exam that she took, on which the mean was 76 and the standard deviation 8. Ferdinand made a 79 on the common final exam that he took, which was more difficult, since the mean was 65 and the standard deviation 12. The one who has a higher z-score did relatively better. Was it Emilia or Ferdinand?
36. Refer to the previous exercise. On the final exam in the same course the following semester, the mean is 68 and the standard deviation is 9. What grade on the exam matches Emilia's performance? Ferdinand's?
37. Rosencrantz and Guildenstern are on a weight-reducing diet. Rosencrantz, who weighs 178 lb, belongs to an age and body-type group for which the mean weight is 145 lb and the standard deviation is 15 lb. Guildenstern, who weighs 204 lb, belongs to an age and body-type group for which the mean weight is 165 lb and the standard deviation is 20 lb. Assuming z-scores are good measures for comparison in this context, who is more overweight for his age and body type?

Large Data Set Exercises

Note: For Large Data Set Exercises below, all of the data sets associated with these questions are missing, but the questions themselves are included here for reference.

38. Large Data Set 1 lists the SAT scores and GPAs of $1,000$ students.
    a. Compute the three quartiles and the interquartile range of the $1,000$ SAT scores.
    b. Compute the three quartiles and the interquartile range of the $1,000$ GPAs.
39. Large Data Set 10 records the scores of 72 students on a statistics exam.
    a. Compute the five-number summary of the data.
    b. Describe in words the performance of the class on the exam in the light of the result in part (a).
40. Large Data Sets 3 and 3A list the heights of 174 customers entering a shoe store.
    a. Compute the five-number summary of the heights, without regard to gender.
    b. Compute the five-number summary of the heights of the men in the sample.
    c. Compute the five-number summary of the heights of the women in the sample.
41. Large Data Sets 7, 7A, and 3B list the survival times in days of 140 laboratory mice with thymic leukemia from onset to death.
    a. Compute the three quartiles and the interquartile range of the survival times for all mice, without regard to gender.
    b. Compute the three quartiles and the interquartile range of the survival times for the 65 male mice (separately recorded in Data Set 7A).
    c. Compute the three quartiles and the interquartile range of the survival times for the 75 female mice (separately recorded in Data Sets 7B).

Answer

1. a. 60
   b. 10
2.
3. a. 59
   b. 23

4.

5. a. 29
   b. 71

6.

7. 50%

8.

9. $x_{min} = 25,\ \ Q_1 = 70,\ \ Q_2 = 77.5\ \ Q_3 = 90,\ \ x_{max} = 100,\ \ IQR = 20$

10.

11. $x_{min} = 1,\ \ Q_1 = 1.5,\ \ Q_2 = 6.5\ \ Q_3 = 8,\ \ x_{max} = 9,\ \ IQR = 6.5$

12.

13. $-1.3,\ 1.39,\ 0.4,\ -0.35,\ -0.11$

14.

15. $z = -0.74$ for $x = 1,\ z = -0.37$ for $x = 2,\ z = 1.48$ for $x = 7$

16.

17. a. 1
    b. 1
    c. 1
    d. $z = -1$ for $x = 0,\ z = 1$ for $x = 2$

18.

19. 16

20.

21. 4.9

22.

23. a. 55
    b. 55

24.

25. a. 93
    b. 0.07

26.

27. a. -1.11
    b. 0.73

28.

29. a. $Q_1 = 59,\ Q_2 = 70,\ Q_3 = 81$
    b. $x_{min} = 39,\ Q_1 = 59,\ Q_2 = 70,\ Q_3 = 81,\ x_{max} = 100$
    c. $R = 61,\ IQR = 22$

30.

31. $x_{min} = 26,\ Q_1 = 28,\ Q_2 = 28,\ Q_3 = 29,\ x_{max} = 32$

32.

33. a. $Q_1 = 1450,\ Q_2 = 2000,\ Q_3 = 2800$
    b. $IQR = 1350$
    c. $x_{min} = 400,\ Q_1 = 1450,\ Q_2 = 2000,\ Q_3 = 2800,\ x_{max} = 3800$

34.

35. Emilia: $z = 0.875$, Ferdinand: $z = 1.1\overline{6}$

36.

37. Rosencrantz: $z = 2.2$, Guildenstern: $z = 1.95$. Rosencrantz is more overweight for his age and body type.

38.

39. a. $x_{min} = 15,\ Q_1 = 51,\ Q_2 = 67,\ Q_3 = 82,\ x_{max} = 97$
    b. The data set appears to be skewed to the left.

40.

41. a. $Q_1 = 440,\ Q_2 = 552.5,\ Q_3 = 661$ and $IQR = 221$
    b. $Q_1 = 641,\ Q_2 = 667,\ Q_3 = 700$ and $IQR = 59$
    c. $Q_1 = 407,\ Q_2 = 448,\ Q_3 = 504$ and $IQR = 97$

2.5 The Empirical Rule and Chebyshev's Theorem

Basic

1. State the Empirical Rule.
2. Describe the conditions under which the Empirical Rule may be applied.
3. State Chebyshev's Theorem.
4. Describe the conditions under which Chebyshev's Theorem may be applied.
5. A sample data set with a bell-shaped distribution has mean $\bar{x} = 6$ and standard deviation $s = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
6. A population data set with a bell-shaped distribution has mean $\mu = 6$ and standard deviation $\sigma = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
7. A population data set with a bell-shaped distribution has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. above 2;
   b. above 3.1;
   c. between 2 and 3.1.
8. A sample data set with a bell-shaped distribution has mean $\bar{x} = 2$ and standard deviation $s = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. below -0.2;
   b. below 3.1;
   c. between -1.3 and 0.9.
9. A population data set with a bell-shaped distribution and size $N = 500$ has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate number of observations in the data set that lie:
   a. above 2;
   b. above 3.1;
   c. between 2 and 3.1.
10. A sample data set with a bell-shaped distribution and size $n = 128$ has mean $\bar{x} = 2$ and standard deviation $s = 1.1$. Find the approximate number of observations in the data set that lie:
    a. below -0.2;
    b. below 3.1;
    c. between -1.3 and 0.9.
11. A sample data set has mean $\bar{x} = 6$ and standard deviation $s = 2$. Find the minimum proportion of observations in the data set that must lie:
    a. between 2 and 10;
    b. between 0 and 12;
    c. between 4 and 8.
12. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the minimum proportion of observations in the data set that must lie:
    a. between -0.2 and 4.2;
    b. between -1.3 and 5.3.

13. A population data set of size $N = 500$ has mean $\mu = 5.2$ and standard deviation $\sigma = 1.1$. Find the minimum number of observations in the data set that must lie:
    a. between 3 and 7.4;
    b. between 1.9 and 8.5.
14. A sample data set of size $n = 128$ has mean $\bar{x} = 2$ and standard deviation $s = 2$. Find the minimum number of observations in the data set that must lie:
    a. between -2 and 6 (including -2 and 6);
    b. between -4 and 8 (including -4 and 8).
15. A sample data set of size $n = 30$ has mean $\bar{x} = 6$ and standard deviation $s = 2$.
    a. What is the maximum proportion of observations in the data set that can lie outside the interval $(2, 10)$?
    b. What can be said about the proportion of observations in the data set that are below 2?
    c. What can be said about the proportion of observations in the data set that are above 10?
    d. What can be said about the number of observations in the data set that are above 10?
16. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$.
    a. What is the maximum proportion of observations in the data set that can lie outside the interval $(-1.3, 5.3)$?
    b. What can be said about the proportion of observations in the data set that are below -1.3?
    c. What can be said about the proportion of observations in the data set that are above 5.3?

Applications

17. Scores on a final exam taken by $1,200$ students have a bell-shaped distribution with mean 72 and standard deviation 9.
    a. What is the median score on the exam?
    b. About how many students scored between 63 and 81?
    c. About how many students scored between 72 and 90?
    d. About how many students scored below 54?
18. Lengths of fish caught by a commercial fishing boat have a bell-shaped distribution with mean 23 inches and standard deviation 1.5 inches.
    a. About what proportion of all fish caught are between 20 inches and 26 inches long?
    b. About what proportion of all fish caught are between 20 inches and 23 inches long?
    c. About how long is the longest fish caught (only a small fraction of a percent are longer)?
19. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped, has mean 5.75 ounces and standard deviation 0.125 ounce, what proportion of the pucks will be usable in professional games?
20. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped and has mean 5.75 ounces, how large can the standard deviation be if 99.7% of the pucks are to be usable in professional games?
21. Speeds of vehicles on a section of highway have a bell-shaped distribution with mean $60\ mph$ and standard deviation $2.5\ mph$.
    a. If the speed limit is $55\ mph$, about what proportion of vehicles are speeding?
    b. What is the median speed for vehicles on this highway?
    c. What is the percentile rank of the speed $65\ mph$?
    d. What speed corresponds to the 16_{th} percentile?
22. Suppose that, as in the previous exercise, speeds of vehicles on a section of highway have mean $60\ mph$ and standard deviation $2.5\ mph$, but now the distribution of speeds is unknown.
    a. If the speed limit is $55\ mph$, at least what proportion of vehicles must speeding?
    b. What can be said about the proportion of vehicles going $65\ mph$ or faster?
23. An instructor announces to the class that the scores on a recent exam had a bell-shaped distribution with mean 75 and standard deviation 5.
    a. What is the median score?
    b. Approximately what proportion of students in the class scored between 70 and 80?
    c. Approximately what proportion of students in the class scored above 85?
    d. What is the percentile rank of the score 85?

24. The GPAs of all currently registered students at a large university have a bell-shaped distribution with mean 2.7 and standard deviation 0.6. Students with a GPA below 1.5 are placed on academic probation. Approximately what percentage of currently registered students at the university are on academic probation?
25. Thirty-six students took an exam on which the average was 80 and the standard deviation was 6. A rumor says that five students had scores 61 or below. Can the rumor be true? Why or why not?

Additional Exercises

26. For the sample data

$$\begin{array}{c|ccccccc} x & 26 & 27 & 28 & 29 & 30 & 31 & 32 \\ \hline f & 3 & 4 & 16 & 12 & 6 & 2 & 1 \end{array} \tag{2.E.78}$$

$$\sum x = 1,256 \text{ and } \sum x^2 = 35,926 \tag{2.E.79}$$

a. Compute the mean and the standard deviation.
b. About how many of the measurements does the Empirical Rule predict will be in the interval $(\bar{x} - s, \bar{x} + s)$, the interval $(\bar{x} - 2s, \bar{x} + 2s)$, and the interval $(\bar{x} - 3s, \bar{x} + 3s)$?
c. Compute the number of measurements that are actually in each of the intervals listed in part (a), and compare to the predicted numbers.

27. A sample of size $n = 80$ has mean 139 and standard deviation 13, but nothing else is known about it.
a. What can be said about the number of observations that lie in the interval $(126, 152)$?
b. What can be said about the number of observations that lie in the interval $(113, 165)$?
c. What can be said about the number of observations that exceed 165?
d. What can be said about the number of observations that either exceed 165 or are less than 113?

28. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 2 & 3 & 4 & 5 \\ \hline f & 84 & 29 & 3 & 3 & 1 \end{array} \tag{2.E.80}$$

$$\sum x = 168 \text{ and } \sum x^2 = 300 \tag{2.E.81}$$

a. Compute the sample mean and the sample standard deviation.
b. Considering the shape of the data set, do you expect the Empirical Rule to apply? Count the number of measurements within one standard deviation of the mean and compare it to the number predicted by the Empirical Rule.
c. What does Chebyshev's Rule say about the number of measurements within one standard deviation of the mean?
d. Count the number of measurements within two standard deviations of the mean and compare it to the minimum number guaranteed by Chebyshev's Theorem to lie in that interval.

29. For the sample data set

$$\begin{array}{c|ccccc} x & 47 & 48 & 49 & 50 & 51 \\ \hline f & 1 & 3 & 18 & 2 & 1 \end{array} \tag{2.E.82}$$

$$\sum x = 1224 \text{ and } \sum x^2 = 59,940 \tag{2.E.83}$$

a. Compute the sample mean and the sample standard deviation.
b. Considering the shape of the data set, do you expect the Empirical Rule to apply? Count the number of measurements within one standard deviation of the mean and compare it to the number predicted by the Empirical Rule.
c. What does Chebyshev's Rule say about the number of measurements within one standard deviation of the mean?
d. Count the number of measurements within two standard deviations of the mean and compare it to the minimum number guaranteed by Chebyshev's Theorem to lie in that interval.

Answers

1. See the displayed statement in the text.
2.
3. See the displayed statement in the text.
4.

5. a. 0.68
   b. 0.95
   c. 0.997
6.
7. a. 0.5
   b. 0.16
   c. 0.34
8.
9. a. 250
   b. 80
   c. 170
10.
11. a. $3/4$
    b. $8/9$
    c. 0
12.
13. a. 375
    b. 445
14.
15. a. At most 0.25.
    b. At most 0.25.
    c. At most 0.25.
    d. At most 7.
16.
17. a. 72
    b. 816
    c. 570
    d. 30
18.
19. 0.95
20.
21. a. 0.975
    b. 60
    c. 97.5
    d. 57.5
22.
23. a. 75
    b. 0.68
    c. 0.025
    d. 0.975
24.
25. By Chebyshev's Theorem at most $1/9$ of the scores can be below 62, so the rumor is impossible.
26.
27. a. Nothing.
    b. It is at least 60.
    c. It is at most 20.
    d. It is at most 20.
28.
29. a. $\bar{x} = 48.96$, $s = 0.7348$.

https://stats.libretexts.org/@go/page/1327

b. Roughly bell-shaped, the Empirical Rule should apply. **True count**: 18, **Predicted**: 17.
c. Nothing.
d. **True count**: 23, **Guaranteed**: at least 18.75, hence at least 19.

Contributor

- Anonymous

This page titled 2.E: Descriptive Statistics (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **2.E: Descriptive Statistics (Exercises)** has no license indicated.

CHAPTER OVERVIEW

3: Basic Concepts of Probability

Suppose a polling organization questions 1,200 voters in order to estimate the proportion of all voters who favor a particular bond issue. We would expect the proportion of the 1,200 voters in the survey who are in favor to be close to the proportion of all voters who are in favor, but this need not be true. There is a degree of randomness associated with the survey result. If the survey result is highly likely to be close to the true proportion, then we have confidence in the survey result. If it is not particularly likely to be close to the population proportion, then we would perhaps not take the survey result too seriously. The likelihood that the survey proportion is close to the population proportion determines our confidence in the survey result. For that reason, we would like to be able to compute that likelihood. The task of computing it belongs to the realm of probability, which we study in this chapter.

3.1: Sample Spaces, Events, and Their Probabilities

3.2: Complements, Intersections, and Unions

3.3: Conditional Probability and Independent Events

3.E: Basic Concepts of Probability (Exercises)

This page titled 3: Basic Concepts of Probability is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

3.1: Sample Spaces, Events, and Their Probabilities

Learning Objectives

- To learn the concept of the sample space associated with a random experiment.
- To learn the concept of an event associated with a random experiment.
- To learn the concept of the probability of an event.

Sample Spaces and Events

Rolling an ordinary six-sided die is a familiar example of a *random experiment*, an action for which all possible outcomes can be listed, but for which the actual outcome on any given trial of the experiment cannot be predicted with certainty. In such a situation we wish to assign to each outcome, such as rolling a two, a number, called the *probability* of the outcome, that indicates how likely it is that the outcome will occur. Similarly, we would like to assign a probability to any *event*, or collection of outcomes, such as rolling an even number, which indicates how likely it is that the event will occur if the experiment is performed. This section provides a framework for discussing probability problems, using the terms just mentioned.

Definition: random experiment

A *random experiment* is a mechanism that produces a definite outcome that cannot be predicted with certainty. The sample space associated with a random experiment is the set of all possible outcomes. An event is a subset of the sample space.

Definition: Element and Occurrence

An event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E.

Example 3.1.1: Sample Space for a single coin

Construct a sample space for the experiment that consists of tossing a single coin.

Solution

The outcomes could be labeled h for heads and t for tails. Then the sample space is the set: $S = \{h, t\}$

Example 3.1.2: Sample Space for a single die

Construct a sample space for the experiment that consists of rolling a single die. Find the events that correspond to the phrases "an even number is rolled" and "a number greater than two is rolled."

Solution

The outcomes could be labeled according to the number of dots on the top face of the die. Then the sample space is the set $S = \{1, 2, 3, 4, 5, 6\}$

The outcomes that are even are $2, 4, \text{ and } 6$, so the event that corresponds to the phrase "an even number is rolled" is the set $\{2, 4, 6\}$, which it is natural to denote by the letter E. We write $E = \{2, 4, 6\}$.

Similarly the event that corresponds to the phrase "a number greater than two is rolled" is the set $T = \{3, 4, 5, 6\}$, which we have denoted T.

A graphical representation of a sample space and events is a *Venn diagram*, as shown in Figure 3.1.1. In general the sample space S is represented by a rectangle, outcomes by points within the rectangle, and events by ovals that enclose the outcomes that compose them.



(a): One Coin Toss (b): One Die Roll

Figure 3.1.1: Venn Diagrams for Two Sample Spaces

Example 3.1.3: Sample Spaces for two coines

A random experiment consists of tossing two coins.

a. Construct a sample space for the situation that the coins are indistinguishable, such as two brand new pennies.
b. Construct a sample space for the situation that the coins are distinguishable, such as one a penny and the other a nickel.

Solution

a. After the coins are tossed one sees either two heads, which could be labeled $2h$, two tails, which could be labeled $2t$, or coins that differ, which could be labeled d Thus a sample space is $S = \{2h, 2t, d\}$.
b. Since we can tell the coins apart, there are now two ways for the coins to differ: the penny heads and the nickel tails, or the penny tails and the nickel heads. We can label each outcome as a pair of letters, the first of which indicates how the penny landed and the second of which indicates how the nickel landed. A sample space is then $S' = \{hh, ht, th, tt\}$.

A device that can be helpful in identifying all possible outcomes of a random experiment, particularly one that can be viewed as proceeding in stages, is what is called a *tree diagram*. It is described in the following example.

Example 3.1.4: Tree diagram

Construct a sample space that describes all three-child families according to the genders of the children with respect to birth order.

Solution

Two of the outcomes are "two boys then a girl," which we might denote bbg, and "a girl then two boys," which we would denote gbb.

Clearly there are many outcomes, and when we try to list all of them it could be difficult to be sure that we have found them all unless we proceed systematically. The tree diagram shown in Figure 3.1.2, gives a systematic approach.

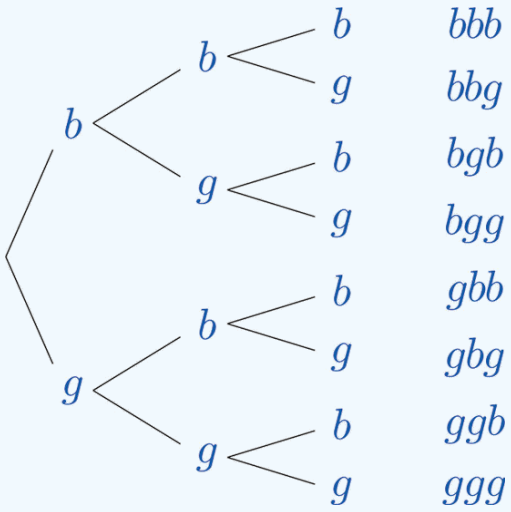


Figure 3.1.2: *Tree Diagram For Three-Child Families*

The diagram was constructed as follows. There are two possibilities for the first child, boy or girl, so we draw two line segments coming out of a starting point, one ending in a b for "boy" and the other ending in a g for "girl." For each of these two possibilities

for the first child there are two possibilities for the second child, “boy” or “girl,” so from each of the b and g we draw two line segments, one segment ending in a b and one in a g. For each of the four ending points now in the diagram there are two possibilities for the third child, so we repeat the process once more.

The line segments are called **branches** of the tree. The right ending point of each branch is called a **node**. The nodes on the extreme right are the **final nodes**; to each one there corresponds an outcome, as shown in the figure.

From the tree it is easy to read off the eight outcomes of the experiment, so the sample space is, reading from the top to the bottom of the final nodes in the tree,

$$S = \{bbb,\ bbg,\ bgb,\ bgg,\ gbb,\ gbg,\ ggb,\ ggg\}$$

Probability

Definition: probability

The probability of an outcome e in a sample space S is a number P between 1 and 0 that measures the likelihood that e will occur on a single trial of the corresponding random experiment. The value $P = 0$ corresponds to the outcome e being impossible and the value $P = 1$ corresponds to the outcome e being certain.

Definition: probability of an event

The *probability of an event* A is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted $P(A)$.

The following formula expresses the content of the definition of the probability of an event:

If an event E is $E = \{e_1, e_2, ..., e_k\}$, then

$$P(E) = P(e_1) + P(e_2) + ... + P(e_k)$$

The following figure expresses the content of the definition of the probability of an event:

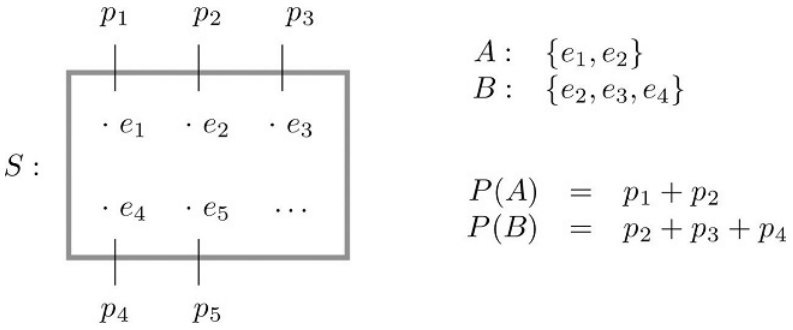


Figure 3.1.3: *Sample Spaces and Probability*

Since the whole sample space S is an event that is certain to occur, the sum of the probabilities of all the outcomes must be the number 1.

In ordinary language probabilities are frequently expressed as percentages. For example, we would say that there is a 70% chance of rain tomorrow, meaning that the probability of rain is 0.70. We will use this practice here, but in all the computational formulas that follow we will use the form 0.70 and not 70%.

Example 3.1.5

A coin is called “balanced” or “fair” if each side is equally likely to land up. Assign a probability to each outcome in the sample space for the experiment that consists of tossing a single fair coin.

Solution

With the outcomes labeled h for heads and t for tails, the sample space is the set

$$S = \{h, t\}$$

Since the outcomes have the same probabilities, which must add up to 1, each outcome is assigned probability $1/2$.

Example 3.1.6

A die is called "balanced" or "fair" if each side is equally likely to land on top. Assign a probability to each outcome in the sample space for the experiment that consists of tossing a single fair die. Find the probabilities of the events E: "an even number is rolled" and T: "a number greater than two is rolled."

Solution

With outcomes labeled according to the number of dots on the top face of the die, the sample space is the set

$$S = \{1, 2, 3, 4, 5, 6\}$$

Since there are six equally likely outcomes, which must add up to 1, each is assigned probability $1/6$.

Since $E = \{2, 4, 6\}$,

$$P(E) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Since $T = \{3, 4, 5, 6\}$,

$$P(T) = \frac{4}{6} = \frac{2}{3}$$

Example 3.1.7

Two fair coins are tossed. Find the probability that the coins match, i.e., either both land heads or both land tails.

Solution

In Example 3.1.3 we constructed the sample space $S = \{2h, 2t, d\}$ for the situation in which the coins are identical and the sample space $S' = \{hh, ht, th, tt\}$ for the situation in which the two coins can be told apart.

The theory of probability does not tell us how to assign probabilities to the outcomes, only what to do with them once they are assigned. Specifically, using sample space S, matching coins is the event $M = \{2h, 2t\}$ which has probability $P(2h) + P(2t)$. Using sample space S', matching coins is the event $M' = \{hh, tt\}$, which has probability $P(hh) + P(tt)$. In the physical world it should make no difference whether the coins are identical or not, and so we would like to assign probabilities to the outcomes so that the numbers $P(M)$ and $P(M')$ are the same and best match what we observe when actual physical experiments are performed with coins that seem to be fair. Actual experience suggests that the outcomes in S' are equally likely, so we assign to each probability $\frac{1}{4}$, and then...

$$P(M') = P(hh) + P(tt) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Similarly, from experience appropriate choices for the outcomes in S are:

$$P(2h) = \frac{1}{4}$$

$$P(2t) = \frac{1}{4}$$

$$P(d) = \frac{1}{2}$$

The previous three examples illustrate how probabilities can be computed simply by counting when the sample space consists of a finite number of equally likely outcomes. In some situations the individual outcomes of any sample space that represents the experiment are unavoidably unequally likely, in which case probabilities cannot be computed merely by counting, but the computational formula given in the definition of the probability of an event must be used.

Example 3.1.8

The breakdown of the student body in a local high school according to race and ethnicity is 51% white, 27% black, 11% Hispanic, 6% Asian, and 5% for all others. A student is randomly selected from this high school. (To select "randomly" means that every student has the same chance of being selected.) Find the probabilities of the following events:

a. B: the student is black,
b. M: the student is minority (that is, not white),
c. N: the student is not black.

Solution

The experiment is the action of randomly selecting a student from the student population of the high school. An obvious sample space is $S = \{w, b, h, a, o\}$. Since 51% of the students are white and all students have the same chance of being selected, $P(w) = 0.51$, and similarly for the other outcomes. This information is summarized in the following table:

Outcome	w	b	h	a	o
Probability	0.51	0.27	0.11	0.06	0.05

a. Since $B = \{b\}$, $P(B) = P(b) = 0.27$
b. Since $M = \{b, h, a, o\}$, $P(M) = P(b) + P(h) + P(a) + P(o) = 0.27 + 0.11 + 0.06 + 0.05 = 0.49$
c. Since $N = \{w, h, a, o\}$, $P(N) = P(w) + P(h) + P(a) + P(o) = 0.51 + 0.11 + 0.06 + 0.05 = 0.73$

Example 3.1.9

The student body in the high school considered in the last example may be broken down into ten categories as follows: 25% white male, 26% white female, 12% black male, 15% black female, 6% Hispanic male, 5% Hispanic female, 3% Asian male, 3% Asian female, 1% male of other minorities combined, and 4% female of other minorities combined. A student is randomly selected from this high school. Find the probabilities of the following events:

a. B: the student is black
b. MF: the student is a non-white female
c. FN: the student is female and is not black

Solution

Now the sample space is $S = \{wm, bm, hm, am, om, wf, bf, hf, af, of\}$. The information given in the example can be summarized in the following table, called a two-way contingency table:

Gender	Race / Ethnicity				
	White	Black	Hispanic	Asian	Others
Male	0.25	0.12	0.06	0.03	0.01
Female	0.26	0.15	0.05	0.03	0.04

a. Since $B = \{bm, bf\}$, $P(B) = P(bm) + P(bf) = 0.12 + 0.15 = 0.27$
b. Since $MF = \{bf, hf, af, of\}$, $P(M) = P(bf) + P(hf) + P(af) + P(of) = 0.15 + 0.05 + 0.03 + 0.04 = 0.27$
c. Since $FN = \{wf, hf, af, of\}$, $P(FN) = P(wf) + P(hf) + P(af) + P(of) = 0.26 + 0.05 + 0.03 + 0.04 = 0.38$

Key Takeaway

- The sample space of a random experiment is the collection of all possible outcomes.
- An event associated with a random experiment is a subset of the sample space.
- The probability of any outcome is a number between 0 and 1. The probabilities of all the outcomes add up to 1.
- The probability of any event A is the sum of the probabilities of the outcomes in A.

This page titled 3.1: Sample Spaces, Events, and Their Probabilities is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

3.2: Complements, Intersections, and Unions

Learning Objectives

- To learn how some events are naturally expressible in terms of other events.
- To learn how to use special formulas for the probability of an event that is expressed in terms of one or more other events.

Some events can be naturally expressed in terms of other, sometimes simpler, events.

Complements

Definition: Complement

The complement of an event A in a sample space S, denoted A^c, is the collection of all outcomes in S that are not elements of the set A. It corresponds to negating any description in words of the event A.

Example 3.2.1

Two events connected with the experiment of rolling a single die are E: ***"the number rolled is even"*** and T: ***"the number rolled is greater than two."*** Find the complement of each.

Solution

In the sample space $S = \{1, 2, 3, 4, 5, 6\}$ the corresponding sets of outcomes are $E = \{2, 4, 6\}$ and $T = \{3, 4, 5, 6\}$. The complements are $E^c = \{1, 3, 5\}$ and $T^c = \{1, 2\}$.

In words the complements are described by "the number rolled is not even" and "the number rolled is not greater than two." Of course easier descriptions would be "the number rolled is odd" and "the number rolled is less than three."

If there is a 60% chance of rain tomorrow, what is the probability of fair weather? The obvious answer, 40%, is an instance of the following general rule.

Definition: Probability Rule for Complements

The Probability Rule for Complements states that

$$P(A^c) = 1 - P(A)$$

This formula is particularly useful when finding the probability of an event directly is difficult.

Example 3.2.2

Find the probability that at least one heads will appear in five tosses of a fair coin.

Solution

Identify outcomes by lists of five hs and ts, such as $tthtt$ and $hhttt$. Although it is tedious to list them all, it is not difficult to count them. Think of using a tree diagram to do so. There are two choices for the first toss. For each of these there are two choices for the second toss, hence $2 \times 2 = 4$ outcomes for two tosses. For each of these four outcomes, there are two possibilities for the third toss, hence $4 \times 2 = 8$ outcomes for three tosses. Similarly, there are $8 \times 2 = 16$ outcomes for four tosses and finally $16 \times 2 = 32$ outcomes for five tosses.

Let O denote the event "at least one heads." There are many ways to obtain at least one heads, but only one way to fail to do so: all tails. Thus although it is difficult to list all the outcomes that form O, it is easy to write $O^c = \{ttttt\}$. Since there are 32 equally likely outcomes, each has probability $\frac{1}{32}$, so $P(O^c) = 1 / 32$, hence $P(O) = 1 - \frac{1}{32} \approx 0.97$ or about a 97% chance.

Intersection of Events

Definition: intersections

The intersection of events A and B, denoted $A \cap B$, is the collection of all outcomes that are elements of both of the sets A and B. It corresponds to combining descriptions of the two events using the word "and."

To say that the event $A \cap B$ occurred means that on a particular trial of the experiment both A and B occurred. A visual representation of the intersection of events A and B in a sample space S is given in Figure 3.2.1. The intersection corresponds to the shaded lens-shaped region that lies within both ovals.

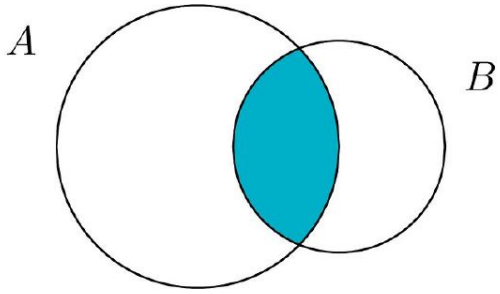


Figure 3.2.1: *The Intersection of Events A and B*

Example 3.2.3

In the experiment of rolling a single die, find the intersection $E \cap T$ of the events E: *"the number rolled is even"* and T: *"the number rolled is greater than two."*

Solution

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Since the outcomes that are common to $E = \{2, 4, 6\}$ and $T = \{3, 4, 5, 6\}$ are 4 and 6, $E \cap T = \{4, 6\}$.

In words the intersection is described by "the number rolled is even and is greater than two." The only numbers between one and six that are both even and greater than two are four and six, corresponding to $E \cap T$ given above.

Example 3.2.4

A single die is rolled.

1. Suppose the die is fair. Find the probability that the number rolled is both even and greater than two.
2. Suppose the die has been "loaded" so that $P(1) = \frac{1}{12}$, $P(6) = \frac{3}{12}$, and the remaining four outcomes are equally likely with one another. Now find the probability that the number rolled is both even and greater than two.

Solution

In both cases the sample space is $S = \{1, 2, 3, 4, 5, 6\}$ and the event in question is the intersection $E \cap T = \{4, 6\}$ of the previous example.

1. Since the die is fair, all outcomes are equally likely, so by counting we have $P(E \cap T) = \frac{2}{6}$.
2. The information on the probabilities of the six outcomes that we have so far is

Outcome	1	2	3	4	5	6
Probablity	$\frac{1}{12}$	p	p	p	p	$\frac{3}{12}$

Since $P(1) + P(6) = \frac{4}{6} = \frac{1}{3}$

$$P(2) + P(3) + P(4) + P(5) = 1 - \frac{1}{3} = \frac{2}{3}$$

Thus $4p = \frac{2}{3}$, so $p = \frac{1}{6}$. In particular $P(4) = \frac{1}{6}$ therefore:

1 3 5

$$P(E \cap T) = P(4) + P(6) = \frac{1}{6} + \frac{1}{12} = \frac{3}{12}$$

Definition: mutually exclusive

Events A and B are mutually exclusive (cannot both occur at once) if they have no elements in common.

For A and B to have no outcomes in common means precisely that it is impossible for both A and B to occur on a single trial of the random experiment. This gives the following rule:

Definition: Probability Rule for Mutually Exclusive Events

Events A and B are mutually exclusive if and only if

$$P(A \cap B) = 0$$

Any event A and its complement A^c are mutually exclusive, but A and B can be mutually exclusive without being complements.

Example 3.2.5

In the experiment of rolling a single die, find three choices for an event A so that the events A and E: "the number rolled is even" are mutually exclusive.

Solution

Since $E = \{2, 4, 6\}$ and we want A to have no elements in common with E, any event that does not contain any even number will do. Three choices are $\{1, 3, 5\}$ (the complement E^c, the odds), $\{1, 3\}$, and $\{5\}$.

Union of Events

Definition: Union of Events

The union of events A and B, denoted $A \cup B$, is the collection of all outcomes that are elements of one or the other of the sets A and B, or of both of them. It corresponds to combining descriptions of the two events using the word "or."

To say that the event $A \cup B$ occurred means that on a particular trial of the experiment either A or B occurred (or both did). A visual representation of the union of events A and B in a sample space S is given in Figure 3.2.2. The union corresponds to the shaded region.

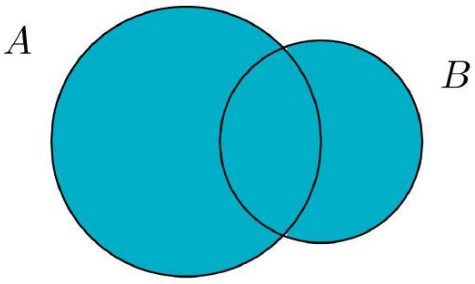


Figure 3.2.2: *The Union of Events A and B*

Example 3.2.6

In the experiment of rolling a single die, find the union of the events E**: "the number rolled is even"** and T**: "the number rolled is greater than two."**

Solution

Since the outcomes that are in either $E = \{2, 4, 6\}$ or $T = \{3, 4, 5, 6\}$ (or both) are $2, 3, 4, 5,$ and 6, that means $E \cup T = \{2, 3, 4, 5, 6\}$.

Note that an outcome such as 4 that is in both sets is still listed only once (although strictly speaking it is not incorrect to list it twice).

In words the union is described by "the number rolled is even or is greater than two." Every number between one and six except the number one is either even or is greater than two, corresponding to $E \cup T$ given above.

Example 3.2.7

A two-child family is selected at random. Let B denote the event that at least one child is a boy, let D denote the event that the genders of the two children differ, and let M denote the event that the genders of the two children match. Find $B \cup D$ and $B \cup M$.

Solution

A sample space for this experiment is $S = \{bb, bg, gb, gg\}$, where the first letter denotes the gender of the firstborn child and the second letter denotes the gender of the second child. The events $B, D,$ and M are $B = \{bb, bg, gb\}$, $D = \{bg, gb\}$, $M = \{bb, gg\}$.

Each outcome in D is already in B, so the outcomes that are in at least one or the other of the sets B and D is just the set B itself: $B \cup D = \{bb, bg, gb\} = B$.

Every outcome in the whole sample space S is in at least one or the other of the sets B and M, so $B \cup M = \{bb, bg, gb, gg\} = S$.

Definition: Additive Rule of Probability

A useful property to know is the Additive Rule of Probability, which is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The next example, in which we compute the probability of a union both by counting and by using the formula, shows why the last term in the formula is needed.

Example 3.2.8

Two fair dice are thrown. Find the probabilities of the following events:

1. both dice show a four
2. at least one die shows a four

Solution

As was the case with tossing two identical coins, actual experience dictates that for the sample space to have equally likely outcomes we should list outcomes as if we could distinguish the two dice. We could imagine that one of them is red and the other is green. Then any outcome can be labeled as a pair of numbers as in the following display, where the first number in the pair is the number of dots on the top face of the green die and the second number in the pair is the number of dots on the top face of the red die.

Illegal pream-token (1)

1. There are 36 equally likely outcomes, of which exactly one corresponds to two fours, so the probability of a pair of fours is $1/36$.
2. From the table we can see that there are 11 pairs that correspond to the event in question: the six pairs in the fourth row (the green die shows a four) plus the additional five pairs other than the pair 44, already counted, in the fourth column (the red die is four), so the answer is $11/36$. To see how the formula gives the same number, let A_G denote the event that the green die is a four and let A_R denote the event that the red die is a four. Then clearly by counting we get: $P(A_G) = 6/36$ and $P(A_R) = 6/36$. Since $A_G \cap A_R = \{44\}$, $P(A_G \cap A_R) = 1/36$. This is the computation from part 1, of course. Thus by the Additive Rule of Probability we get:

$$P(A_G \cap A_R) = P(A_G) + P(A_R) - P(A_G - A_R) = 6/36 + 6/36 - 1/36 = \frac{11}{}$$

36

Example 3.2.9

A tutoring service specializes in preparing adults for high school equivalence tests. Among all the students seeking help from the service, 63% need help in mathematics, 34% need help in English, and 27% need help in both mathematics and English. What is the percentage of students who need help in either mathematics or English?

Solution

Imagine selecting a student at random, that is, in such a way that every student has the same chance of being selected. Let M denote the event "the student needs help in mathematics" and let E denote the event "the student needs help in English." The information given is that $P(M) = 0.63$, $P(E) = 0.34$ and $P(M \cap E) = 0.27$. Thus the Additive Rule of Probability gives:

$$P(M \cup E) = P(M) + P(E) - P(M \cap E) = 0.63 + 0.34 - 0.27 = 0.70$$

Note how the naïve reasoning that if 63% need help in mathematics and 34% need help in English then 63 plus 34 or 97% need help in one or the other gives a number that is too large. The percentage that need help in both subjects must be subtracted off, else the people needing help in both are counted twice, once for needing help in mathematics and once again for needing help in English. The simple sum of the probabilities would work if the events in question were mutually exclusive, for then $P(A \cap B)$ is zero, and makes no difference.

Example 3.2.10

Volunteers for a disaster relief effort were classified according to both specialty (C: construction, E: education, M: medicine) and language ability (S: speaks a single language fluently, T: speaks two or more languages fluently). The results are shown in the following two-way classification table:

Specialty	Language Ability	
	S	T
C	12	1
E	4	3
M	6	2

The first row of numbers means that 12 volunteers whose specialty is construction speak a single language fluently, and 1 volunteer whose specialty is construction speaks at least two languages fluently. Similarly for the other two rows.

A volunteer is selected at random, meaning that each one has an equal chance of being chosen. Find the probability that:

1. his specialty is medicine and he speaks two or more languages;
2. either his specialty is medicine or he speaks two or more languages;
3. his specialty is something other than medicine.

Solution

When information is presented in a two-way classification table it is typically convenient to adjoin to the table the row and column totals, to produce a new table like this:

Specialty	Language Ability		Total
	S	T	
C	12	1	13
E	4	3	7
M	6	2	8

Specialty	Language Ability		Total
	S	T	
Total	22	6	28

1. The probability sought is $P(M \cap T)$. The table shows that there are 2 such people, out of 28 in all, hence $P(M \cap T) = 2/28 \approx 0.07$ or about a 7% chance.
2. The probability sought is $P(M \cup T)$. The third row total and the grand total in the sample give $P(M) = 8/28$. The second column total and the grand total give $P(T) = 6/28$. Thus using the result from part (1),

$$P(M \cup T) = P(M) + P(T) - P(M \cap T) = 828 + 628 - 228 = 1228 \approx 0.43$$

   or about a 43% chance.

3. This probability can be computed in two ways. Since the event of interest can be viewed as the event $C \cup E$ and the events C and E are mutually exclusive, the answer is, using the first two row totals,

$$P(C \cup E) = P(C) + P(E) - P(C \cap E) = 1328 + 728 - 028 = 2028 \approx 0.71$$

On the other hand, the event of interest can be thought of as the complement M^c of M, hence using the value of $P(M)$computed in part (2),

$$P(M^c) = 1 - P(M) = 1 - 828 = 2028 \approx 0.71$$

as before.

Key Takeaway

- The probability of an event that is a complement or union of events of known probability can be computed using formulas.

This page titled 3.2: Complements, Intersections, and Unions is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

3.3: Conditional Probability and Independent Events

Learning Objectives

- To learn the concept of a conditional probability and how to compute it.
- To learn the concept of independence of events, and how to apply it.

Suppose a fair die has been rolled and you are asked to give the probability that it was a five. There are six equally likely outcomes, so your answer is $1/6$. But suppose that before you give your answer you are given the extra information that the number rolled was odd. Since there are only three odd numbers that are possible, one of which is five, you would certai: nly revise your estimate of the likelihood that a five was rolled from $1/6$ to $1/3$. In general, the *revised* probability that an event A has occurred, taking into account the additional information that another event B has definitely occurred on this trial of the experiment, is called the *conditional probability of A given B* and is denoted by $P(A \mid B)$. The reasoning employed in this example can be generalized to yield the computational formula in the following definition.

Definition: conditional probability

The *conditional probability* of A given B, denoted $P(A \mid B)$, is the probability that event A has occurred in a trial of a random experiment for which it is known that event B has definitely occurred. It may be computed by means of the following formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3.3.1}$$

Example 3.3.1: Rolling a Die

A fair (unbiased) die is rolled.

a. Find the probability that the number rolled is a five, given that it is odd.
b. Find the probability that the number rolled is odd, given that it is a five.

Solution

The sample space for this experiment is the set $S = 1, 2, 3, 4, 5, 6$ consisting of six equally likely outcomes. Let F denote the event "a five is rolled" and let O denote the event "an odd number is rolled," so that

$$F = 5 \text{ and } O = 1, 3, 5$$

a. This is the introductory example, so we already know that the answer is $1/3$. To use Equation 3.3.1 to confirm this we must replace A in the formula (the event whose likelihood we seek to estimate) by F and replace B (the event we know for certain has occurred) by O:

$$P(F \mid O) = \frac{P(F \cap O)}{P(O)}$$

Since

$$F \cap O = 5 \cap 1, 3, 5 = 5, \; P(F \cap O) = 1/6$$

Since

$$O = 1, 3, 5, \; P(O) = 3/6.$$

Thus

$$P(F \mid O) = \frac{P(F \cap O)}{P(O)} = \frac{1/6}{3/6} = \frac{1}{3}$$

b. This is the same problem, but with the roles of F and O reversed. Since we are given that the number that was rolled is five, which is odd, the probability in question must be 1. To apply Equation 3.3.1 to this case we must now replace A (the event whose likelihood we seek to estimate) by O and B (the event we know for certain has occurred) by F:

$$P(O \cap F)$$

$$P(O \mid F) = \frac{P(O \cap F)}{P(F)}$$

Obviously $P(F) = 1/6$. In part (a) we found that $P(F \mid O) = 1/6$. Thus

$$P(O \mid F) = \frac{P(O \cap F)}{P(F)} = \frac{1/6}{1/6} = 1$$

Just as we did not need the computational formula in this example, we do not need it when the information is presented in a *two-way classification table*, as in the next example.

Example 3.3.2: Marriage and Gender

In a sample of 902 individuals under 40 who were or had previously been married, each person was classified according to gender and age at first marriage. The results are summarized in the following two-way classification table, where the meaning of the labels is:

- M: male
- F: female
- E: a teenager when first married
- W: in one's twenties when first married
- H: in one's thirties when first married

	E	W	H	**Total**
M	43	293	114	450
F	82	299	71	452
Total	125	592	185	902

The numbers in the first row mean that 43 people in the sample were men who were first married in their teens, 293 were men who were first married in their twenties, 114 men who were first married in their thirties, and a total of 450 people in the sample were men. Similarly for the numbers in the second row. The numbers in the last row mean that, irrespective of gender, 125 people in the sample were married in their teens, 592 in their twenties, 185 in their thirties, and that there were 902 people in the sample in all. Suppose that the proportions in the sample accurately reflect those in the population of all individuals in the population who are under 40 and who are or have previously been married. Suppose such a person is selected at random.

1. Find the probability that the individual selected was a teenager at first marriage.
2. Find the probability that the individual selected was a teenager at first marriage, given that the person is male.

Solution

It is natural to let E also denote the event that the person selected was a teenager at first marriage and to let M denote the event that the person selected is male.

1. According to the table, the proportion of individuals in the sample who were in their teens at their first marriage is $125/902$. This is the relative frequency of such people in the population, hence $P(E) = 125/902 \approx 0.139$ or about 14%.
2. Since it is known that the person selected is male, all the females may be removed from consideration, so that only the row in the table corresponding to men in the sample applies:

	E	W	H	**Total**
M	43	293	114	450

The proportion of males in the sample who were in their teens at their first marriage is $43/450$. This is the relative frequency of such people in the population of males, hence $P(E/M) = 43/450 \approx 0.096$ or about 10%.

In the next example, the computational formula in the definition must be used.

Example 3.3.3: Body Weigth and hypertension

Suppose that in an adult population the proportion of people who are both overweight and suffer hypertension is 0.09; the proportion of people who are not overweight but suffer hypertension is 0.11; the proportion of people who are overweight but do not suffer hypertension is 0.02; and the proportion of people who are neither overweight nor suffer hypertension is 0.78. An adult is randomly selected from this population.

a. Find the probability that the person selected suffers hypertension given that he is overweight.
b. Find the probability that the selected person suffers hypertension given that he is not overweight.
c. Compare the two probabilities just found to give an answer to the question as to whether overweight people tend to suffer from hypertension.

Solution:

Let H denote the event "the person selected suffers hypertension." Let O denote the event "the person selected is overweight." The probability information given in the problem may be organized into the following contingency table:

	O	O^c
H	0.09	0.11
H^c	0.02	0.78

a. Using the formula in the definition of conditional probability (Equation 3.3.1),

$$P(H|O) = \frac{P(H \cap O)}{P(O)} = \frac{0.09}{0.09 + 0.02} = 0.8182$$

b. Using the formula in the definition of conditional probability (Equation 3.3.1),

$$P(H|O) = \frac{P(H \cap O^c)}{P(O^c)} = \frac{0.11}{0.11 + 0.78} = 0.1236$$

c. $P(H|O) = 0.8182$ is over six times as large as $P(H|O^c) = 0.1236$, which indicates a much higher rate of hypertension among people who are overweight than among people who are not overweight. It might be interesting to note that a direct comparison of $P(H \cap O) = 0.09$ and $P(H \cap O^c) = 0.11$ does not answer the same question.

Independent Events

Although typically we expect the conditional probability $P(A \mid B)$ to be different from the probability $P(A)$ of A, it does not have to be different from $P(A)$. When $P(A \mid B) = P(A)$, the occurrence of B has no effect on the likelihood of A. Whether or not the event A has occurred is independent of the event B.

Using algebra it can be shown that the equality $P(A \mid B) = P(A)$ holds if and only if the equality $P(A \cap B) = P(A) \cdot P(B)$ holds, which in turn is true if and only if $P(B \mid A) = P(B)$. This is the basis for the following definition.

Definition: Independent and Dependent Events

Events A and B are *independent* (i.e., events whose probability of occurring together is the product of their individual probabilities). if

$$P(A \cap B) = P(A) \cdot P(B)$$

If A and B are not independent then they are *dependent*.

The formula in the definition has two practical but exactly opposite uses:

- In a situation in which we can compute all three probabilities $P(A), P(B)$ and $P(A \cap B)$, it is used to check whether or not the events A and B are independent:
  - If $P(A \cap B) = P(A) \cdot P(B)$, then A and B are independent.
  - If $P(A \cap B) \neq P(A) \cdot P(B)$, then A and B are not independent.

- In a situation in which each of $P(A)$ and $P(B)$ can be computed and it is known that A and B are independent, then we can compute $P(A \cap B)$ by multiplying together $P(A)$ and $P(B)$: $P(A \cap B) = P(A) \cdot P(B)$.

Example 3.3.4: Rolling a Die again

A single fair die is rolled. Let $A = \{3\}$ and $B = \{1, 3, 5\}$. Are A and B independent?

Solution

In this example we can compute all three probabilities $P(A) = 1/6$, $P(B) = 1/2$, and $P(A \cap B) = P(\{3\}) = 1/6$. Since the product $P(A) \cdot P(B) = (1/6)(1/2) = 1/12$ is not the same number as $P(A \cap B) = 1/6$, the events A and B are not independent.

Example 3.3.5

The two-way classification of married or previously married adults under 40 according to gender and age at first marriage produced the table

	E	W	H	Total
M	43	293	114	450
F	82	299	71	452
Total	125	592	185	902

Determine whether or not the events F: "female" and E: "was a teenager at first marriage" are independent.

Solution

The table shows that in the sample of 902 such adults, 452 were female, 125 were teenagers at their first marriage, and 82 were females who were teenagers at their first marriage, so that

$$P(F) = \frac{452}{902},$$
$$P(E) = \frac{125}{902}$$
$$P(F \cap E) = \frac{82}{902}$$

Since

$$P(F) \cdot P(E) = \frac{452}{902} \cdot \frac{125}{902} = 0.069$$

is not the same as

$$P(F \cap E) = \frac{82}{902} = 0.091$$

we conclude that the two events are not independent.

Example 3.3.6

Many diagnostic tests for detecting diseases do not test for the disease directly but for a chemical or biological product of the disease, hence are not perfectly reliable. The *sensitivity* of a test is the probability that the test will be positive when administered to a person who has the disease. The higher the sensitivity, the greater the detection rate and the lower the false negative rate.

Suppose the sensitivity of a diagnostic procedure to test whether a person has a particular disease is 92%. A person who actually has the disease is tested for it using this procedure by two independent laboratories.

a. What is the probability that both test results will be positive?

b. What is the probability that at least one of the two test results will be positive?

Solution

a. Let A_1 denote the event "the test by the first laboratory is positive" and let A_2 denote the event "the test by the second laboratory is positive." Since A_1 and A_2 are independent,

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2) \\ &= 0.92 \times 0.92 \\ &= 0.8464 \end{aligned}$$

b. Using the Additive Rule for Probability and the probability just computed,

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.92 + 0.92 - 0.8464 \\ &= 0.9936 \end{aligned}$$

Example 3.3.7: specificity of a diagnostic test

The *specificity* of a diagnostic test for a disease is the probability that the test will be negative when administered to a person who does not have the disease. The higher the specificity, the lower the false positive rate. Suppose the specificity of a diagnostic procedure to test whether a person has a particular disease is 89%.

a. A person who does not have the disease is tested for it using this procedure. What is the probability that the test result will be positive?
b. A person who does not have the disease is tested for it by two independent laboratories using this procedure. What is the probability that both test results will be positive?

Solution

a. Let B denote the event "the test result is positive." The complement of B is that the test result is negative, and has probability the specificity of the test, 0.89. Thus

$$P(B) = 1 - P(B^c) = 1 - 0.89 = 0.11$$

b. Let B_1 denote the event "the test by the first laboratory is positive" and let B_2 denote the event "the test by the second laboratory is positive." Since B_1 and B_2 are independent, by part (a) of the example

$$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2) = 0.11 \times 0.11 = 0.0121$$

The concept of independence applies to any number of events. For example, three events $A,\ B,$ and C are independent if $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$. Note carefully that, as is the case with just two events, this is not a formula that is always valid, but holds precisely when the events in question are independent.

Example 3.3.8: redundancy

The reliability of a system can be enhanced by redundancy, which means building two or more independent devices to do the same job, such as two independent braking systems in an automobile. Suppose a particular species of trained dogs has a 90% chance of detecting contraband in airline luggage. If the luggage is checked three times by three different dogs independently of one another, what is the probability that contraband will be detected?

Solution

Let D_1 denote the event that the contraband is detected by the first dog, D_2 the event that it is detected by the second dog, and D_3 the event that it is detected by the third. Since each dog has a 90% of detecting the contraband, by the Probability Rule for Complements it has a 10% chance of failing. In symbols,

$$P(D_1^c) = 0.10, \quad P(D_2^c) = 0.10, \quad P(D_3^c) = 0.10$$

Let D denote the event that the contraband is detected. We seek $P(D)$. It is easier to find $P(D^c)$, because although there are several ways for the contraband to be detected, there is only one way for it to go undetected: all three dogs must fail. Thus D^c

$= D_1^c \cap D_2^c \cap D_3^c$ and

$$P(D) = 1 - P(D^c) = 1 - P(D_1^c \cap D_2^c \cap D_3^c)$$

But the events D_1, D_2, and D_3 are independent, which implies that their complements are independent, so

$$P(D_1^c \cap D_2^c \cap D_3^c) = P(D_1^c) \cdot P(D_2^c) \cdot P(D_3^c) = 0.10 \times 0.10 \times 0.10 = 0.001$$

Using this number in the previous display we obtain

$$P(D) = 1 - 0.001 = 0.999$$

That is, although any one dog has only a 90% chance of detecting the contraband, three dogs working independently have a 99.9% chance of detecting it.

Probabilities on Tree Diagrams

Some probability problems are made much simpler when approached using a tree diagram. The next example illustrates how to place probabilities on a tree diagram and use it to solve a problem.

Example 3.3.9: A jar of Marbles

A jar contains 10 marbles, 7 black and 3 white. Two marbles are drawn without replacement, which means that the first one is not put back before the second one is drawn.

a. What is the probability that both marbles are black?
b. What is the probability that exactly one marble is black?
c. What is the probability that at least one marble is black?

Solution

A tree diagram for the situation of drawing one marble after the other without replacement is shown in Figure 3.3.1. The circle and rectangle will be explained later, and should be ignored for now.

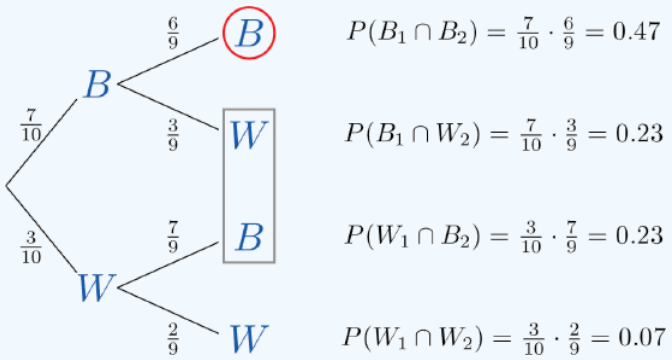


Figure 3.3.1: Tree Diagram for Drawing Two Marbles

The numbers on the two leftmost branches are the probabilities of getting either a black marble, 7 out of 10, or a white marble, 3 out of 10, on the first draw. The number on each remaining branch is the probability of the event corresponding to the node on the right end of the branch occurring, given that the event corresponding to the node on the left end of the branch has occurred. Thus for the top branch, connecting the two Bs, it is $P(B_2 \mid B_1)$, where B_1 denotes the event "the first marble drawn is black" and B_2 denotes the event "the second marble drawn is black." Since after drawing a black marble out there are 9 marbles left, of which 6 are black, this probability is $6/9$.

The number to the right of each final node is computed as shown, using the principle that if the formula in the Conditional Rule for Probability is multiplied by $P(B)$, then the result is

$$P(B \cap A) = P(B) \cdot P(A \mid B)$$

a. The event "both marbles are black" is $B_1 \cap B_2$ and corresponds to the top right node in the tree, which has been circled. Thus as indicated there, it is 0.47.

b. The event “exactly one marble is black” corresponds to the two nodes of the tree enclosed by the rectangle. The events that correspond to these two nodes are mutually exclusive: black followed by white is incompatible with white followed by black. Thus in accordance with the Additive Rule for Probability we merely add the two probabilities next to these nodes, since what would be subtracted from the sum is zero. Thus the probability of drawing exactly one black marble in two tries is $0.23 + 0.23 = 0.46$.

c. The event “at least one marble is black” corresponds to the three nodes of the tree enclosed by either the circle or the rectangle. The events that correspond to these nodes are mutually exclusive, so as in part (b) we merely add the probabilities next to these nodes. Thus the probability of drawing at least one black marble in two tries is $0.47 + 0.23 + 0.23 = 0.93$.

Of course, this answer could have been found more easily using the Probability Law for Complements, simply subtracting the probability of the complementary event, “two white marbles are drawn,” from 1 to obtain $1 - 0.07 = 0.93$.

As this example shows, finding the probability for each branch is fairly straightforward, since we compute it knowing everything that has happened in the sequence of steps so far. Two principles that are true in general emerge from this example:

Probabilities on Tree Diagrams

- The probability of the event corresponding to any node on a tree is the product of the numbers on the unique path of branches that leads to that node from the start.
- If an event corresponds to several final nodes, then its probability is obtained by *adding* the numbers next to those nodes.

Key Takeaway

- A conditional probability is the probability that an event has occurred, taking into account additional information about the result of the experiment.
- A conditional probability can always be computed using the formula in the definition. Sometimes it can be computed by discarding part of the sample space.
- Two events A and B are independent if the probability $P(A \cap B)$ of their intersection $A \cap B$ is equal to the product $P(A) \cdot P(B)$ of their individual probabilities.

This page titled 3.3: Conditional Probability and Independent Events is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

3.E: Basic Concepts of Probability (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

3.1: Sample Spaces, Events, and Their Probabilities

Basic

Q3.1.1

A box contains 10 white and 10 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time. (To draw “with replacement” means that the first marble is put back before the second marble is drawn.)

Q3.1.2

A box contains 16 white and 16 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time. (To draw “with replacement” means that each marble is put back before the next marble is drawn.)

Q3.1.3

A box contains 8 red, 8 yellow, and 8 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time.

Q3.1.4

A box contains 6 red, 6 yellow, and 6 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time.

Q3.1.5

In the situation of **Exercise 1**, list the outcomes that comprise each of the following events.

a. At least one marble of each color is drawn.
b. No white marble is drawn.

Q3.1.6

In the situation of **Exercise 2**, list the outcomes that comprise each of the following events.

a. At least one marble of each color is drawn.
b. No white marble is drawn.
c. More black than white marbles are drawn.

Q3.1.7

In the situation of **Exercise 3**, list the outcomes that comprise each of the following events.

a. No yellow marble is drawn.
b. The two marbles drawn have the same color.
c. At least one marble of each color is drawn.

Q3.1.8

In the situation of **Exercise 4**, list the outcomes that comprise each of the following events.

a. No yellow marble is drawn.
b. The three marbles drawn have the same color.
c. At least one marble of each color is drawn.

Q3.1.9

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 5**.

Q3.1.10

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 6**.

Q3.1.11

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 7**.

Q3.1.12

Assuming that each outcome is equally likely, find the probability of each event in **Exercise 8**.

Q3.1.13

A sample space is $S = \{a, b, c, d, e\}$. Identify two events as $U = \{a, b, d\}$ and $V = \{b, c, d\}$. Suppose $P(a)$ and $P(b)$ are each 0.2 and $P(c)$ and $P(d)$ are each 0.1.

a. Determine what $P(e)$ must be.
b. Find $P(U)$.
c. Find $P(V)$

Q3.1.14

A sample space is $S = \{u, v, w, x\}$. Identify two events as $A = \{v, w\}$ and $B = \{u, w, x\}$. Suppose $P(u) = 0.22$, $P(w) = 0.36$, and $P(x) = 0.27$.

a. Determine what $P(v)$ must be.
b. Find $P(A)$.
c. Find $P(B)$.

Q3.1.15

A sample space is $S = \{m, n, q, r, s\}$. Identify two events as $U = \{m, q, s\}$ and $V = \{n, q, r\}$. The probabilities of some of the outcomes are given by the following table:

Outcome	m	n	q	r	s
Probability	0.18	0.16		0.24	0.21

(3.E.1)

a. Determine what $P(q)$ must be.
b. Find $P(U)$.
c. Find $P(V)$.

Q3.1.16

A sample space is $S = \{d, e, f, g, h\}$. Identify two events as $M = \{e, f, g, h\}$ and $N = \{d, g\}$. The probabilities of some of the outcomes are given by the following table:

Outcome	d	e	f	g	h
Probability	0.22	0.13	0.27		0.19

(3.E.2)

a. Determine what $P(g)$ must be.
b. Find $P(M)$.
c. Find $P(N)$.

Applications

Q3.1.17

The sample space that describes all three-child families according to the genders of the children with respect to birth order was constructed in "Example 3.1.4". Identify the outcomes that comprise each of the following events in the experiment of selecting a three-child family at random.

a. At least one child is a girl.
b. At most one child is a girl.
c. All of the children are girls.

d. Exactly two of the children are girls.
e. The first born is a girl.

Q3.1.18

The sample space that describes three tosses of a coin is the same as the one constructed in "Example 3.1.4" with “boy” replaced by “heads” and “girl” replaced by “tails.” Identify the outcomes that comprise each of the following events in the experiment of tossing a coin three times.

a. The coin lands heads more often than tails.
b. The coin lands heads the same number of times as it lands tails.
c. The coin lands heads at least twice.
d. The coin lands heads on the last toss.

Q3.1.19

Assuming that the outcomes are equally likely, find the probability of each event in **Exercise 17**.

Q3.1.20

Assuming that the outcomes are equally likely, find the probability of each event in **Exercise 18**.

Additional Exercises

Q3.1.21

The following two-way contingency table gives the breakdown of the population in a particular locale according to age and tobacco usage:

Age	Tobacco Use	
	Smoker	Non-smoker
Under 30	0.05	0.20
Over 30	0.20	0.55

A person is selected at random. Find the probability of each of the following events.

a. The person is a smoker.
b. The person is under 30.
c. The person is a smoker who is under 30.

Q3.1.22

The following two-way contingency table gives the breakdown of the population in a particular locale according to party affiliation ($A, B, C,$ or None) and opinion on a bond issue:

Affiliation	Opinion		
	Favors	Opposes	Undecided
A	0.12	0.09	0.07
B	0.16	0.12	0.14
C	0.04	0.03	0.06
None	0.08	0.06	0.03

A person is selected at random. Find the probability of each of the following events.

a. The person is affiliated with party B.
b. The person is affiliated with some party.
c. The person is in favor of the bond issue.

d. The person has no party affiliation and is undecided about the bond issue.

Q3.1.23

The following two-way contingency table gives the breakdown of the population of married or previously married women beyond child-bearing age in a particular locale according to age at first marriage and number of children:

Age	Number of Children		
	0	1 *or* 2	3 or More
Under 20	0.02	0.14	0.08
20 – 29	0.07	0.37	0.11
30 and above	0.10	0.10	0.01

A woman is selected at random. Find the probability of each of the following events.

a. The woman was in her twenties at her first marriage.
b. The woman was 20 or older at her first marriage.
c. The woman had no children.
d. The woman was in her twenties at her first marriage and had at least three children.

Q3.1.24

The following two-way contingency table gives the breakdown of the population of adults in a particular locale according to highest level of education and whether or not the individual regularly takes dietary supplements:

Education	Use of Supplements	
	Takes	Does Not Take
No High School Diploma	0.04	0.06
High School Diploma	0.06	0.44
Undergraduate Degree	0.09	0.28
Graduate Degree	0.01	0.02

An adult is selected at random. Find the probability of each of the following events.

a. The person has a high school diploma and takes dietary supplements regularly.
b. The person has an undergraduate degree and takes dietary supplements regularly.
c. The person takes dietary supplements regularly.
d. The person does not take dietary supplements regularly.

Large Data Set Exercises

Q3.1.25

Large Data Set 4 and Data Set 4A record the results of 500 tosses of a coin. Find the relative frequency of each outcome $1, 2, 3, 4, 5,$ *and* 6. Does the coin appear to be "balanced" or "fair"?

Q3.1.26

Large Data Set 6, Data Set 6A, and Data Set 6B record results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate.

a. Find the probability that a randomly selected voter among these 400 prefers Candidate A.
b. Find the probability that a randomly selected voter among the 200 who live in Region 1 prefers Candidate A (separately recorded in Large Data Set 6A).

c. Find the probability that a randomly selected voter among the 200 who live in Region 2 prefers Candidate A (separately recorded in Large Data Set 6B).

Answers

S3.1.1

$S = \{bb, bw, wb, ww\}$

S3.1.3

$S = \{rr, ry, rg, yr, yy, yg, gr, gy, gg\}$

S3.1.5

a. $\{bw, wb\}$
b. $\{bb\}$

S3.1.7

a. $\{rr, rg, gr, gg\}$
b. $\{rr, yy, gg\}$
c. $\varnothing$

S3.1.9

a. $1/4$
b. $2/4$

S3.1.11

a. $4/9$
b. $3/9$
c. 0

S3.1.13

a. 0.4
b. 0.5
c. 0.4

S3.1.15

a. 0.61
b. 0.6
c. 0.21

S3.1.17

a. $\{gbb, gbg, ggb, ggg\}$
b. $\{bgg, gbg, ggb\}$
c. $\{ggg\}$
d. $\{bbb, bbg, bgb, gbb\}$
e. $\{bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$

S3.1.19

a. $4/8$
b. $3/8$
c. $1/8$
d. $4/8$
e. $7/8$

S3.1.21

a. 0.05
b. 0.25
c. 0.25

S3.1.23

a. 0.11
b. 0.19
c. 0.76
d. 0.55

S3.1.25

The relative frequencies for 1 through 6 are $0.16, 0.194, 0.162, 0.164, 0.154$ *and* 0.166. It would appear that the die is not balanced.

3.2: Complements, Intersections and Unions

Basic

1. For the sample space $S = \{a, b, c, d, e\}$ identify the complement of each event given.
   a. $A = \{a, d, e\}$
   b. $B = \{b, c, d, e\}$
   c. S
2. For the sample space $S = \{r, s, t, u, v\}$ identify the complement of each event given.
   a. $R = \{t, u\}$
   b. $T = \{r\}$
   c. $\varnothing$ (the "empty" set that has no elements)
3. The sample space for three tosses of a coin is $S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$ Define events

$$\begin{array}{c}\text{H:at least one head is observed} \\ \text{M:more heads than tails are observed}\end{array} \tag{3.E.3}$$

   a. List the outcomes that comprise H and M.
   b. List the outcomes that comprise $H \cap M$, $H \cup M$, and H^c.
   c. Assuming all outcomes are equally likely, find $P(H \cap M)$, $P(H \cup M)$, and $P(H^c)$.
   d. Determine whether or not H^c and M are mutually exclusive. Explain why or why not.
4. For the experiment of rolling a single six-sided die once, define events

$$\begin{array}{c}\text{T:the number rolled is three} \\ \text{G:the number rolled is four or greater}\end{array} \tag{3.E.4}$$

   a. List the outcomes that comprise T and G.
   b. List the outcomes that comprise $T \cap G$, $T \cup G$, T^c, and $(T \cup G)^c$.
   c. Assuming all outcomes are equally likely, find $P(T \cap G)$, $P(T \cup G)$, and $P(T^c)$.
   d. Determine whether or not T and G are mutually exclusive. Explain why or why not.
5. A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Define events

$$\begin{array}{c}\text{B:the card is blue} \\ \text{R:the card is red} \\ \text{N:the number on the card is at most two}\end{array} \tag{3.E.5}$$

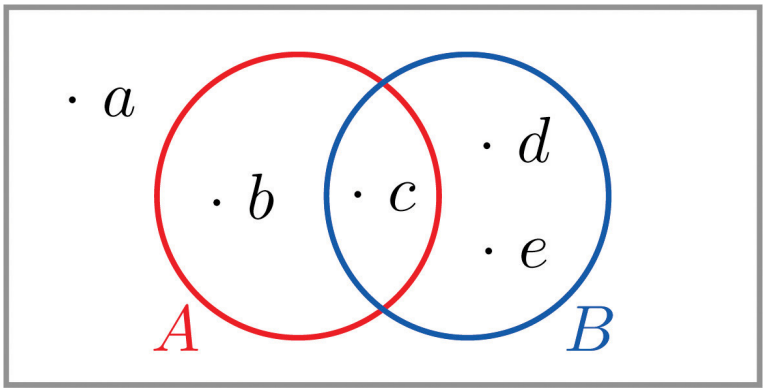
   a. List the outcomes that comprise B, R, and N.
   b. List the outcomes that comprise $B \cap R$, $B \cup R$, $B \cap N$, $R \cup N$, B^c, and $(B \cup R)^c$.
   c. Assuming all outcomes are equally likely, find the probabilities of the events in the previous part.
   d. Determine whether or not B and N are mutually exclusive. Explain why or why not.
6. In the context of the previous problem, define events

$$\text{Y:the card is yellow} \tag{3.E.6}$$

I:the number on the card is not a one
J:the number on the card is a two or a four

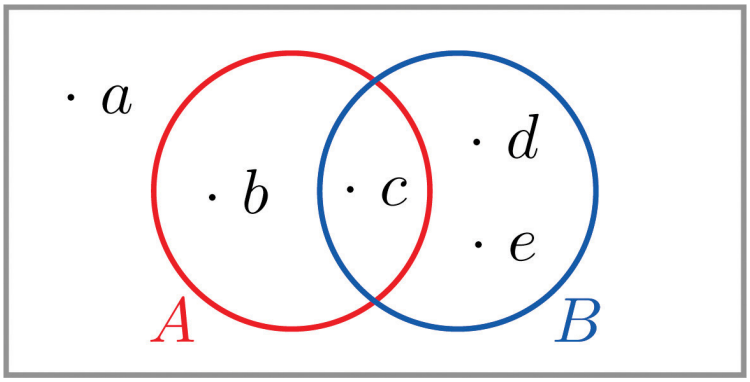
a. List the outcomes that comprise Y, I, and J.
b. List the outcomes that comprise $Y \cap I$, $Y \cup J$, $I \cap J$, I^c, and $(Y \cup J)^c$.
c. Assuming all outcomes are equally likely, find the probabilities of the events in the previous part.
d. Determine whether or not I^c and J are mutually exclusive. Explain why or why not.

7. The Venn diagram provided shows a sample space and two events A and B. Suppose $P(a) = 0.13, P(b) = 0.09, P(c) = 0.27, P(d) = 0.20,$ and $P(e) = 0.31$. Confirm that the probabilities of the outcomes add up to 1, then compute the following probabilities.



a. $P(A)$.
b. $P(B)$.
c. $P(A^c)$. Two ways: (i) by finding the outcomes in A^c and adding their probabilities, and (ii) using the Probability Rule for Complements.
d. $P(A \cap B)$.
e. $P(A \cup B)$ Two ways: (i) by finding the outcomes in $A \cup B$ and adding their probabilities, and (ii) using the Additive Rule of Probability.

8. The Venn diagram provided shows a sample space and two events A and B. Suppose $P(a) = 0.32, P(b) = 0.17, P(c) = 0.28,$ and $P(d) = 0.23$. Confirm that the probabilities of the outcomes add up to 1, then compute the following probabilities.



a. $P(A)$.
b. $P(B)$.
c. $P(A^c)$. Two ways: (i) by finding the outcomes in A^c and adding their probabilities, and (ii) using the Probability Rule for Complements.
d. $P(A \cap B)$.
e. $P(A \cup B)$ Two ways: (i) by finding the outcomes in $A \cup B$ and adding their probabilities, and (ii) using the Additive Rule of Probability.

9. Confirm that the probabilities in the two-way contingency table add up to 1, then use it to find the probabilities of the events indicated.

	U	V	W
A	0.15	0.00	0.23
B	0.22	0.30	0.10

a. $P(A), P(B), P(A \cap B)$.
b. $P(U), P(W), P(U \cap W)$.
c. $P(U \cup W)$.
d. $P(V^c)$.
e. Determine whether or not the events A and U are mutually exclusive; the events A and V.

10. Confirm that the probabilities in the two-way contingency table add up to 1, then use it to find the probabilities of the events indicated.

	R	S	T
M	0.09	0.25	0.19
N	0.31	0.16	0.00

a. $P(R), P(S), P(R \cap S)$.
b. $P(M), P(N), P(M \cap N)$.
c. $P(R \cup S)$.
d. $P(R^c)$.
e. Determine whether or not the events N and S are mutually exclusive; the events N and T.

Applications

11. Make a statement in ordinary English that describes the complement of each event (do not simply insert the word "not").
    a. In the roll of a die: "five or more."
    b. In a roll of a die: "an even number."
    c. In two tosses of a coin: "at least one heads."
    d. In the random selection of a college student: "Not a freshman."
12. Make a statement in ordinary English that describes the complement of each event (do not simply insert the word "not").
    a. In the roll of a die: "two or less."
    b. In the roll of a die: "one, three, or four."
    c. In two tosses of a coin: "at most one heads."
    d. In the random selection of a college student: "Neither a freshman nor a senior."
13. The sample space that describes all three-child families according to the genders of the children with respect to birth order is $S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$. For each of the following events in the experiment of selecting a three-child family at random, state the complement of the event in the simplest possible terms, then find the outcomes that comprise the event and its complement.
    a. At least one child is a girl.
    b. At most one child is a girl.
    c. All of the children are girls.
    d. Exactly two of the children are girls.
    e. The first born is a girl.
14. The sample space that describes the two-way classification of citizens according to gender and opinion on a political issue is $S = \{mf, ma, mn, ff, fa, fn\}$, where the first letter denotes gender (m: male, f: female) and the second opinion (f: for, a: against, n: neutral). For each of the following events in the experiment of selecting a citizen at random, state the complement of the event in the simplest possible terms, then find the outcomes that comprise the event and its complement.
    a. The person is male.
    b. The person is not in favor.
    c. The person is either male or in favor.

d. The person is female and neutral.

15. A tourist who speaks English and German but no other language visits a region of Slovenia. If 35% of the residents speak English, 15% speak German, and 3% speak both English and German, what is the probability that the tourist will be able to talk with a randomly encountered resident of the region?
16. In a certain country 43% of all automobiles have airbags, 27% have anti-lock brakes, and 13% have both. What is the probability that a randomly selected vehicle will have both airbags and anti-lock brakes?
17. A manufacturer examines its records over the last year on a component part received from outside suppliers. The breakdown on source (supplier A, supplier B) and quality (H: high, U: usable, D: defective) is shown in the two-way contingency table.

	H	U	D
A	0.6937	0.0049	0.0014
B	0.2982	0.0009	0.0009

The record of a part is selected at random. Find the probability of each of the following events.

a. The part was defective.
b. The part was either of high quality or was at least usable, in two ways: (i) by adding numbers in the table, and (ii) using the answer to (a) and the Probability Rule for Complements.
c. The part was defective and came from supplier B.
d. The part was defective or came from supplier B, in two ways: by finding the cells in the table that correspond to this event and adding their probabilities, and (ii) using the Additive Rule of Probability.

18. Individuals with a particular medical condition were classified according to the presence (T) or absence (N) of a potential toxin in their blood and the onset of the condition (E: early, M: midrange, L: late). The breakdown according to this classification is shown in the two-way contingency table.

	E	M	L
T	0.012	0.124	0.013
N	0.170	0.638	0.043

One of these individuals is selected at random. Find the probability of each of the following events.

a. The person experienced early onset of the condition.
b. The onset of the condition was either midrange or late, in two ways: (i) by adding numbers in the table, and (ii) using the answer to (a) and the Probability Rule for Complements.
c. The toxin is present in the person's blood.
d. The person experienced early onset of the condition and the toxin is present in the person's blood.
e. The person experienced early onset of the condition or the toxin is present in the person's blood, in two ways: (i) by finding the cells in the table that correspond to this event and adding their probabilities, and (ii) using the Additive Rule of Probability.

19. The breakdown of the students enrolled in a university course by class (F: freshman, So: sophomore, J: junior, Se: senior) and academic major (S: science, mathematics, or engineering, L: liberal arts, O: other) is shown in the two-way classification table.

Major	Class			
	F	So	J	Se
S	92	42	20	13
L	368	167	80	53
O	460	209	100	67

A student enrolled in the course is selected at random. Adjoin the row and column totals to the table and use the expanded table to find the probability of each of the following events.

a. The student is a freshman.
b. The student is a liberal arts major.
c. The student is a freshman liberal arts major.
d. The student is either a freshman or a liberal arts major.
e. The student is not a liberal arts major.

20. The table relates the response to a fund-raising appeal by a college to its alumni to the number of years since graduation.

Response	Years Since Graduation			
	$0-5$	$6-20$	$21-35$	Over 35
Positive	120	440	210	90
None	1380	3560	3290	910

An alumnus is selected at random. Adjoin the row and column totals to the table and use the expanded table to find the probability of each of the following events.

a. The alumnus responded.
b. The alumnus did not respond.
c. The alumnus graduated at least 21 years ago.
d. The alumnus graduated at least 21 years ago and responded.

Additional Exercises

21. The sample space for tossing three coins is $S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$

a. List the outcomes that correspond to the statement "All the coins are heads."
b. List the outcomes that correspond to the statement "Not all the coins are heads."
c. List the outcomes that correspond to the statement "All the coins are not heads."

Answers

1. a. $\{b, c\}$
   b. $\{a\}$
   c. $\varnothing$
2.
3. a. $H = \{hhh, hht, hth, htt, thh, tht, tth\},\ M = \{hhh, hht, hth, thh\}$
   b. $H \cap M = \{hhh, hht, hth, thh\}, H \cup M = H, H^c = \{ttt\}$
   c. $P(H \cap M) = 4/8, P(H \cup M) = 7/8, P(H^c) = 1/8$
   d. Mutually exclusive because they have no elements in common.
4.
5. a. $B = \{b1, b2, b3, b4\},\ R = \{r1, r2, r3, r4\},\ N = \{b1, b2, y1, y2, g1, g2, r1, r2\}$
   b. $B \cap R = \varnothing, B \cup R = \{b1, b2, b3, b4, r1, r2, r3, r4\},\ B \cap N = \{b1, b2\},$
   $R \cup N = \{b1, b2, y1, y2, g1, g2, r1, r2, r3, r4\},$
   $B^c = \{y1, y2, y3, y4, g1, g2, g3, g4, r1, r2, r3, r4\},\ (B \cup R)^c = \{y1, y2, y3, y4, g1, g2, g3, g4\}$
   c. $P(B \cap R) = 0,\ P(B \cup R) = 8/16,\ P(B \cap N) = 2/16,\ P(R \cup N) = 10/16,\ P(B^c) = 12/16,\ P((B \cup R)^c) = 8/16$
   d. Not mutually exclusive because they have an element in common.
6.
7. a. 0.36
   b. 0.78
   c. 0.64
   d. 0.27
   e. 0.87

8.

9. a. $P(A) = 0.38,\ P(B) = 0.62,\ P(A \cap B) = 0$
   b. $P(U) = 0.37,\ P(W) = 0.33,\ P(U \cap W) = 0$
   c. 0.7
   d. 0.7
   e. A and U are not mutually exclusive because $P(A \cap U)$ is the nonzero number 0.15. A and V are mutually exclusive because $P(A \cap V) = 0$.

10.

11. a. “four or less”
    b. “an odd number”
    c. “no heads” or “all tails”
    d. “a freshman”

12.

13. a. “All the children are boys.” Event: $\{bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$, Complement: $\{bbb\}$
    b. “At least two of the children are girls” or “There are two or three girls.” Event: $\{bbb, bbg, bgb, gbb\}$, Complement: $\{bgg, gbg, ggb, ggg\}$
    c. “At least one child is a boy.” Event: $\{ggg\}$, Complement: $\{bbb, bbg, bgb, bgg, gbb, gbg, ggb\}$
    d. “There are either no girls, exactly one girl, or three girls.” Event: $\{bgg, gbg, ggb\}$, Complement: $\{bbb, bbg, bgb, gbb, ggg\}$
    e. “The first born is a boy.” Event: $\{gbb, gbg, ggb, ggg\}$, Complement: $\{bbb, bbg, bgb, bgg\}$

14.

15. 0.47

16.

17. a. 0.0023
    b. 0.9977
    c. 0.0009
    d. 0.3014

18.

19. a. $920/1671$
    b. $668/1671$
    c. $368/1671$
    d. $1220/1671$
    e. $1003/1671$

20.

21. a. $\{hhh\}$
    b. $\{hht, hth, htt, thh, tht, tth, ttt\}$
    c. $\{ttt\}$

3.3: Conditional Probability and Independent Events

Basic

1. Q3.3.1For two events A and B, $P(A) = 0.73,\ P(B) = 0.48$ and $P(A \cap B) = 0.29$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
   c. Determine whether or not A and B are independent.
2. Q3.3.1For two events A and B, $P(A) = 0.26,\ P(B) = 0.37$ and $P(A \cap B) = 0.11$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
   c. Determine whether or not A and B are independent.
3. Q3.3.1For independent events A and B, $P(A) = 0.81$ and $P(B) = 0.27$.
   a. $P(A \cap B)$.

b. Find $P(A \mid B)$.
c. Find $P(B \mid A)$.

4. Q3.3.1For independent events A and B, $P(A) = 0.68$ and $P(B) = 0.37$.
   a. $P(A \cap B)$.
   b. Find $P(A \mid B)$.
   c. Find $P(B \mid A)$.
5. Q3.3.1For mutually exclusive events A and B, $P(A) = 0.17$ and $P(B) = 0.32$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
6. Q3.3.1For mutually exclusive events A and B, $P(A) = 0.45$ and $P(B) = 0.09$.
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
7. Q3.3.1Compute the following probabilities in connection with the roll of a single fair die.
   a. The probability that the roll is even.
   b. The probability that the roll is even, given that it is not a two.
   c. The probability that the roll is even, given that it is not a one.
8. Q3.3.1Compute the following probabilities in connection with two tosses of a fair coin.
   a. The probability that the second toss is heads.
   b. The probability that the second toss is heads, given that the first toss is heads.
   c. The probability that the second toss is heads, given that at least one of the two tosses is heads.
9. Q3.3.1A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Find the following probabilities.
   a. The probability that the card drawn is red.
   b. The probability that the card is red, given that it is not green.
   c. The probability that the card is red, given that it is neither red nor yellow.
   d. The probability that the card is red, given that it is not a four.
10. Q3.3.1A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Find the following probabilities.
    a. The probability that the card drawn is a two or a four.
    b. The probability that the card is a two or a four, given that it is not a one.
    c. The probability that the card is a two or a four, given that it is either a two or a three.
    d. The probability that the card is a two or a four, given that it is red or green.
11. Q3.3.1A random experiment gave rise to the two-way contingency table shown. Use it to compute the probabilities indicated.

	R	S
A	0.12	0.18
B	0.28	0.42

    a. $P(A),\ P(R),\ P(A \cap B)$.
    b. Based on the answer to (a), determine whether or not the events A and R are independent.
    c. Based on the answer to (b), determine whether or not $P(A \mid R)$ can be predicted without any computation. If so, make the prediction. In any case, compute $P(A \mid R)$ using the Rule for Conditional Probability.
12. Q3.3.1A random experiment gave rise to the two-way contingency table shown. Use it to compute the probabilities indicated.

	R	S
A	0.13	0.07
B	0.61	0.19

a. $P(A),\ P(R),\ P(A \cap B)$.
b. Based on the answer to (a), determine whether or not the events A and R are independent.
c. Based on the answer to (b), determine whether or not $P(A \mid R)$ can be predicted without any computation. If so, make the prediction. In any case, compute $P(A \mid R)$ using the Rule for Conditional Probability.

13. Q3.3.1Suppose for events A and B in a random experiment $P(A) = 0.70$ and $P(B) = 0.30$.Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B)$.
    b. $P(A \cap B)$, with the extra information that A and B are independent.
    c. $P(A \cap B)$, with the extra information that A and B are mutually exclusive.
14. Q3.3.1Suppose for events A and B in a random experiment $P(A) = 0.50$ and $P(B) = 0.50$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B)$.
    b. $P(A \cap B)$, with the extra information that A and B are independent.
    c. $P(A \cap B)$, with the extra information that A and B are mutually exclusive.
15. Q3.3.1Suppose for events $A,\ B,\ and\ C$ connected to some random experiment, $A,\ B,\ and\ C$ are independent and $P(A) = 0.50$, $P(B) = 0.50$ and $P(C) = 0.44$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B \cap C)$.
    b. $P(A^c \cap B^c \cap C^c)$.
16. Q3.3.1Suppose for events $A,\ B,\ and\ C$ connected to some random experiment, $A,\ B,\ and\ C$ are independent and $P(A) = 0.95$, $P(B) = 0.73$ and $P(C) = 0.62$. Compute the indicated probability, or explain why there is not enough information to do so.
    a. $P(A \cap B \cap C)$.
    b. $P(A^c \cap B^c \cap C^c)$.

Applications

Q3.3.17

The sample space that describes all three-child families according to the genders of the children with respect to birth order is

$$S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\} \tag{3.E.7}$$

In the experiment of selecting a three-child family at random, compute each of the following probabilities, assuming all outcomes are equally likely.

a. The probability that the family has at least two boys.
b. The probability that the family has at least two boys, given that not all of the children are girls.
c. The probability that at least one child is a boy.
d. The probability that at least one child is a boy, given that the first born is a girl.

Q3.3.18

The following two-way contingency table gives the breakdown of the population in a particular locale according to age and number of vehicular moving violations in the past three years:

Age	Violations		
	0	1	2+
Under 21	0.04	0.06	0.02
21 – 40	0.25	0.16	0.01
41 – 60	0.23	0.10	0.02
60+	0.08	0.03	0.00

A person is selected at random. Find the following probabilities.

a. The person is under 21.

b. The person has had at least two violations in the past three years.
c. The person has had at least two violations in the past three years, given that he is under 21.
d. The person is under 21, given that he has had at least two violations in the past three years.
e. Determine whether the events “the person is under 21” and “the person has had at least two violations in the past three years” are independent or not.

Q3.3.19

The following two-way contingency table gives the breakdown of the population in a particular locale according to party affiliation ($A, B, C,$ or None) and opinion on a bond issue:

Affiliation	Opinion		
	Favors	Opposes	Undecided
A	0.12	0.09	0.07
B	0.16	0.12	0.14
C	0.04	0.03	0.06
None	0.08	0.06	0.03

A person is selected at random. Find each of the following probabilities.

a. The person is in favor of the bond issue.
b. The person is in favor of the bond issue, given that he is affiliated with party A.
c. The person is in favor of the bond issue, given that he is affiliated with party B.

Q3.3.20

The following two-way contingency table gives the breakdown of the population of patrons at a grocery store according to the number of items purchased and whether or not the patron made an impulse purchase at the checkout counter:

Number of Items	Impulse Purchase	
	Made	Not Made
Few	0.01	0.19
Many	0.04	0.76

A patron is selected at random. Find each of the following probabilities.

a. The patron made an impulse purchase.
b. The patron made an impulse purchase, given that the total number of items purchased was many.
c. Determine whether or not the events “few purchases” and “made an impulse purchase at the checkout counter” are independent.

Q3.3.21

The following two-way contingency table gives the breakdown of the population of adults in a particular locale according to employment type and level of life insurance:

Employment Type	Level of Insurance		
	Low	Medium	High
Unskilled	0.07	0.19	0.00
Semi-skilled	0.04	0.28	0.08
Skilled	0.03	0.18	0.05
Professional	0.01	0.05	0.02

An adult is selected at random. Find each of the following probabilities.

a. The person has a high level of life insurance.
b. The person has a high level of life insurance, given that he does not have a professional position.
c. The person has a high level of life insurance, given that he has a professional position.
d. Determine whether or not the events “has a high level of life insurance” and “has a professional position” are independent.

Q3.3.22

The sample space of equally likely outcomes for the experiment of rolling two fair dice is

$$\begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array} \tag{3.E.8}$$

Identify the events N: the sum is at least nine, T: at least one of the dice is a two, and F: at least one of the dice is a five.

a. Find $P(N)$.
b. Find $P(N \mid F)$.
c. Find $P(N \mid T)$.
d. Determine from the previous answers whether or not the events N and F are independent; whether or not N and T are.

Q3.3.23

The *sensitivity* of a drug test is the probability that the test will be positive when administered to a person who has actually taken the drug. Suppose that there are two independent tests to detect the presence of a certain type of banned drugs in athletes. One has sensitivity 0.75; the other has sensitivity 0.85. If both are applied to an athlete who has taken this type of drug, what is the chance that his usage will go undetected?

Q3.3.24

A man has two lights in his well house to keep the pipes from freezing in winter. He checks the lights daily. Each light has probability 0.002 of burning out before it is checked the next day (independently of the other light).

a. If the lights are wired in parallel one will continue to shine even if the other burns out. In this situation, compute the probability that at least one light will continue to shine for the full 24 hours. Note the greatly increased reliability of the system of two bulbs over that of a single bulb.
b. If the lights are wired in series neither one will continue to shine even if only one of them burns out. In this situation, compute the probability that at least one light will continue to shine for the full 24 hours. Note the slightly decreased reliability of the system of two bulbs over that of a single bulb.

Q3.3.25

An accountant has observed that 5% of all copies of a particular two-part form have an error in Part I, and 2% have an error in Part II. If the errors occur independently, find the probability that a randomly selected form will be error-free.

Q3.3.26

A box contains 20 screws which are identical in size, but 12 of which are zinc coated and 8 of which are not. Two screws are selected at random, without replacement.

a. Find the probability that both are zinc coated.
b. Find the probability that at least one is zinc coated.

Additional Exercises

Q3.3.27

Events A and B are mutually exclusive. Find $P(A \mid B)$.

Q3.3.28

The city council of a particular city is composed of five members of party A, four members of party B, and three independents. Two council members are randomly selected to form an investigative committee.

a. Find the probability that both are from party A.
b. Find the probability that at least one is an independent.
c. Find the probability that the two have different party affiliations (that is, not both A, not both B, and not both independent).

Q3.3.29

A basketball player makes 60% of the free throws that he attempts, except that if he has just tried and missed a free throw then his chances of making a second one go down to only 30%. Suppose he has just been awarded two free throws.

a. Find the probability that he makes both.
b. Find the probability that he makes at least one. (A tree diagram could help.)

Q3.3.30

An economist wishes to ascertain the proportion p of the population of individual taxpayers who have purposely submitted fraudulent information on an income tax return. To truly guarantee anonymity of the taxpayers in a random survey, taxpayers questioned are given the following instructions.

a. Flip a coin.
b. If the coin lands heads, answer "Yes" to the question "Have you ever submitted fraudulent information on a tax return?" even if you have not.
c. If the coin lands tails, give a truthful "Yes" or "No" answer to the question "Have you ever submitted fraudulent information on a tax return?"

The questioner is not told how the coin landed, so he does not know if a "Yes" answer is the truth or is given only because of the coin toss.

a. Using the Probability Rule for Complements and the independence of the coin toss and the taxpayers' status fill in the empty cells in the two-way contingency table shown. Assume that the coin is fair. Each cell except the two in the bottom row will contain the unknown proportion (or probability) p.

Status	Coin		Probability
	H	T	
Fraud			p
No fraud			
Probability			1

b. The only information that the economist sees are the entries in the following table:

$Response$	$"Yes"$	$"No"$
$Proportion$	r	s

(3.E.9)

Equate the entry in the one cell in the table in (a) that corresponds to the answer "No" to the number s to obtain the formula that expresses the unknown number p in terms of the known number s.
c. Equate the sum of the entries in the three cells in the table in (a) that together correspond to the answer "Yes" to the number r to obtain the formula that expresses the unknown number p in terms of the known number r.
d. Use the fact that $r + s = 1$(since they are the probabilities of complementary events) to verify that the formulas in (b) and (c) give the same value for p. (For example, insert $s = 1 - r$into the formula in (b) to obtain the formula in (c)).
e. Suppose a survey of $1,200$ taxpayers is conducted and 690 respond "Yes" (truthfully or not) to the question "Have you ever submitted fraudulent information on a tax return?" Use the answer to either (b) or (c) to estimate the true proportion p of all individual taxpayers who have purposely submitted fraudulent information on an income tax return.

Answers

1. a. 0.6
   b. 0.4
   c. not independent
2.
3. a. 0.22
   b. 0.81
   c. 0.27
4.
5. a. 0
   b. 0
6.
7. a. 0.5
   b. 0.4
   c. 0.6
8.
9. a. 0.25
   b. 0.33
   c. 0
   d. 0.25
10.
11. a. $P(A) = 0.3,\ P(R) = 0.4,\ P(A \cap R) = 0.12$
    b. independent
    c. without computation 0.3
12.
13. a. Insufficient information. The events *A* and *B* are not known to be either independent or mutually exclusive.
    b. 0.21
    c. 0
14.
15. a. 0.25
    b. 0.02
16.
17. a. 0.5
    b. 0.57
    c. 0.875
    d. 0.75
18.
19. a. 0.4
    b. 0.43
    c. 0.38
20.
21. a. 0.15
    b. 0.14
    c. 0.25
    d. not independent
22.
23. 0.0375
24.

25. 0.931
26.
27. 0
28.
29. a. 0.36
    b. 0.72

Contributor

- Anonymous

This page titled 3.E: Basic Concepts of Probability (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **3.E: Basic Concepts of Probability (Exercises)** has no license indicated.

CHAPTER OVERVIEW

4: Discrete Random Variables

It is often the case that a number is naturally associated to the outcome of a random experiment: the number of boys in a three-child family, the number of defective light bulbs in a case of 100 bulbs, the length of time until the next customer arrives at the drive-through window at a bank. Such a number varies from trial to trial of the corresponding experiment, and does so in a way that cannot be predicted with certainty; hence, it is called a random variable. In this chapter and the next we study such variables.

4.1: Random Variables

4.2: Probability Distributions for Discrete Random Variables

4.3: The Binomial Distribution

4.E: Discrete Random Variables (Exercises)

This page titled 4: Discrete Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Random Variables

Learning Objectives

- To learn the concept of a random variable.
- To learn the distinction between discrete and continuous random variables.

Definition: random variable

A random variable is a numerical quantity that is generated by a random experiment.

We will denote random variables by capital letters, such as X or Z, and the actual values that they can take by lowercase letters, such as x and z.

Table 4.1.1 gives four examples of random variables. In the second example, the three dots indicates that every counting number is a possible value for X. Although it is highly unlikely, for example, that it would take 50 tosses of the coin to observe heads for the first time, nevertheless it is conceivable, hence the number 50 is a possible value. The set of possible values is infinite, but is still at least countable, in the sense that all possible values can be listed one after another. In the last two examples, by way of contrast, the possible values cannot be individually listed, but take up a whole interval of numbers. In the fourth example, since the light bulb could conceivably continue to shine indefinitely, there is no natural greatest value for its lifetime, so we simply place the symbol ∞ for infinity as the right endpoint of the interval of possible values.

Table 4.1.1: Four Random Variables

Experiment	Number X	Possible Values of X
Roll two fair dice	Sum of the number of dots on the top faces	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Flip a fair coin repeatedly	Number of tosses until the coin lands heads	1, 2, 3,4, …
Measure the voltage at an electrical outlet	Voltage measured	$118 \le x \le 122$
Operate a light bulb until it burns out	Time until the bulb burns out	$0 \le x < \infty$

Definition: discrete random variable

A random variable is called discrete if it has either a finite or a countable number of possible values. A random variable is called continuous if its possible values contain a whole interval of numbers.

The examples in the table are typical in that discrete random variables typically arise from a counting process, whereas continuous random variables typically arise from a measurement.

Key Takeaway

- A random variable is a number generated by a random experiment.
- A random variable is called discrete if its possible values form a finite or countable set.
- A random variable is called continuous if its possible values contain a whole interval of numbers.

This page titled 4.1: Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

4.2: Probability Distributions for Discrete Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a discrete random variable.
- To learn the concepts of the mean, variance, and standard deviation of a discrete random variable, and how to compute them.

Associated to each possible value x of a discrete random variable X is the probability $P(x)$ that X will take the value x in one trial of the experiment.

Definition: probability distribution

The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial of the experiment.

The probabilities in the probability distribution of a random variable X must satisfy the following two conditions:

- Each probability $P(x)$ must be between 0 and 1:

$$0 \le P(x) \le 1.$$

- The sum of all the possible probabilities is 1:

$$\sum P(x) = 1.$$

Example 4.2.1: two Fair Coins

A fair coin is tossed twice. Let X be the number of heads that are observed.

a. Construct the probability distribution of X.
b. Find the probability that at least one head is observed.

Solution

a. The possible values that X can take are 0, 1, and 2. Each of these numbers corresponds to an event in the sample space $S = \{hh, ht, th, tt\}$ of equally likely outcomes for this experiment:

$$X = 0 \text{ to } \{tt\},\ X = 1 \text{ to } \{ht, th\},\ \text{and } X = 2 \text{ to } hh.$$

The probability of each of these events, hence of the corresponding value of X, can be found simply by counting, to give

x	0	1	2
$P(x)$	0.25	0.50	0.25

This table is the probability distribution of X.

b. "At least one head" is the event $X \ge 1$, which is the union of the mutually exclusive events $X = 1$ and $X = 2$. Thus

$$P(X \ge 1) = P(1) + P(2) = 0.50 + 0.25$$
$$= 0.75$$

A histogram that graphically illustrates the probability distribution is given in Figure 4.2.1.

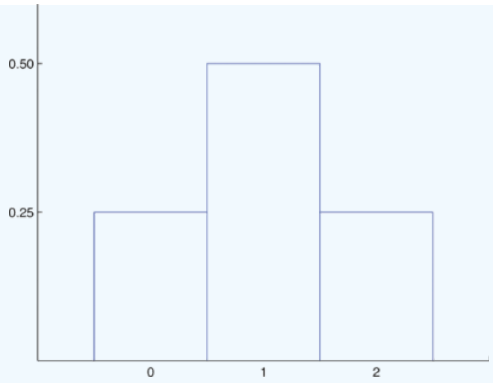


Figure 4.2.1: *Probability Distribution for Tossing a Fair Coin Twice*

Example 4.2.2: Two Fair Dice

A pair of fair dice is rolled. Let X denote the sum of the number of dots on the top faces.

a. Construct the probability distribution of X for a paid of fair dice.
b. Find $P(X \geq 9)$.
c. Find the probability that X takes an even value.

Solution

The sample space of equally likely outcomes is

$$\begin{array}{cccccc} 11 & 12 & 13 & 14 & 15 & 16 \\ 21 & 22 & 23 & 24 & 25 & 26 \\ 31 & 32 & 33 & 34 & 35 & 36 \\ 41 & 42 & 43 & 44 & 45 & 46 \\ 51 & 52 & 53 & 54 & 55 & 56 \\ 61 & 62 & 63 & 64 & 65 & 66 \end{array}$$

where the first digit is die 1 and the second number is die 2.

a. The possible values for X are the numbers 2 through 12. $X = 2$ is the event $\{11\}$, so $P(2) = 1/36$. $X = 3$ is the event $\{12, 21\}$, so $P(3) = 2/36$. Continuing this way we obtain the following table

x	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This table is the probability distribution of X.

b. The event $X \geq 9$ is the union of the mutually exclusive events $X = 9$, $X = 10$, $X = 11$, and $X = 12$. Thus

$$\begin{aligned} P(X \geq 9) &= P(9) + P(10) + P(11) + P(12) \\ &= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \\ &= \frac{10}{36} \\ &= 0.2\overline{7} \end{aligned}$$

c. Before we immediately jump to the conclusion that the probability that X takes an even value must be 0.5, note that X takes six different even values but only five different odd values. We compute

$$\begin{aligned} P(X \text{ is even}) &= P(2) + P(4) + P(6) + P(8) + P(10) + P(12) \\ &= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} \\ &= \frac{18}{36} \\ &= 0.5 \end{aligned}$$

A histogram that graphically illustrates the probability distribution is given in Figure 4.2.2.

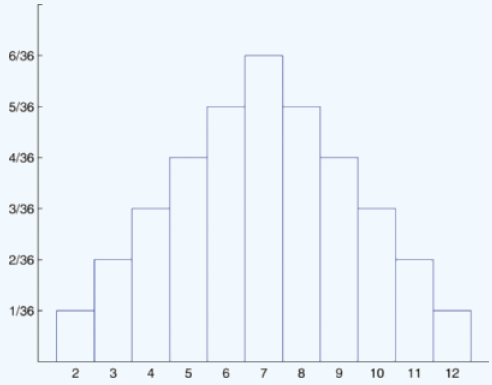


Figure 4.2.2: *Probability Distribution for Tossing Two Fair Dice*

The Mean and Standard Deviation of a Discrete Random Variable

Definition: mean

The *mean* (also called the "expectation value" or "expected value") of a discrete random variable X is the number

$$\mu = E(X) = \sum xP(x) \tag{4.2.1}$$

The mean of a random variable may be interpreted as the average of the values assumed by the random variable in repeated trials of the experiment.

Example 4.2.3

Find the mean of the discrete random variable X whose probability distribution is

x	-2	1	2	3.5
$P(x)$	0.21	0.34	0.24	0.21

Solution

Using the definition of mean (Equation 4.2.1) gives

$$\begin{aligned}\mu &= \sum xP(x) \\ &= (-2)(0.21) + (1)(0.34) + (2)(0.24) + (3.5)(0.21) \\ &= 1.135\end{aligned}$$

Example 4.2.4

A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for $\$1$ each. Each has an equal chance of winning. First prize is $\$300$, second prize is $\$200$, and third prize is $\$100$. Let X denote the net gain from the purchase of one ticket.

a. Construct the probability distribution of X.
b. Find the probability of winning any money in the purchase of one ticket.
c. Find the expected value of X, and interpret its meaning.

Solution

a. If a ticket is selected as the first prize winner, the net gain to the purchaser is the $\$300$ prize less the $\$1$ that was paid for the ticket, hence $X = 300 - 11 = 299$. There is one such ticket, so $P(299) = 0.001$. Applying the same "income minus outgo" principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:

x	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

x	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

b. Let W denote the event that a ticket is selected to win one of the prizes. Using the table

$$\begin{aligned} P(W) &= P(299) + P(199) + P(99) = 0.001 + 0.001 + 0.001 \\ &= 0.003 \end{aligned}$$

c. Using the definition of expected value (Equation 4.2.1),

$$\begin{aligned} E(X) &= (299) \cdot (0.001) + (199) \cdot (0.001) + (99) \cdot (0.001) + (-1) \cdot (0.997) \\ &= -0.4 \end{aligned}$$

The negative value means that one loses money on the average. In particular, if someone were to buy tickets repeatedly, then although he would win now and then, on average he would lose 40 cents per ticket purchased.

The concept of expected value is also basic to the insurance industry, as the following simplified example illustrates.

Example 4.2.5

A life insurance company will sell a $\$200,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$195$. Find the expected value to the company of a single policy if a person in this risk group has a 99.97% chance of surviving one year.

Solution

Let X denote the net gain to the company from the sale of one such policy. There are two possibilities: the insured person lives the whole year or the insured person dies before the year is up. Applying the "income minus outgo" principle, in the former case the value of X is $195 - 0$; in the latter case it is $195 - 200,000 = -199,805$. Since the probability in the first case is 0.9997 and in the second case is $1 - 0.9997 = 0.0003$, the probability distribution for X is:

x	195	$-199,805$
$P(x)$	0.9997	0.0003

Therefore

$$\begin{aligned} E(X) &= \sum xP(x) \\ &= (195) \cdot (0.9997) + (-199,805) \cdot (0.0003) \\ &= 135 \end{aligned}$$

Occasionally (in fact, 3 times in $10,000$) the company loses a large amount of money on a policy, but typically it gains $\$195$, which by our computation of $E(X)$ works out to a net gain of $\$135$ per policy sold, on average.

Definition: variance

The *variance* (σ^2) of a discrete random variable X is the number

$$\sigma^2 = \sum (x - \mu)^2 P(x) \tag{4.2.2}$$

which by algebra is equivalent to the formula

$$\sigma^2 = \left[\sum x^2 P(x)\right] - \mu^2 \tag{4.2.3}$$

Definition: standard deviation

The standard deviation, σ, *of a discrete random variable X is the square root of its variance, hence is given by the formulas*

$$\sigma = \sqrt{\sum (x - \mu)^2 P(x)} = \sqrt{\left[\sum x^2 P(x)\right] - \mu^2} \tag{4.2.4}$$

The variance and standard deviation of a discrete random variable X may be interpreted as measures of the variability of the values assumed by the random variable in repeated trials of the experiment. The units on the standard deviation match those of X.

Example 4.2.6

A discrete random variable X has the following probability distribution:

x	-1	0	1	4
$P(x)$	0.2	0.5	a	0.1

(4.2.5)

A histogram that graphically illustrates the probability distribution is given in Figure 4.2.3.

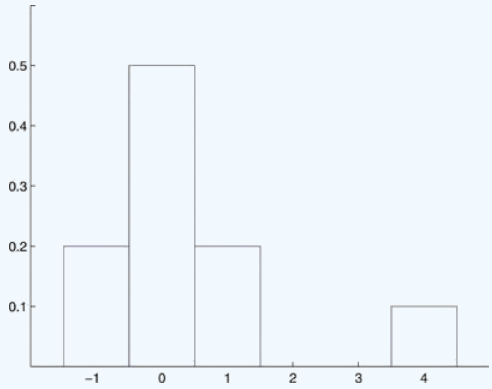


Figure 4.2.3: Probability Distribution of a Discrete Random Variable

Compute each of the following quantities.

a. a.
b. $P(0)$.
c. $P(X > 0)$.
d. $P(X \geq 0)$.
e. $P(X \leq -2)$.
f. The mean μ of X.
g. The variance σ^2 of X.
h. The standard deviation σ of X.

Solution

a. Since all probabilities must add up to 1,

$$a = 1 - (0.2 + 0.5 + 0.1) = 0.2$$

b. Directly from the table, P(0)=0.5

$$P(0) = 0.5$$

c. From Table 4.2.5,

$$P(X > 0) = P(1) + P(4) = 0.2 + 0.1 = 0.3$$

d. From Table 4.2.5,

$$P(X \geq 0) = P(0) + P(1) + P(4) = 0.5 + 0.2 + 0.1 = 0.8$$

e. Since none of the numbers listed as possible values for X is less than or equal to -2, the event $X \leq -2$ is impossible, so

$$P(X \leq -2) = 0$$

f. Using the formula in the definition of μ (Equation 4.2.1)

$$\begin{aligned}\mu &= \sum xP(x) \\ &= (-1)\cdot(0.2) + (0)\cdot(0.5) + (1)\cdot(0.2) + (4)\cdot(0.1) \\ &= 0.4\end{aligned}$$

g. Using the formula in the definition of σ^2 (Equation 4.2.2) and the value of μ that was just computed,

$$\begin{aligned}\sigma^2 &= \sum (x-\mu)^2 P(x) \\ &= (-1-0.4)^2\cdot(0.2) + (0-0.4)^2\cdot(0.5) + (1-0.4)^2\cdot(0.2) + (4-0.4)^2\cdot(0.1) \\ &= 1.84\end{aligned}$$

h. Using the result of part (g), $\sigma = \sqrt{1.84} = 1.3565$

Summary

- The probability distribution of a discrete random variable X is a listing of each possible value x taken by X along with the probability $P(x)$ that X takes that value in one trial of the experiment.
- The mean μ of a discrete random variable X is a number that indicates the average value of X over numerous trials of the experiment. It is computed using the formula $\mu = \sum xP(x)$.
- The variance σ^2 and standard deviation σ of a discrete random variable X are numbers that indicate the variability of X over numerous trials of the experiment. They may be computed using the formula $\sigma^2 = [\sum x^2 P(x)] - \mu^2$.

This page titled 4.2: Probability Distributions for Discrete Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

4.3: The Binomial Distribution

Learning Objectives

- To learn the concept of a binomial random variable.
- To learn how to recognize a random variable as being a binomial random variable.

The experiment of tossing a fair coin three times and the experiment of observing the genders according to birth order of the children in a randomly selected three-child family are completely different, but the random variables that count the number of heads in the coin toss and the number of boys in the family (assuming the two genders are equally likely) are the same random variable, the one with probability distribution

x	0	1	2	3
$P(x)$	0.125	0.375	0.375	0.125

A histogram that graphically illustrates this probability distribution is given in Figure 4.3.1. What is common to the two experiments is that we perform three identical and independent trials of the same action, each trial has only two outcomes (heads or tails, boy or girl), and the probability of success is the same number, 0.5, on every trial. The random variable that is generated is called the binomial random variable with parameters $n = 3$ and $p = 0.5$. This is just one case of a general situation.

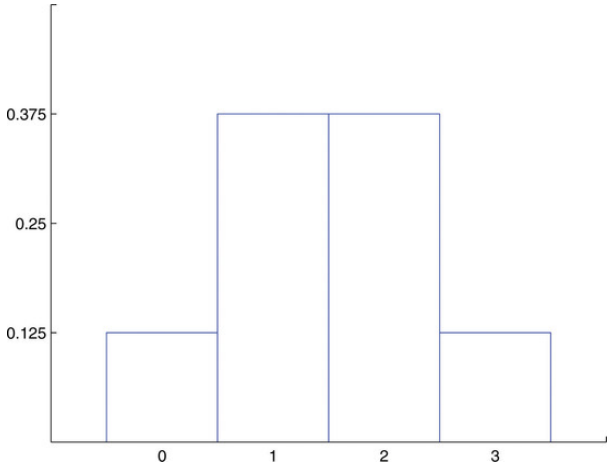


Figure 4.3.1: Probability Distribution for Three Coins and Three Children

Definition: binomial distribution

Suppose a random experiment has the following characteristics.

- There are n identical and independent trials of a common procedure.
- There are exactly two possible outcomes for each trial, one termed "success" and the other "failure."
- The probability of success on any one trial is the same number p.

Then the discrete random variable X that counts the number of successes in the n trials is the binomial random variable with parameters n and p. We also say that X has a binomial distribution with parameters n and p.

The following four examples illustrate the definition. Note how in every case "success" is the outcome that is counted, not the outcome that we prefer or think is better in some sense.

1. A random sample of 125 students is selected from a large college in which the proportion of students who are females is 57%. Suppose X denotes the number of female students in the sample. In this situation there are $n = 125$ identical and independent trials of a common procedure, selecting a student at random; there are exactly two possible outcomes for each trial, "success" (what we are counting, that the student be female) and "failure;" and finally the probability of success on any one trial is the same number $p = 0.57$. X is a binomial random variable with parameters $n = 125$ and $p = 0.57$.
2. A multiple-choice test has 15 questions, each of which has five choices. An unprepared student taking the test answers each of the questions completely randomly by choosing an arbitrary answer from the five provided. Suppose X denotes the number of

answers that the student gets right. X is a binomial random variable with parameters $n = 15$ and $p = 1/5 = 0.20$.

3. In a survey of $1,000$ registered voters each voter is asked if he intends to vote for a candidate Titania Queen in the upcoming election. Suppose X denotes the number of voters in the survey who intend to vote for Titania Queen. X is a binomial random variable with $n = 1000$ and p equal to the true proportion of voters (surveyed or not) who intend to vote for Titania Queen.
4. An experimental medication was given to 30 patients with a certain medical condition. Suppose X denotes the number of patients who develop severe side effects. X is a binomial random variable with $n = 30$ and p equal to the true probability that a patient with the underlying condition will experience severe side effects if given that medication.

Probability Formula for a Binomial Random Variable

Often the most difficult aspect of working a problem that involves the binomial random variable is recognizing that the random variable in question has a binomial distribution. Once that is known, probabilities can be computed using the following formula.

If X is a binomial random variable with parameters n and p, then

$$P(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

where $q = 1 - p$ and where for any counting number m, $m!$ (read "m factorial") is defined by

$$0! = 1, 1! = 1, 2! = 1 \cdot 2, 3! = 1 \cdot 2 \cdot 3$$

and in general

$$m! = 1 \cdot 2 \cdot \cdots \cdot (m-1) \cdot m$$

Example 4.3.1

Seventeen percent of victims of financial fraud know the perpetrator of the fraud personally.

a. Use the formula to construct the probability distribution for the number X of people in a random sample of five victims of financial fraud who knew the perpetrator personally.
b. A investigator examines five cases of financial fraud every day. Find the most frequent number of cases each day in which the victim knew the perpetrator.
c. A investigator examines five cases of financial fraud every day. Find the average number of cases per day in which the victim knew the perpetrator.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$; $q = 1 - p = 0.83$. The possible values of X are $0, 1, 2, 3, 4,$ and 5.

$$\begin{aligned} P(0) &= \frac{5!}{0!5!}(0.17)^0(0.83)^5 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)} 1 \cdot (0.3939040643) \\ &= 0.3939040643 \approx 0.3939 \end{aligned}$$

$$\begin{aligned} P(1) &= \frac{5!}{1!4!}(0.17)^1(0.83)^4 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4)}(0.17) \cdot (0.47458321) \\ &= 5 \cdot (0.17) \cdot (0.47458321) \\ &= 0.4033957285 \approx 0.4034 \end{aligned}$$

$$\begin{aligned} P(2) &= \frac{5!}{2!3!}(0.17)^2(0.83)^3 \\ &= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1 \cdot 2)(1 \cdot 2 \cdot 3)}(0.0289) \cdot (0.571787) \\ &= 10 \cdot (0.0289) \cdot (0.571787) \\ &= 0.165246443 \approx 0.1652 \end{aligned}$$

The remaining three probabilities are computed similarly, to give the probability distribution

x	0	1	2	3	4	5
$P(x)$	0.3939	0.4034	0.1652	0.0338	0.0035	0.0001

The probabilities do not add up to exactly 1 because of rounding.

This probability distribution is represented by the histogram in Figure 4.3.2, which graphically illustrates just how improbable the events $X = 4$ and $X = 5$ are. The corresponding bar in the histogram above the number 4 is barely visible, if visible at all, and the bar above 5 is far too short to be visible.

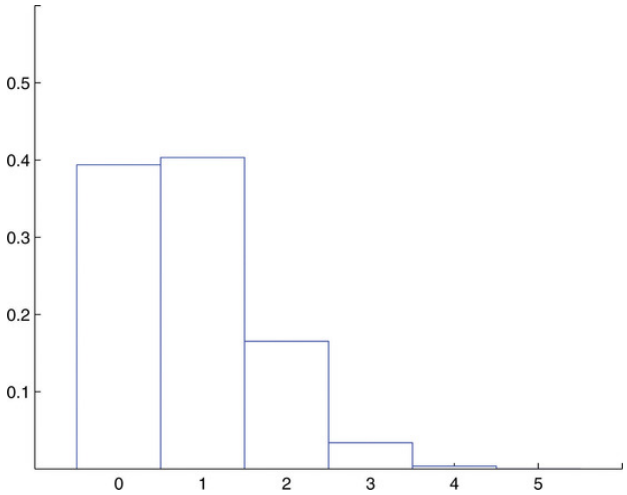


Figure 4.3.2: Probability Distribution of the Binomial Random Variable in Example 4.3.1

The value of X that is most likely is $X = 1$, so the most frequent number of cases seen each day in which the victim knew the perpetrator is one.

The average number of cases per day in which the victim knew the perpetrator is the mean of X, which is

$$\mu = \sum xP(x) \tag{4.3.1}$$
$$= 0 \cdot 0.3939 + 1 \cdot 0.4034 + 2 \cdot 0.1652 + 3 \cdot 0.0338 + 4 \cdot 0.0035 + 5 \cdot 0.0001 \tag{4.3.2}$$
$$= 0.8497 \tag{4.3.3}$$

Special Formulas for the Mean and Standard Deviation of a Binomial Random Variable

Since a binomial random variable is a discrete random variable, the formulas for its mean, variance, and standard deviation given in the previous section apply to it, as we just saw in Example 4.3.2 in the case of the mean. However, for the binomial random variable there are much simpler formulas.

If X is a binomial random variable with parameters n and p, then

$$\mu = np$$

$$\sigma^2 = npq$$

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

where $q = 1 - p$.

Example 4.3.2

Find the mean and standard deviation of the random variable X of Example 4.3.1.

Solution

The random variable X is binomial with parameters $n = 5$ and $p = 0.17$, and $q = 1 - p = 0.83$. Thus its mean and standard deviation are

$$\mu = np = (5) \cdot (0.17) = 0.85 \text{ (exactly)}$$

and

$$\sigma = \sqrt{npq} = \sqrt{(5) \cdot (0.17) \cdot (0.83)} = \sqrt{0.7055} \approx 0.8399$$

The Cumulative Probability Distribution of a Binomial Random Variable

In order to allow a broader range of more realistic problems contains probability tables for binomial random variables for various choices of the parameters n and p. These tables are not the probability distributions that we have seen so far, but are **cumulative probability distributions**. In the place of the probability $P(x)$ the table contains the probability

$$P(X \leq x) = P(0) + P(1) + ... + P(x)$$

This is illustrated in Figure 4.3.3. The probability entered in the table corresponds to the area of the shaded region. The reason for providing a cumulative table is that in practical problems that involve a binomial random variable typically the probability that is sought is of the form $P(X \leq x)$ or $P(X \geq x)$. The cumulative table is much easier to use for computing $P(X \leq x)$ since all the individual probabilities have already been computed and added. The one table suffices for both $P(X \leq x)$ or $P(X \geq x)$ and can be used to readily obtain probabilities of the form $P(x)$, too, because of the following formulas. The first is just the Probability Rule for Complements.

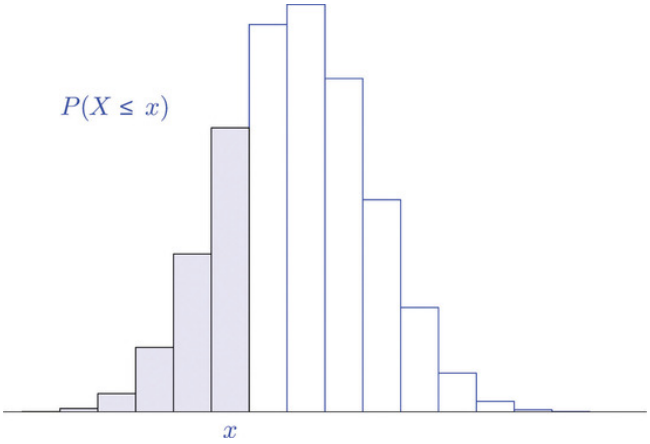


Figure 4.3.3: Cumulative Probabilities

If X is a discrete random variable, then

$$P(X \geq x) = 1 - P(X \leq x - 1)$$

and

$$P(x) = P(X \leq x) - P(X \leq x - 1)$$

Example 4.3.3

A student takes a ten-question true/false exam.

a. Find the probability that the student gets exactly six of the questions right simply by guessing the answer on every question.
b. Find the probability that the student will obtain a passing grade of 60% or greater simply by guessing.

Solution

Let X denote the number of questions that the student guesses correctly. Then X is a binomial random variable with parameters $n = 10$ and $p = 0.50$.

a. The probability sought is $P(6)$. The formula gives

$$P(6) = 10!(6!)(4!)(.5)6.54 = 0.205078125$$

Using the table,

$$P(6) = P(X \leq 6) - P(X \leq 5) = 0.8281 - 0.6230 = 0.2051$$

b. The student must guess correctly on at least 60% of the questions, which is $(0.60) \cdot (10) = 6$ questions. The probability sought is not $P(6)$ (an easy mistake to make), but

$$P(X \geq 6) = P(6) + P(7) + P(8) + P(9) + P(10)$$

Instead of computing each of these five numbers using the formula and adding them we can use the table to obtain

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6230 = 0.3770$$

which is much less work and of sufficient accuracy for the situation at hand.

Example 4.3.4

An appliance repairman services five washing machines on site each day. One-third of the service calls require installation of a particular part.

a. The repairman has only one such part on his truck today. Find the probability that the one part will be enough today, that is, that at most one washing machine he services will require installation of this particular part.
b. Find the minimum number of such parts he should take with him each day in order that the probability that he have enough for the day's service calls is at least 95%.

Solution

Let X denote the number of service calls today on which the part is required. Then X is a binomial random variable with parameters $n = 5$ and $p = 1/3 = 0.\overline{3}$

a. Note that the probability in question is not $P(1)$, but rather $P(X \leq 1)$. Using the cumulative distribution table,

$$P(X \leq 1) = 0.4609$$

b. The answer is the smallest number x such that the table entry $P(X \leq x)$ is at least 0.9500. Since $P(X \leq 2) = 0.7901$ is less than 0.95, two parts are not enough. Since $P(X \leq 3) = 0.9547$ is as large as 0.95, three parts will suffice at least 95% of the time. Thus the minimum needed is three.

Summary

- The discrete random variable X that counts the number of successes in n identical, independent trials of a procedure that always results in either of two outcomes, “success” or “failure,” and in which the probability of success on each trial is the same number p, is called the binomial random variable with parameters n and p.
- There is a formula for the probability that the binomial random variable with parameters n and p will take a particular value x.
- There are special formulas for the mean, variance, and standard deviation of the binomial random variable with parameters n and p that are much simpler than the general formulas that apply to all discrete random variables.
- Cumulative probability distribution tables, when available, facilitate computation of probabilities encountered in typical practical situations.

This page titled 4.3: The Binomial Distribution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

4.E: Discrete Random Variables (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

4.1: Random Variables

Basic

1. Classify each random variable as either discrete or continuous.
   a. The number of arrivals at an emergency room between midnight and $6:00\ a.m.$
   b. The weight of a box of cereal labeled "18 ounces."
   c. The duration of the next outgoing telephone call from a business office.
   d. The number of kernels of popcorn in a 1-pound container.
   e. The number of applicants for a job.
2. Classify each random variable as either discrete or continuous.
   a. The time between customers entering a checkout lane at a retail store.
   b. The weight of refuse on a truck arriving at a landfill.
   c. The number of passengers in a passenger vehicle on a highway at rush hour.
   d. The number of clerical errors on a medical chart.
   e. The number of accident-free days in one month at a factory.
3. Classify each random variable as either discrete or continuous.
   a. The number of boys in a randomly selected three-child family.
   b. The temperature of a cup of coffee served at a restaurant.
   c. The number of no-shows for every 100 reservations made with a commercial airline.
   d. The number of vehicles owned by a randomly selected household.
   e. The average amount spent on electricity each July by a randomly selected household in a certain state.
4. Classify each random variable as either discrete or continuous.
   a. The number of patrons arriving at a restaurant between $5:00\ p.m.$ and $6:00\ p.m.$
   b. The number of new cases of influenza in a particular county in a coming month.
   c. The air pressure of a tire on an automobile.
   d. The amount of rain recorded at an airport one day.
   e. The number of students who actually register for classes at a university next semester.
5. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of heads in two tosses of a coin.
   b. The average weight of newborn babies born in a particular county one month.
   c. The amount of liquid in a 12-ounce can of soft drink.
   d. The number of games in the next World Series (best of up to seven games).
   e. The number of coins that match when three coins are tossed at once.
6. Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)
   a. The number of hearts in a five-card hand drawn from a deck of 52 cards that contains 13 hearts in all.
   b. The number of pitches made by a starting pitcher in a major league baseball game.
   c. The number of breakdowns of city buses in a large city in one week.
   d. The distance a rental car rented on a daily rate is driven each day.
   e. The amount of rainfall at an airport next month.

Answers

1. a. discrete
   b. continuous
   c. continuous
   d. discrete

e. discrete

2.

3. a. discrete
   b. continuous
   c. discrete
   d. discrete
   e. continuous

4.

5. a. $\{0.1.2\}$
   b. an interval (a, b) (answers vary)
   c. an interval (a, b) (answers vary)
   d. $\{4, 5, 6, 7\}$
   e. $\{2, 3\}$

4.2: Probability Distributioins for Discrete Random Variables

Basic

1. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.
   $$\begin{array}{c|cccc} x & -2 & 0 & 2 & 4 \\ \hline P(x) & 0.3 & 0.5 & 0.2 & 0.1 \end{array} \tag{4.E.1}$$

   b.
   $$\begin{array}{c|ccc} x & 0.5 & 0.25 & 0.25 \\ \hline P(x) & -0.4 & 0.6 & 0.8 \end{array} \tag{4.E.2}$$

   c.
   $$\begin{array}{c|ccccc} x & 1.1 & 2.5 & 4.1 & 4.6 & 5.3 \\ \hline P(x) & 0.16 & 0.14 & 0.11 & 0.27 & 0.22 \end{array} \tag{4.E.3}$$

2. Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

   a.
   $$\begin{array}{c|ccccc} x & 0 & 1 & 2 & 3 & 4 \\ \hline P(x) & -0.25 & 0.50 & 0.35 & 0.10 & 0.30 \end{array} \tag{4.E.4}$$

   b.
   $$\begin{array}{c|ccc} x & 1 & 2 & 3 \\ \hline P(x) & 0.325 & 0.406 & 0.164 \end{array} \tag{4.E.5}$$

   c.
   $$\begin{array}{c|ccccc} x & 25 & 26 & 27 & 28 & 29 \\ \hline P(x) & 0.13 & 0.27 & 0.28 & 0.18 & 0.14 \end{array} \tag{4.E.6}$$

3. A discrete random variable X has the following probability distribution:

   $$\begin{array}{c|ccccc} x & 77 & 78 & 79 & 80 & 81 \\ \hline P(x) & 0.15 & 0.15 & 0.20 & 0.40 & 0.10 \end{array} \tag{4.E.7}$$

   Compute each of the following quantities.

   a. $P(80)$.
   b. $P(X > 80)$.
   c. $P(X \leq 80)$.
   d. The mean μ of X.
   e. The variance σ^2 of X.
   f. The standard deviation σ of X.

4. A discrete random variable X has the following probability distribution:

   $$\begin{array}{c|ccccc} x & 13 & 18 & 20 & 24 & 27 \\ \hline P(x) & 0.22 & 0.25 & 0.20 & 0.17 & 0.16 \end{array} \tag{4.E.8}$$

   Compute each of the following quantities.

a. $P(18)$.
b. $P(X > 18)$.
c. $P(X \leq 18)$.
d. The mean μ of X.
e. The variance σ^2 of X.
f. The standard deviation σ of X.

5. If each die in a pair is "loaded" so that one comes up half as often as it should, six comes up half again as often as it should, and the probabilities of the other faces are unaltered, then the probability distribution for the sum X of the number of dots on the top faces when the two are rolled is

x	2	3	4	5	6	7
$P(x)$	$\frac{1}{144}$	$\frac{4}{144}$	$\frac{8}{144}$	$\frac{12}{144}$	$\frac{16}{144}$	$\frac{22}{144}$

(4.E.9)

x	8	9	10	11	12
$P(x)$	$\frac{24}{144}$	$\frac{20}{144}$	$\frac{16}{144}$	$\frac{12}{144}$	$\frac{9}{144}$

(4.E.10)

Compute each of the following.

a. $P(5 \leq X \leq 9)$.
b. $P(X \geq 7)$.
c. The mean μ of X. (For fair dice this number is 7).
d. The standard deviation σ of X. (For fair dice this number is about 2.415).

Applications

6. Borachio works in an automotive tire factory. The number X of sound but blemished tires that he produces on a random day has the probability distribution

x	2	3	4	5
$P(x)$	0.48	0.36	0.12	0.04

(4.E.11)

a. Find the probability that Borachio will produce more than three blemished tires tomorrow.
b. Find the probability that Borachio will produce at most two blemished tires tomorrow.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

7. In a hamster breeder's experience the number X of live pups in a litter of a female not over twelve months in age who has not borne a litter in the past six weeks has the probability distribution

x	3	4	5	6	7	8	9
$P(x)$	0.04	0.10	0.26	0.31	0.22	0.05	0.02

(4.E.12)

a. Find the probability that the next litter will produce five to seven live pups.
b. Find the probability that the next litter will produce at least six live pups.
c. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

8. The number X of days in the summer months that a construction crew cannot work because of the weather has the probability distribution

x	6	7	8	9	10
$P(x)$	0.03	0.08	0.15	0.20	0.19

(4.E.13)

x	11	12	13	14
$P(x)$	0.16	0.10	0.07	0.02

(4.E.14)

a. Find the probability that no more than ten days will be lost next summer.
b. Find the probability that from 8 to 12 days will be lost next summer.
c. Find the probability that no days at all will be lost next summer.
d. Compute the mean and standard deviation of X. Interpret the mean in the context of the problem.

9. Let X denote the number of boys in a randomly selected three-child family. Assuming that boys and girls are equally likely, construct the probability distribution of X.

10. Let X denote the number of times a fair coin lands heads in three tosses. Construct the probability distribution of X.
11. Five thousand lottery tickets are sold for $\$1$ each. One ticket will win $\$1,000$, two tickets will win $\$500$ each, and ten tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
12. Seven thousand lottery tickets are sold for $\$5$ each. One ticket will win $\$2,000$, two tickets will win $\$750$ each, and five tickets will win $\$100$ each. Let X denote the net gain from the purchase of a randomly selected ticket.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X. Interpret its meaning.
    c. Compute the standard deviation σ of X.
13. An insurance company will sell a $\$90,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$478$. Find the expected value to the company of a single policy if a person in this risk group has a 99.62% chance of surviving one year.
14. An insurance company will sell a $\$10,000$ one-year term life insurance policy to an individual in a particular risk group for a premium of $\$368$. Find the expected value to the company of a single policy if a person in this risk group has a 97.25% chance of surviving one year.
15. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.9825. Such a person wishes to buy a $\$150,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$250$ per policy on all such policies.
16. An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.99. Such a person wishes to buy a $\$75,000$ one-year term life insurance policy. Let C denote how much the insurance company charges such a person for such a policy.
    a. Construct the probability distribution of X. (Two entries in the table will contain C).
    b. Compute the expected value $E(X)$ of X.
    c. Determine the value C must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).
    d. Determine the value C must have in order for the company to average a net gain of $\$150$ per policy on all such policies.
17. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; half of them are red and half are black. The remaining two slots are numbered 0 and 00 and are green. In a $\$1$ bet on red, the bettor pays $\$1$ to play. If the ball lands in a red slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on red he loses his dollar. Let X denote the net gain to the bettor on one play of the game.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X, and interpret its meaning in the context of the problem.
    c. Compute the standard deviation of X.
18. A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; the remaining two slots are numbered 0 and 00. Suppose the "number" 00 is considered not to be even, but the number 0 is still even. In a $\$1$ bet on even, the bettor pays $\$1$ to play. If the ball lands in an even numbered slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on an even numbered slot, he loses his dollar. Let X denote the net gain to the bettor on one play of the game.
    a. Construct the probability distribution of X.
    b. Compute the expected value $E(X)$ of X, and explain why this game is not offered in a casino (where 0 is not considered even).
    c. Compute the standard deviation of X.
19. The time, to the nearest whole minute, that a city bus takes to go from one end of its route to the other has the probability distribution shown. As sometimes happens with probabilities computed as empirical relative frequencies, probabilities in the

table add up only to a value other than 1.00 because of round-off error.

x	42	43	44	45	46	47
$P(x)$	0.10	0.23	0.34	0.25	0.05	0.02

(4.E.15)

a. Find the average time the bus takes to drive the length of its route.
b. Find the standard deviation of the length of time the bus takes to drive the length of its route.

20. Tybalt receives in the mail an offer to enter a national sweepstakes. The prizes and chances of winning are listed in the offer as: $\$5$ million, one chance in 65 million; $\$150,000$, one chance in 6.5 million; $\$5,000$, one chance in $650,000$; and $\$1,000$, one chance in $65,000$. If it costs Tybalt 44 cents to mail his entry, what is the expected value of the sweepstakes to him?

Additional Exercises

21. The number X of nails in a randomly selected 1-pound box has the probability distribution shown. Find the average number of nails per pound.

x	100	101	102
$P(x)$	0.01	0.96	0.03

(4.E.16)

22. Three fair dice are rolled at once. Let X denote the number of dice that land with the same number of dots on top as at least one other die. The probability distribution for X is

x	0	u	3
$P(x)$	p	$\frac{15}{36}$	$\frac{1}{36}$

(4.E.17)

a. Find the missing value u of X.
b. Find the missing probability p.
c. Compute the mean of X.
d. Compute the standard deviation of X.

23. Two fair dice are rolled at once. Let X denote the difference in the number of dots that appear on the top faces of the two dice. Thus for example if a one and a five are rolled, $X = 4$, and if two sixes are rolled, $X = 0$.

a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

24. A fair coin is tossed repeatedly until either it lands heads or a total of five tosses have been made, whichever comes first. Let X denote the number of tosses made.

a. Construct the probability distribution for X.
b. Compute the mean μ of X.
c. Compute the standard deviation σ of X.

25. A manufacturer receives a certain component from a supplier in shipments of 100 units. Two units in each shipment are selected at random and tested. If either one of the units is defective the shipment is rejected. Suppose a shipment has 5 defective units.

a. Construct the probability distribution for the number X of defective units in such a sample. (A tree diagram is helpful).
b. Find the probability that such a shipment will be accepted.

26. Shylock enters a local branch bank at $4:30\ p.m.$ every payday, at which time there are always two tellers on duty. The number X of customers in the bank who are either at a teller window or are waiting in a single line for the next available teller has the following probability distribution.

x	0	1	2	3
$P(x)$	0.135	0.192	0.284	0.230

(4.E.18)

x	4	5	6
$P(x)$	0.103	0.051	0.005

(4.E.19)

a. What number of customers does Shylock most often see in the bank the moment he enters?
b. What number of customers waiting in line does Shylock most often see the moment he enters?
c. What is the average number of customers who are waiting in line the moment Shylock enters?

27. The owner of a proposed outdoor theater must decide whether to include a cover that will allow shows to be performed in all weather conditions. Based on projected audience sizes and weather conditions, the probability distribution for the revenue X per night if the cover is not installed is

$Weather$	x	$P(x)$
$Clear$	$\$3,000$	0.61
$Threatening$	$\$2,800$	0.17
$LightRain$	$\$1,975$	0.11
$Show-cancelling\ rain$	$\$0$	0.11

(4.E.20)

The additional cost of the cover is $410,000. The owner will have it built if this cost can be recovered from the increased revenue the cover affords in the first ten 90-night seasons.

a. Compute the mean revenue per night if the cover is not installed.
b. Use the answer to (a) to compute the projected total revenue per 90-night season if the cover is not installed.
c. Compute the projected total revenue per season when the cover is in place. To do so assume that if the cover were in place the revenue each night of the season would be the same as the revenue on a clear night.
d. Using the answers to (b) and (c), decide whether or not the additional cost of the installation of the cover will be recovered from the increased revenue over the first ten years. Will the owner have the cover installed?

Answers

1. a. no: the sum of the probabilities exceeds 1
   b. no: a negative probability
   c. no: the sum of the probabilities is less than 1
2.
3. a. 0.4
   b. 0.1
   c. 0.9
   d. 79.15
   e. $\sigma^2 = 1.5275$
   f. $\sigma = 1.2359$
4.
5. a. 0.6528
   b. 0.7153
   c. $\mu = 7.8333$
   d. $\sigma^2 = 5.4866$
   e. $\sigma = 2.3424$
6.
7. a. 0.79
   b. 0.60
   c. $\mu = 5.8$, $\sigma = 1.2570$
8.
9.

x	0	1	2	3
$P(x)$	1/8	3/8	3/8	1/8

(4.E.21)

10.
11. a.

x	-1	999	499	99
$P(x)$	$\frac{4987}{5000}$	$\frac{1}{5000}$	$\frac{2}{5000}$	$\frac{10}{5000}$

(4.E.22)

    b. -0.4
    c. 17.8785
12.
13. 136

14.

15. a.

x	C	C	$-150,000$
$P(x)$	0.9825		0.0175

(4.E.23)

b. $C - 2625$
c. $C \geq 2625$
d. $C \geq 2875$

16.

17. a.

x	-1	1
$P(x)$	$\frac{20}{38}$	$\frac{18}{38}$

(4.E.24)

b. $E(X) = -0.0526$. In many bets the bettor sustains an average loss of about 5.25 cents per bet.
c. 0.9986

18.

19. a. 43.54
b. 1.2046

20.

21. 101.02

22.

23. a.

x	0	1	2	3	4	5
$P(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

(4.E.25)

b. 1.9444
c. 1.4326

24.

25. a.

x	0	1	2
$P(x)$	0.902	0.096	0.002

(4.E.26)

b. 0.902

26.

27. a. 2523.25
b. $227,092.5$
c. $270,000$
d. The owner will install the cover.

4.3: The Binomial Distribution

Basic

1. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of dots on the top face of fair die that is rolled.
   b. X is the number of hearts in a five-card hand drawn (without replacement) from a well-shuffled ordinary deck.
   c. X is the number of defective parts in a sample of ten randomly selected parts coming from a manufacturing process in which 0.02% of all parts are defective.
   d. X is the number of times the number of dots on the top face of a fair die is even in six rolls of the die.
   e. X is the number of dice that show an even number of dots on the top face when six dice are rolled at once.
2. Determine whether or not the random variable X is a binomial random variable. If so, give the values of n and p. If not, explain why not.
   a. X is the number of black marbles in a sample of 5 marbles drawn randomly and without replacement from a box that contains 25 white marbles and 15 black marbles.

b. X is the number of black marbles in a sample of 5 marbles drawn randomly and with replacement from a box that contains 25 white marbles and 15 black marbles.
c. X is the number of voters in favor of proposed law in a sample $1,200$ randomly selected voters drawn from the entire electorate of a country in which 35% of the voters favor the law.
d. X is the number of fish of a particular species, among the next ten landed by a commercial fishing boat, that are more than 13 inches in length, when 17% of all such fish exceed 13 inches in length.
e. X is the number of coins that match at least one other coin when four coins are tossed at once.

3. X is a binomial random variable with parameters $n = 12$ and $p = 0.82$. Compute the probability indicated.
   a. $P(11)$
   b. $P(9)$
   c. $P(0)$
   d. $P(13)$
4. X is a binomial random variable with parameters $n = 16$ and $p = 0.74$. Compute the probability indicated.
   a. $P(14)$
   b. $P(4)$
   c. $P(0)$
   d. $P(20)$
5. X is a binomial random variable with parameters $n = 5$, $p = 0.5$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 3)$
   b. $P(X \geq 3)$
   c. $P(3)$
   d. $P(0)$
   e. $P(5)$
6. X is a binomial random variable with parameters $n = 5$, $p = 0.\overline{3}$. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $P(X \leq 2)$
   b. $P(X \geq 2)$
   c. $P(2)$
   d. $P(0)$
   e. $P(5)$
7. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $n = 10, p = 0.25, P(X \leq 6)$
   b. $n = 10, p = 0.75, P(X \leq 6)$
   c. $n = 15, p = 0.75, P(X \leq 6)$
   d. $n = 15, p = 0.75, P(12)$
   e. $n = 15, p = 0.\overline{6}, P(10 \leq X \leq 12)$
8. X is a binomial random variable with the parameters shown. Use the tables in 7.1: Large Sample Estimation of a Population Mean to compute the probability indicated.
   a. $n = 5, p = 0.05, P(X \leq 1)$
   b. $n = 5, p = 0.5, P(X \leq 1)$
   c. $n = 10, p = 0.75, P(X \leq 5)$
   d. $n = 10, p = 0.75, P(12)$
   e. $n = 10, p = 0.\overline{6}, P(5 \leq X \leq 8)$
9. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
   a. $n = 8, p = 0.43$
   b. $n = 47, p = 0.82$
   c. $n = 1200, p = 0.44$

d. $n = 2100, p = 0.62$

10. X is a binomial random variable with the parameters shown. Use the special formulas to compute its mean μ and standard deviation σ.
    a. $n = 14, p = 0.55$
    b. $n = 83, p = 0.05$
    c. $n = 957, p = 0.35$
    d. $n = 1750, p = 0.79$
11. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 5, p = 0.\overline{3}$
    b. $n = 10, p = 0.75$
12. X is a binomial random variable with the parameters shown. Compute its mean μ and standard deviation σ in two ways, first using the tables in 7.1: Large Sample Estimation of a Population Mean in conjunction with the general formulas $\mu = \sum xP(x)$ and $\sigma = \sqrt{[\sum x^2 P(x)] - \mu^2}$, then using the special formulas $\mu = np$ and $\sigma = \sqrt{npq}$.
    a. $n = 10, p = 0.25$
    b. $n = 15, p = 0.1$
13. X is a binomial random variable with parameters $n = 10$ and $p = 1/3$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
14. X is a binomial random variable with parameters $n = 15$ and $p = 1/2$. Use the cumulative probability distribution for X that is given in 7.1: Large Sample Estimation of a Population Mean to construct the probability distribution of X.
15. In a certain board game a player's turn begins with three rolls of a pair of dice. If the player rolls doubles all three times there is a penalty. The probability of rolling doubles in a single roll of a pair of fair dice is $1/6$. Find the probability of rolling doubles all three times.
16. A coin is bent so that the probability that it lands heads up is $2/3$. The coin is tossed ten times.
    a. Find the probability that it lands heads up at most five times.
    b. Find the probability that it lands heads up more times than it lands tails up.

Applications

17. An English-speaking tourist visits a country in which 30% of the population speaks English. He needs to ask someone directions.
    a. Find the probability that the first person he encounters will be able to speak English.
    b. The tourist sees four local people standing at a bus stop. Find the probability that at least one of them will be able to speak English.
18. The probability that an egg in a retail package is cracked or broken is 0.025.
    a. Find the probability that a carton of one dozen eggs contains no eggs that are either cracked or broken.
    b. Find the probability that a carton of one dozen eggs has (i) at least one that is either cracked or broken; (ii) at least two that are cracked or broken.
    c. Find the average number of cracked or broken eggs in one dozen cartons.
19. An appliance store sells 20 refrigerators each week. Ten percent of all purchasers of a refrigerator buy an extended warranty. Let X denote the number of the next 20 purchasers who do so.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that X is zero.
    c. Find the probability that X is two, three, or four.
    d. Find the probability that X is at least five.
20. Adverse growing conditions have caused 5% of grapefruit grown in a certain region to be of inferior quality. Grapefruit are sold by the dozen.
    a. Find the average number of inferior quality grapefruit per box of a dozen.
    b. A box that contains two or more grapefruit of inferior quality will cause a strong adverse customer reaction. Find the probability that a box of one dozen grapefruit will contain two or more grapefruit of inferior quality.

21. The probability that a 7-ounce skein of a discount worsted weight knitting yarn contains a knot is 0.25. Goneril buys ten skeins to crochet an afghan.
    a. Find the probability that (i) none of the ten skeins will contain a knot; (ii) at most one will.
    b. Find the expected number of skeins that contain knots.
    c. Find the most likely number of skeins that contain knots.
22. One-third of all patients who undergo a non-invasive but unpleasant medical test require a sedative. A laboratory performs 20 such tests daily. Let X denote the number of patients on any given day who require a sedative.
    a. Verify that X satisfies the conditions for a binomial random variable, and find n and p.
    b. Find the probability that on any given day between five and nine patients will require a sedative (include five and nine).
    c. Find the average number of patients each day who require a sedative.
    d. Using the cumulative probability distribution for X in 7.1: Large Sample Estimation of a Population Mean find the minimum number x_{min} of doses of the sedative that should be on hand at the start of the day so that there is a 99% chance that the laboratory will not run out.
23. About 2% of alumni give money upon receiving a solicitation from the college or university from which they graduated. Find the average number monetary gifts a college can expect from every $2,000$ solicitations it sends.
24. Of all college students who are eligible to give blood, about 18% do so on a regular basis. Each month a local blood bank sends an appeal to give blood to 250 randomly selected students. Find the average number of appeals in such mailings that are made to students who already give blood.
25. About 12% of all individuals write with their left hands. A class of 130 students meets in a classroom with 130 individual desks, exactly 14 of which are constructed for people who write with their left hands. Find the probability that exactly 14 of the students enrolled in the class write with their left hands.
26. A traveling salesman makes a sale on 65% of his calls on regular customers. He makes four sales calls each day.
    a. Construct the probability distribution of X, the number of sales made each day.
    b. Find the probability that, on a randomly selected day, the salesman will make a sale.
    c. Assuming that the salesman makes 20 sales calls per week, find the mean and standard deviation of the number of sales made *per week*.
27. A corporation has advertised heavily to try to insure that over half the adult population recognizes the brand name of its products. In a random sample of 20 adults, 14 recognized its brand name. What is the probability that 14 or more people in such a sample would recognize its brand name if the actual proportion p of all adults who recognize the brand name were only 0.50?

Additional Exercises

28. When dropped on a hard surface a thumbtack lands with its sharp point touching the surface with probability $2/3$; it lands with its sharp point directed up into the air with probability $1/3$. The tack is dropped and its landing position observed 15 times.
    a. Find the probability that it lands with its point in the air at least 7 times.
    b. If the experiment of dropping the tack 15 times is done repeatedly, what is the average number of times it lands with its point in the air?
29. A professional proofreader has a 98% chance of detecting an error in a piece of written work (other than misspellings, double words, and similar errors that are machine detected). A work contains four errors.
    a. Find the probability that the proofreader will miss at least one of them.
    b. Show that two such proofreaders working independently have a 99.96% chance of detecting an error in a piece of written work.
    c. Find the probability that two such proofreaders working independently will miss at least one error in a work that contains four errors.
30. A multiple choice exam has 20 questions; there are four choices for each question.
    a. A student guesses the answer to every question. Find the chance that he guesses correctly between four and seven times.
    b. Find the minimum score the instructor can set so that the probability that a student will pass just by guessing is 20% or less.
31. In spite of the requirement that all dogs boarded in a kennel be inoculated, the chance that a healthy dog boarded in a clean, well-ventilated kennel will develop kennel cough from a carrier is 0.008.
    a. If a carrier (not known to be such, of course) is boarded with three other dogs, what is the probability that at least one of the three healthy dogs will develop kennel cough?

b. If a carrier is boarded with four other dogs, what is the probability that at least one of the four healthy dogs will develop kennel cough?
c. The pattern evident from parts (a) and (b) is that if $K+1$ dogs are boarded together, one a carrier and K healthy dogs, then the probability that at least one of the healthy dogs will develop kennel cough is $P(X \geq 1) = 1-(0.992)^K$, where X is the binomial random variable that counts the number of healthy dogs that develop the condition. Experiment with different values of K in this formula to find the maximum number $K+1$ of dogs that a kennel owner can board together so that if one of the dogs has the condition, the chance that another dog will be infected is less than 0.05.

32. Investigators need to determine which of 600 adults have a medical condition that affects 2% of the adult population. A blood sample is taken from each of the individuals.
a. Show that the expected number of diseased individuals in the group of 600 is 12 individuals.
b. Instead of testing all 600 blood samples to find the expected 12 diseased individuals, investigators group the samples into 60 groups of 10 each, mix a little of the blood from each of the 10 samples in each group, and test each of the 60 mixtures. Show that the probability that any such mixture will contain the blood of at least one diseased person, hence test positive, is about 0.18.
c. Based on the result in (b), show that the expected number of mixtures that test positive is about 11. (Supposing that indeed 11 of the 60 mixtures test positive, then we know that none of the 490 persons whose blood was in the remaining 49 samples that tested negative has the disease. We have eliminated 490 persons from our search while performing only 60 tests.)

Answers

1. a. not binomial; not success/failure.
b. not binomial; trials are not independent.
c. binomial; $n=10, p=0.0002$
d. binomial; $n=6, p=0.5$
e. binomial; $n=6, p=0.5$
2.
3. a. 0.2434
b. 0.2151
c. $0.18^{12} \approx 0$
d. 0
4.
5. a. 0.8125
b. 0.5000
c. 0.3125
d. 0.0313
e. 0.0312
6.
7. a. 0.9965
b. 0.2241
c. 0.0042
d. 0.2252
e. 0.5390
8.
9. a. $\mu = 3.44, \sigma = 1.4003$
b. $\mu = 38.54, \sigma = 2.6339$
c. $\mu = 528, \sigma = 17.1953$
d. $\mu = 1302, \sigma = 22.2432$
10.
11. a. $\mu = 1.6667, \sigma = 1.0541$
b. $\mu = 7.5, \sigma = 1.3693$
12.

13.

x	0	1	2	3
$P(x)$	0.0173	0.0867	0.1951	0.2602

(4.E.27)

x	4	5	6	7
$P(x)$	0.2276	0.1365	0.0569	0.0163

(4.E.28)

x	8	9	10
$P(x)$	0.0030	0.0004	0.0000

(4.E.29)

14.
15. 0.0046
16.
17. a. 0.3
    b. 0.7599
18.
19. a. $n = 20, p = 0.1$
    b. 0.1216
    c. 0.5651
    d. 0.0432
20.
21. a. 0.0563 and 0.2440
    b. 2.5
    c. 2
22.
23. 40
24.
25. 0.1019
26.
27. 0.0577
28.
29. a. 0.0776
    b. 0.9996
    c. 0.0016
30.
31. a. 0.0238
    b. 0.0316
    c. 6

Contributor

- Anonymous

This page titled 4.E: Discrete Random Variables (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **4.E: Discrete Random Variables (Exercises)** has no license indicated.

CHAPTER OVERVIEW

5: Continuous Random Variables

A random variable is called *continuous* if its set of possible values contains a whole interval of decimal numbers. In this chapter we investigate such random variables.

5.1: Continuous Random Variables
5.2: The Standard Normal Distribution
5.3: Probability Computations for General Normal Random Variables
5.4: Areas of Tails of Distributions
5.E: Continuous Random Variables (Exercises)

This page titled 5: Continuous Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Continuous Random Variables

Learning Objectives

- To learn the concept of the probability distribution of a continuous random variable, and how it is used to compute probabilities.
- To learn basic facts about the family of normally distributed random variables.

The Probability Distribution of a Continuous Random Variable

For a discrete random variable X the probability that X assumes one of its possible values on a single trial of the experiment makes good sense. This is not the case for a continuous random variable. For example, suppose X denotes the length of time a commuter just arriving at a bus stop has to wait for the next bus. If buses run every 30 minutes without fail, then the set of possible values of X is the interval denoted $[0, 30]$, the set of all decimal numbers between 0 and 30. But although the number 7.211916 is a possible value of X, there is little or no meaning to the concept of the probability that the commuter will wait precisely 7.211916 minutes for the next bus. If anything the probability should be zero, since if we could meaningfully measure the waiting time to the nearest millionth of a minute it is practically inconceivable that we would ever get exactly 7.211916 minutes. More meaningful questions are those of the form: What is the probability that the commuter's waiting time is less than 10 minutes, or is between 5 and 10 minutes? In other words, with continuous random variables one is concerned not with the event that the variable assumes a single particular value, but with the event that the random variable assumes a value in a particular interval.

Definition: density function

The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$, called a density function, in the following way: the probability that X assumes a value in the interval $[a, b]$ is equal to the area of the region that is bounded above by the graph of the equation $y = f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b, as illustrated in Figure 5.1.1.

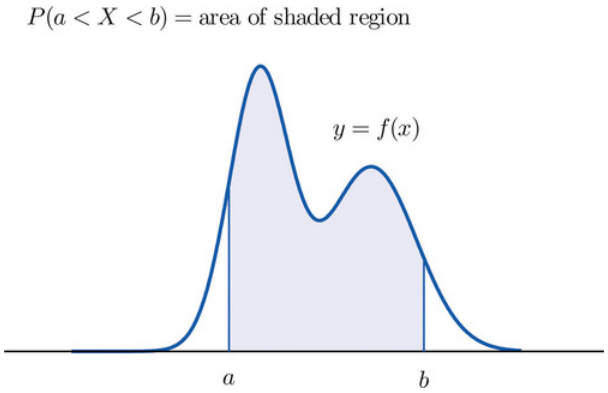


Figure 5.1.1: Probability Given as Area of a Region under a Curve

This definition can be understood as a natural outgrowth of the discussion in Section 2.1.3. There we saw that if we have in view a population (or a very large sample) and make measurements with greater and greater precision, then as the bars in the relative frequency histogram become exceedingly fine their vertical sides merge and disappear, and what is left is just the curve formed by their tops, as shown in Figure 2.1.5. Moreover the total area under the curve is 1, and the proportion of the population with measurements between two numbers a and b is the area under the curve and between a and b, as shown in Figure 2.1.6. If we think of X as a measurement to infinite precision arising from the selection of any one member of the population at random, then $P(a < X < b)$is simply the proportion of the population with measurements between a and b, the curve in the relative frequency histogram is the density function for X, and we arrive at the definition just above.

- Every density function $f(x)$ must satisfy the following two conditions:
- For all numbers x, $f(x) \geq 0$, so that the graph of $y = f(x)$ never drops below the x-axis.
- The area of the region under the graph of $y = f(x)$ and above the x-axis is 1.

Because the area of a line segment is 0, the definition of the probability distribution of a continuous random variable implies that for any particular decimal number, say a, the probability that X assumes the exact value a is 0. This property implies that whether or not the endpoints of an interval are included makes no difference concerning the probability of the interval.

For any continuous random variable X:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Example 5.1.1

A random variable X has the uniform distribution on the interval $[0, 1]$: the density function is $f(x) = 1$ if x is between 0 and 1 and $f(x) = 0$ for all other values of x, as shown in Figure 5.1.2.

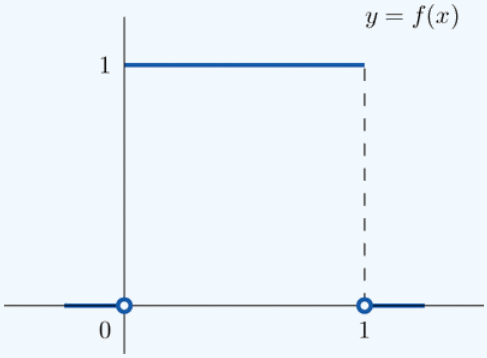


Figure 5.1.2: Uniform Distribution on [0,1].

1. Find $P(X > 0.75)$, the probability that X assumes a value greater than 0.75.
2. Find $P(X \leq 0.2)$, the probability that X assumes a value less than or equal to 0.2.
3. Find $P(0.4 < X < 0.7)$, the probability that X assumes a value between 0.4 and 0.7.

Solution

1. $P(X > 0.75)$ is the area of the rectangle of height 1 and base length $1 - 0.75 = 0.25$, hence is $base \times height = (0.25) \cdot (1) = 0.25$. See Figure 5.1.3$a$.
2. $P(X \leq 0.2)$ is the area of the rectangle of height 1 and base length $0.2 - 0 = 0.2$, hence is $base \times height = (0.2) \cdot (1) = 0.2$. See Figure 5.1.3$b$.
3. $P(0.4 < X < 0.7)$ is the area of the rectangle of height 1 and length $0.7 - 0.4 = 0.3$, hence is $base \times height = (0.3) \cdot (1) = 0.3$. See Figure 5.1.3$c$.

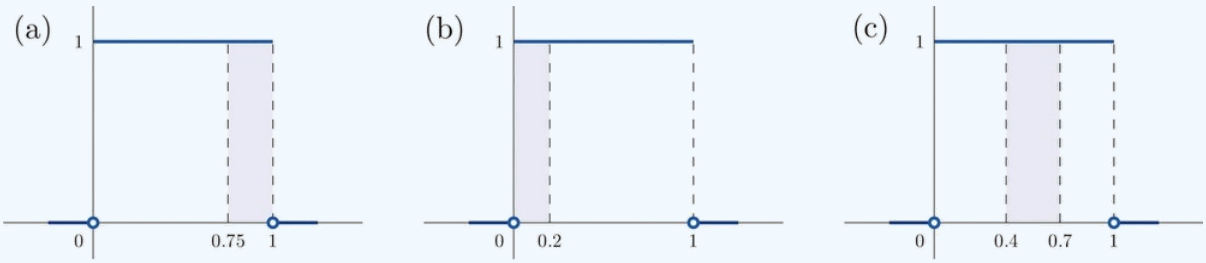


Figure 5.1.3: Probabilities from the Uniform Distribution on [0,1]

Example 5.1.2

A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Solution

The graph of the density function is a horizontal line above the interval from 0 to 30 and is the x-axis everywhere else. Since the total area under the curve must be 1, the height of the horizontal line is $1/30$ (Figure 5.1.4). The probability sought is $P(0 \leq X \leq 10)$.By definition, this probability is the area of the rectangular region bounded above by the horizontal line $f(x) = 1/30$,

bounded below by the x-axis, bounded on the left by the vertical line at 0 (the y-axis), and bounded on the right by the vertical line at 10. This is the shaded region in Figure $5.1.4$. Its area is the base of the rectangle times its height, $(10) \cdot (1/30) = 1/3$. Thus $P(0 \le X \le 10) = 1/3$.

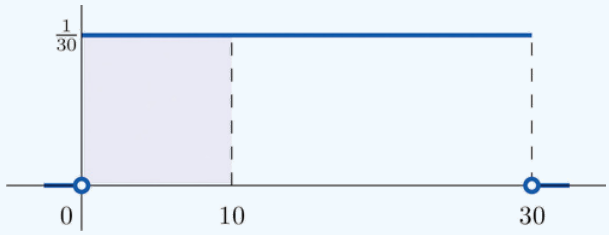


Figure $5.1.4$: Probability of Waiting At Most 10 Minutes for a Bus

Normal Distributions

Most people have heard of the "bell curve." It is the graph of a specific density function $f(x)$ that describes the behavior of continuous random variables as different as the heights of human beings, the amount of a product in a container that was filled by a high-speed packing machine, or the velocities of molecules in a gas. The formula for $f(x)$ contains two parameters μ and σ that can be assigned any specific numerical values, so long as σ is positive. We will not need to know the formula for $f(x)$, but for those who are interested it is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\mu - x)^2/\sigma^2}$$

where $\pi \approx 3.14159$ and $e \approx 2.71828$ is the base of the natural logarithms.

Each different choice of specific numerical values for the pair μ and σ gives a different bell curve. The value of μ determines the location of the curve, as shown in Figure $5.1.5$. In each case the curve is symmetric about μ.

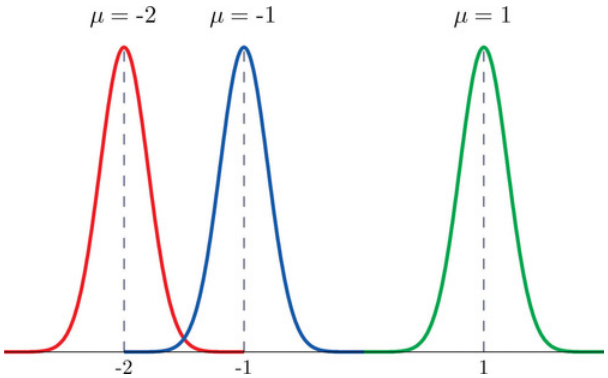


Figure $5.1.5$: Bell Curves with σ = 0.25 and Different Values of μ

The value of σ determines whether the bell curve is tall and thin or short and squat, subject always to the condition that the total area under the curve be equal to 1. This is shown in Figure $5.1.6$, where we have arbitrarily chosen to center the curves at $\mu = 6$.

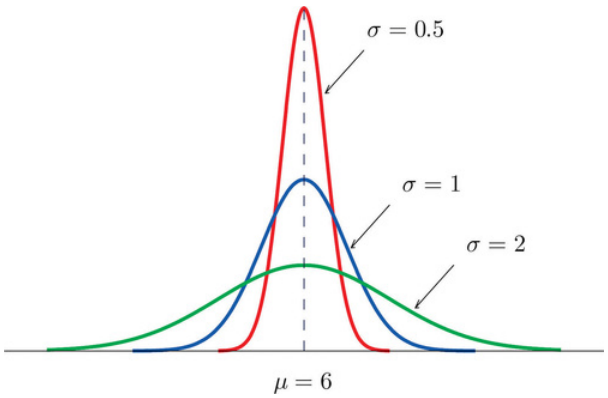


Figure $5.1.6$: Bell Curves with $\mu = 6$ and Different Values of σ.

Definition: normal distribution

The probability distribution corresponding to the density function for the bell curve with parameters μ and σ is called the normal distribution with mean μ and standard deviation σ.

Definition: normally distributed random variable

A continuous random variable whose probabilities are described by the normal distribution with mean μ and standard deviation σ is called a normally distributed random variable, or a normal random variable for short, with mean μ and standard deviation σ.

Figure 5.1.7 shows the density function that determines the normal distribution with mean μ and standard deviation σ. We repeat an important fact about this curve: **The density curve for the normal distribution is symmetric about the mean.**

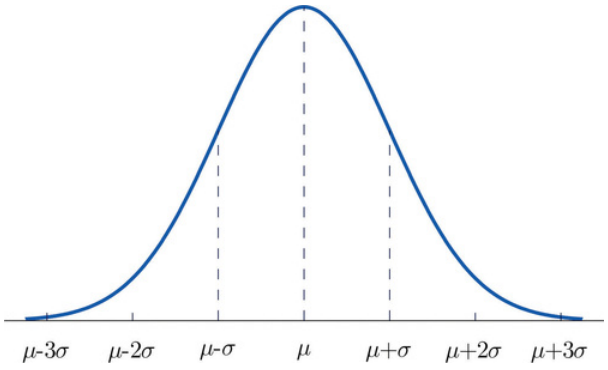


Figure 5.1.7: Density Function for a Normally Distributed Random Variable with Mean μ and Standard Deviation σ

Example 5.1.3

Heights of 25-year-old men in a certain region have mean 69.75 inches and standard deviation 2.59 inches. These heights are approximately normally distributed. Thus the height X of a randomly selected 25-year-old man is a normal random variable with mean $\mu = 69.75$ and standard deviation $\sigma = 2.59$. Sketch a qualitatively accurate graph of the density function for X. Find the probability that a randomly selected 25-year-old man is more than 69.75 inches tall.

Solution

The distribution of heights looks like the bell curve in Figure 5.1.8. The important point is that it is centered at its mean, 69.75, and is symmetric about the mean.

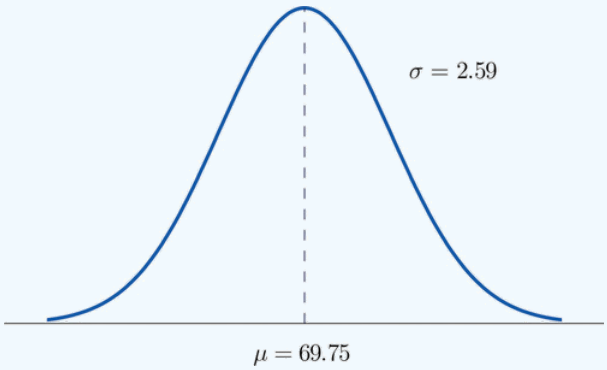


Figure 5.1.8: Density Function for Heights of 25-Year-Old Men

Since the total area under the curve is 1, by symmetry the area to the right of 69.75 is half the total, or 0.5. But this area is precisely the probability $P(X > 69.75)$, the probability that a randomly selected 25-year-old man is more than 69.75 inches tall. We will learn how to compute other probabilities in the next two sections.

Key Takeaway

- For a continuous random variable X the only probabilities that are computed are those of X taking a value in a specified interval.
- The probability that X take a value in a particular interval is the same whether or not the endpoints of the interval are included.
- The probability $P(a < X < b)$, that X take a value in the interval from a to b, is the area of the region between the vertical lines through a and b, above the x-axis, and below the graph of a function $f(x)$ called the density function.
- A normally distributed random variable is one whose density function is a bell curve.
- Every bell curve is symmetric about its mean and lies everywhere above the x-axis, which it approaches asymptotically (arbitrarily closely without touching).

This page titled 5.1: Continuous Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

5.2: The Standard Normal Distribution

Learning Objectives

- To learn what a standard normal random variable is.
- To learn how to compute probabilities related to a standard normal random variable.

Definition: standard normal random variable

A *standard normal random variable* is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. It will always be denoted by the letter Z.

The density function for a standard normal random variable is shown in Figure 5.2.1.

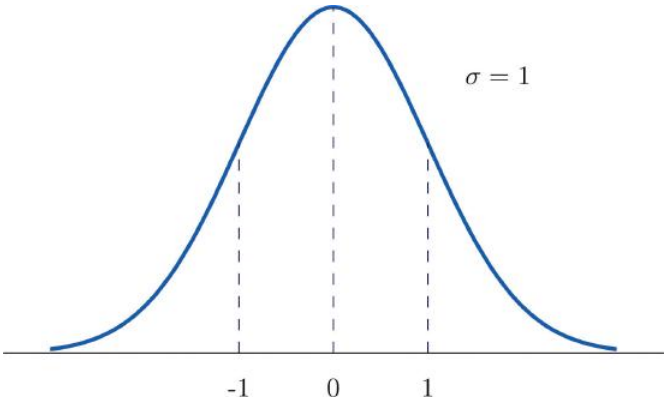


Figure 5.2.1: Density Curve for a Standard Normal Random Variable

To compute probabilities for Z we will not work with its density function directly but instead read probabilities out of Figure 5.2.2. The tables are tables of *cumulative* probabilities; their entries are probabilities of the form $P(Z < z)$. The use of the tables will be explained by the following series of examples.

Example 5.2.1

Find the probabilities indicated, where as always Z denotes a standard normal random variable.

1. $P(Z < 1.48)$.
2. $P(Z < -0.25)$.

Solution

1. Figure 5.2.3 shows how this probability is read directly from the table without any computation required. The digits in the ones and tenths places of 1.48, namely 1.4, are used to select the appropriate row of the table; the hundredths part of 1.48, namely 0.08, is used to select the appropriate column of the table. The four decimal place number in the interior of the table that lies in the intersection of the row and column selected, 0.9306, is the probability sought:

$$P(Z < 1.48) = 0.9306$$

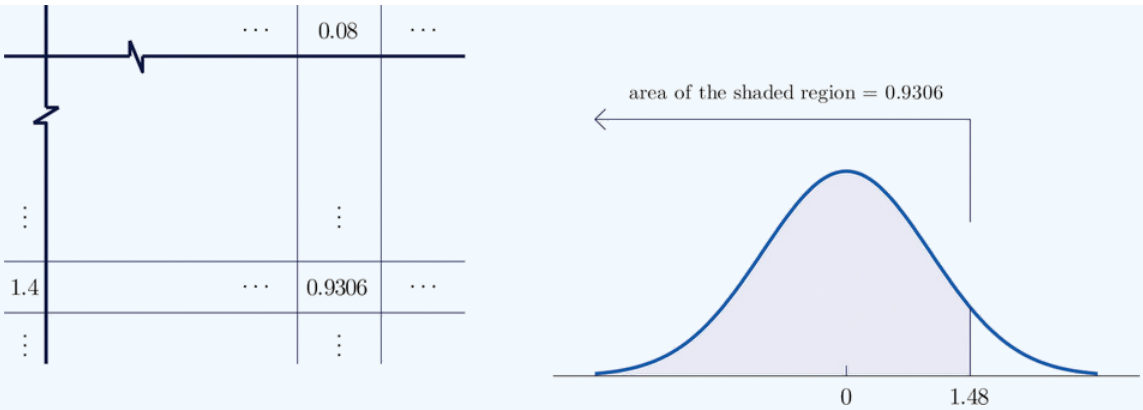


Figure 5.2.3: *Computing Probabilities Using the Cumulative Table*

2. The minus sign in -0.25 makes no difference in the procedure; the table is used in exactly the same way as in part (a): the probability sought is the number that is in the intersection of the row with heading -0.2 and the column with heading 0.05, the number 0.4013. Thus $P(Z < -0.25) = 0.4013$.

Example 5.2.2

Find the probabilities indicated.

1. $P(Z > 1.60)$.
2. $P(Z > -1.02)$.

Solution

1. Because the events $Z > 1.60$ and $Z \leq 1.60$ are complements, the *Probability Rule for Complements* implies that

$$P(Z > 1.60) = 1 - P(Z \leq 1.60)$$

Since inclusion of the endpoint makes no difference for the continuous random variable Z, $P(Z \leq 1.60) = P(Z < 1.60)$, which we know how to find from the table in Figure 5.2.2. The number in the row with heading 1.6 and in the column with heading 0.00 is 0.9452. Thus $P(Z < 1.60) = 0.9452$ so

$$P(Z > 1.60) = 1 - P(Z \leq 1.60) = 1 - 0.9452 = 0.0548$$

Figure 5.2.4 illustrates the ideas geometrically. Since the total area under the curve is 1 and the area of the region to the left of 1.60 is (from the table) 0.9452, the area of the region to the right of 1.60 must be $1 - 0.9452 = 0.0548$.

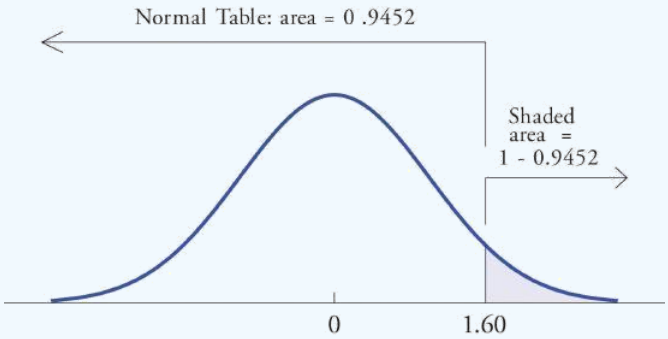


Figure 5.2.4: Computing a Probability for a Right Half-Line

2. The minus sign in -1.02 makes no difference in the procedure; the table is used in exactly the same way as in part (a). The number in the intersection of the row with heading -1.0 and the column with heading 0.02 is 0.1539. This means that $P(Z < -1.02) = P(Z \leq -1.02) = 0.1539$. Hence

$$P(Z > -1.02) = P(Z \leq -1.02) = 1 - 0.1539 = 0.8461$$

Example 5.2.3

Find the probabilities indicated.

1. $P(0.5 < Z < 1.57)$.
2. $P(-2.55 < Z < 0.09)$.

Solution

1. Figure 5.2.5 illustrates the ideas involved for intervals of this type. First look up the areas in the table that correspond to the numbers 0.5 (which we think of as 0.50 to use the table) and 1.57. We obtain 0.6915 and 0.9418, respectively. From the figure it is apparent that we must take the difference of these two numbers to obtain the probability desired. In symbols,

$$P(0.5 < Z < 1.57) = P(Z < 1.57) - P(Z < 0.50) = 0.9418 - 0.6915 = 0.2503$$

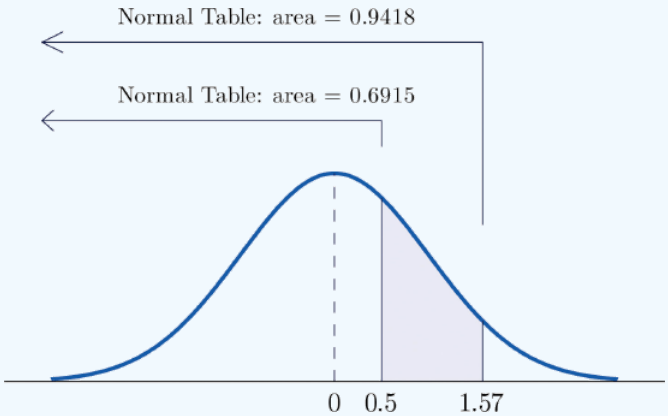


Figure 5.2.5: Computing a Probability for an Interval of Finite Length

2. The procedure for finding the probability that Z takes a value in a finite interval whose endpoints have opposite signs is exactly the same procedure used in part (a), and is illustrated in Figure 5.2.6 "Computing a Probability for an Interval of Finite Length". In symbols the computation is

$$P(-2.55 < Z < 0.09) = P(Z < 0.09) - P(Z < -2.55) = 0.5359 - 0.0054 = 0.5305$$

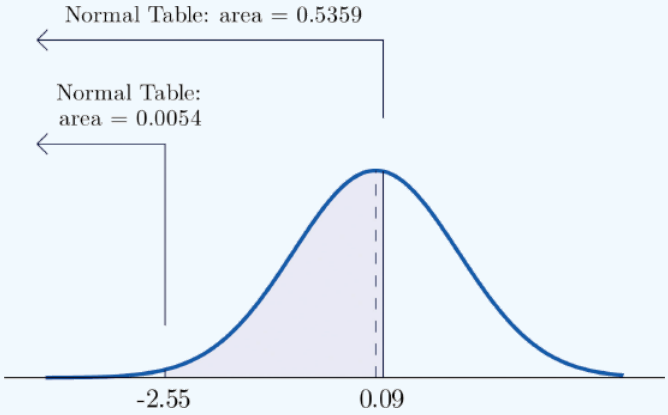


Figure 5.2.6: Computing a Probability for an Interval of Finite Length

The next example shows what to do if the value of Z that we want to look up in the table is not present there.

Example 5.2.4

Find the probabilities indicated.

1. $P(1.13 < Z < 4.16)$.
2. $P(-5.22 < Z < 2.15)$.

Solution

1. We attempt to compute the probability exactly as in Example 5.2.3 by looking up the numbers 1.13 and 4.16 in the table. We obtain the value 0.8708 for the area of the region under the density curve to left of 1.13 without any problem, but when we go to look up the number 4.16 in the table, it is not there. We can see from the last row of numbers in the table that the area to the left of 4.16 must be so close to 1 that to four decimal places it rounds to 1.0000. Therefore

$$P(1.13 < Z < 4.16) = 1.0000 - 0.8708 = 0.1292$$

2. Similarly, here we can read directly from the table that the area under the density curve and to the left of 2.15 is 0.9842, but −5.22 is too far to the left on the number line to be in the table. We can see from the first line of the table that the area to the left of −5.22 must be so close to 0 that to four decimal places it rounds to 0.0000. Therefore

$$P(-5.22 < Z < 2.15) = 0.9842 - 0.0000 = 0.9842$$

The final example of this section explains the origin of the proportions given in the Empirical Rule.

Example 5.2.5

Find the probabilities indicated.

1. $P(-1 < Z < 1)$.
2. $P(-2 < Z < 2)$.
3. $P(-3 < Z < 3)$.

Solution

1. Using the table as was done in Example 5.2.3 we obtain

$$P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826$$

Since Z has mean 0 and standard deviation 1, for Z to take a value between −1 and 1 means that Z takes a value that is within one standard deviation of the mean. Our computation shows that the probability that this happens is about 0.68, the proportion given by the Empirical Rule for histograms that are mound shaped and symmetrical, like the bell curve.

2. Using the table in the same way,

$$P(-2 < Z < 2) = 0.9772 - 0.0228 = 0.9544$$

This corresponds to the proportion 0.95 for data within two standard deviations of the mean.

3. Similarly,

$$P(-3 < Z < 3) = 0.9987 - 0.0013 = 0.9974$$

which corresponds to the proportion 0.997 for data within three standard deviations of the mean.

Key takeaway

- A standard normal random variable Z is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- Probabilities for a standard normal random variable are computed using Figure 5.2.2.

This page titled 5.2: The Standard Normal Distribution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Probability Computations for General Normal Random Variables

Learning Objectives

- To learn how to compute probabilities related to any normal random variable.

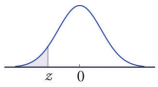
If X is any normally distributed normal random variable then Figure 5.3.1 can also be used to compute a probability of the form $P(a < X < b)$ by means of the following equality.

equality

If X is a normally distributed random variable with mean μ and standard deviation σ, then

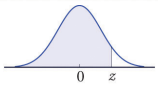
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

where Z denotes a standard normal random variable. a can be any decimal number or $-\infty$; b can be any decimal number or ∞.



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8304	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

https://stats.libretexts.org/@go/page/526

2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 5.3.1: Cumulative Normal Probability

The new endpoints $\frac{(a-\mu)}{\sigma}$ and $\frac{(b-\mu)}{\sigma}$ are the z-scores of a and b as defined in Chapter 2.

Figure 5.3.2 illustrates the meaning of the equality geometrically: the two shaded regions, one under the density curve for X and the other under the density curve for Z, have the same area. Instead of drawing both bell curves, though, we will always draw a single generic bell-shaped curve with both an x-axis and a z-axis below it.

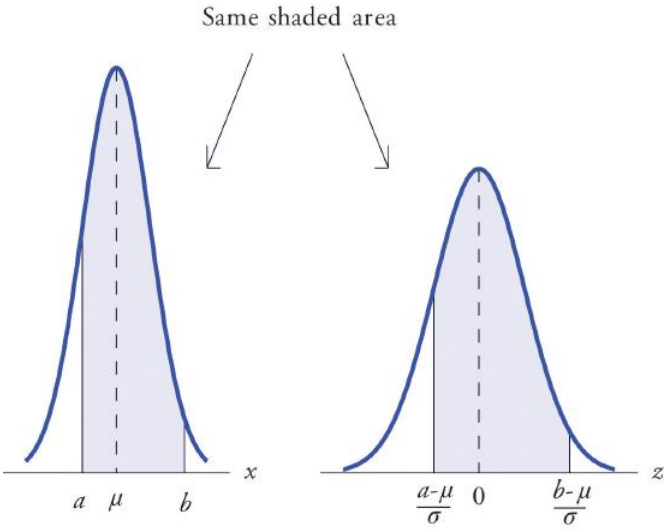


Figure 5.3.2: Probability for an Interval of Finite Length

Example 5.3.1

Let X be a normal random variable with mean $\mu = 10$ and standard deviation $\sigma = 2.5$. Compute the following probabilities.

1. $P(X < 14)$.
2. $P(8 < X < 14)$.

Solution

1. See Figure 5.3.3 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(X < 14) &= P\left(Z < \frac{14-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{14-10}{2.5}\right) \\ &= P(Z < 1.60) \\ &= 0.9452 \end{aligned}$$

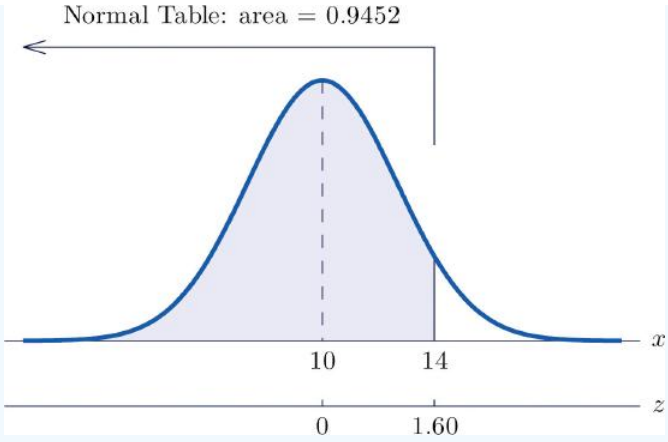


Figure 5.3.3: *Probability Computation for a General Normal Random Variable*

2. See Figure 5.3.4 "Probability Computation for a General Normal Random Variable".

$$\begin{aligned} P(8 < X < 14) &= P\left(\frac{8-10}{2.5} < Z < \frac{14-10}{2.5}\right) \\ &= P\left(-0.80 < Z < 1.60\right) \\ &= 0.9452 - 0.2119 \\ &= 0.7333 \end{aligned}$$

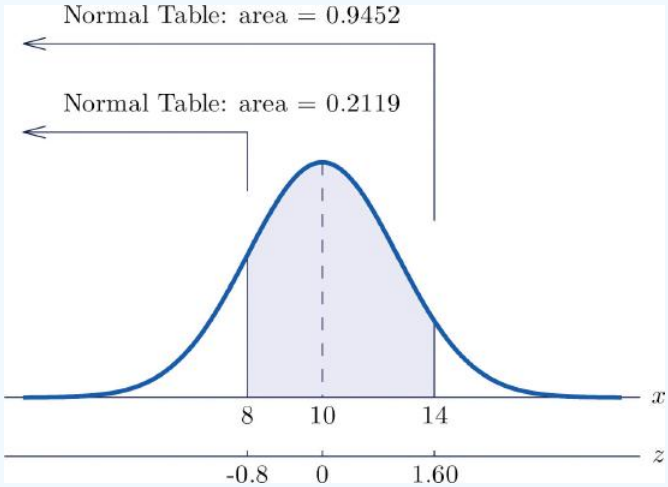


Figure 5.3.4: *Probability Computation for a General Normal Random Variable*

Example 5.3.2

The lifetimes of the tread of a certain automobile tire are normally distributed with mean $37,500$ miles and standard deviation $4,500$ miles. Find the probability that the tread life of a randomly selected tire will be between $30,000$ and $40,000$ miles.

Solution

Let X denote the tread life of a randomly selected tire. To make the numbers easier to work with we will choose thousands of miles as the units. Thus $\mu = 37.5,\ \sigma = 4.5$, and the problem is to compute $P(30 < X < 40)$. Figure 5.3.5 "Probability Computation for Tire Tread Wear" illustrates the following computation:

$$\begin{aligned} P(30 < X < 40) &= P\left(\frac{30-\mu}{\sigma} < Z < \frac{40-\mu}{\sigma}\right) \\ &= P\left(\frac{30-37.5}{4.5} < Z < \frac{40-37.5}{4.5}\right) \\ &= P\left(-1.67 < Z < 0.56\right) \\ &= 0.7123 - 0.0475 \\ &= 0.6648 \end{aligned}$$

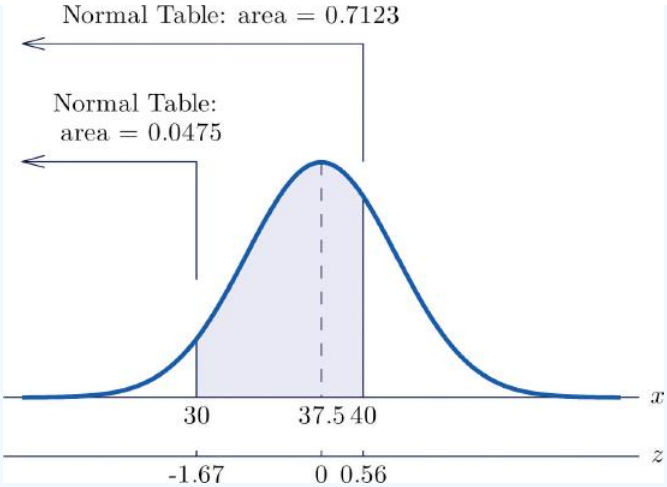


Figure 5.3.5: Probability Computation for Tire Tread Wear

Note that the two z-scores were rounded to two decimal places in order to use Figure 5.3.1 "Cumulative Normal Probability".

Example 5.3.3

Scores on a standardized college entrance examination (*CEE*) are normally distributed with mean 510 and standard deviation 60. A selective university considers for admission only applicants with *CEE* scores over 650. Find percentage of all individuals who took the *CEE* who meet the university's *CEE* requirement for consideration for admission.

Solution

Let X denote the score made on the *CEE* by a randomly selected individual. Then X is normally distributed with mean 510 and standard deviation 60. The probability that X lie in a particular interval is the same as the proportion of all exam scores that lie in that interval. Thus the solution to the problem is $P(X > 650)$, expressed as a percentage. Figure 5.3.6 "Probability Computation for Exam Scores" illustrates the following computation:

$$\begin{aligned} P(X > 650) &= P\left(Z > \frac{650-\mu}{\sigma}\right) \\ &= P\left(Z > \frac{650-510}{60}\right) \\ &= P(Z > 2.33) \\ &= 1 - 0.9901 \\ &= 0.0099 \end{aligned}$$

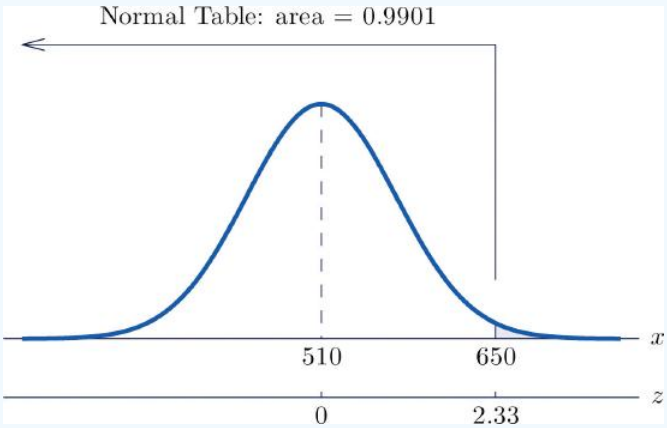


Figure 5.3.6: *Probability Computation for Exam Scores*

The proportion of all *CEE* scores that exceed 650 is 0.0099, hence 0.99% or about 1% do.

- Probabilities for a general normal random variable are computed using Figure 5.3.1 after converting x-values to z-scores.

This page titled 5.3: Probability Computations for General Normal Random Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

5.4: Areas of Tails of Distributions

Learning Objectives

- To learn how to find, for a normal random variable X and an area a, the value x^* of X so that $P(X < x^*) = a$ or that $P(X > x^*) = a$, whichever is required.

Definition: Left and Right Tails

The left tail of a density curve $y = f(x)$ of a continuous random variable X cut off by a value x^* of X is the region under the curve that is to the left of x^*, as shown by the shading in Figure 5.4.1(a). The right tail cut off by x^* is defined similarly, as indicated by the shading in Figure 5.4.1(b).

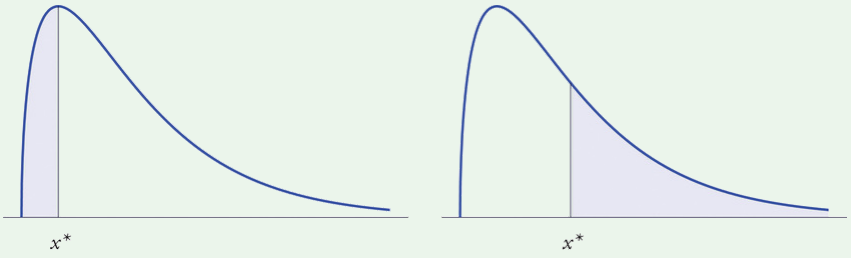


Figure 5.4.1: Right and Left Tails of a Distribution

The probabilities tabulated in Figure 5.3.1 are areas of left tails in the standard normal distribution.

Tails of the Standard Normal Distribution

At times it is important to be able to solve the kind of problem illustrated by Figure 5.4.2. We have a certain specific area in mind, in this case the area 0.0125 of the shaded region in the figure, and we want to find the value z^* of Z that produces it. This is exactly the reverse of the kind of problems encountered so far. Instead of knowing a value z^* of Z and finding a corresponding area, we know the area and want to find z^*. In the case at hand, in the terminology of the definition just above, we wish to find the value z^* that cuts off a left tail of area 0.0125 in the standard normal distribution.

The idea for solving such a problem is fairly simple, although sometimes its implementation can be a bit complicated. In a nutshell, one reads the cumulative probability table for Z in reverse, looking up the relevant area in the interior of the table and reading off the value of Z from the margins.

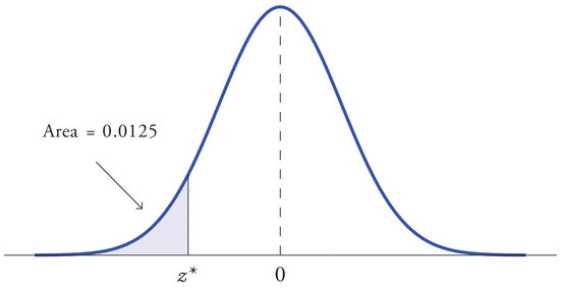


Figure 5.4.2: Z Value that Produces a Known Area

Example 5.4.1

Find the value z^* of Z as determined by Figure 5.4.2: the value z^* that cuts off a left tail of area 0.0125 in the standard normal distribution. In symbols, find the number z^* such that $P(Z < z^*) = 0.0125$.

Solution

The number that is known, 0.0125, is the area of a left tail, and as already mentioned the probabilities tabulated in Figure 5.3.1 are areas of left tails. Thus to solve this problem we need only search in the interior of Figure 5.3.1 for the number 0.0125. It lies in the row with the heading -2.2 and in the column with the heading 0.04. This means that $P(Z < -2.24) = 0.0125$, hence $z^* = -2.24$.

Example 5.4.2

Find the value z^* of Z as determined by Figure 5.4.3: the value z^* that cuts off a right tail of area 0.0250 in the standard normal distribution. In symbols, find the number z^* such that $P(Z > z^*) = 0.0250$.

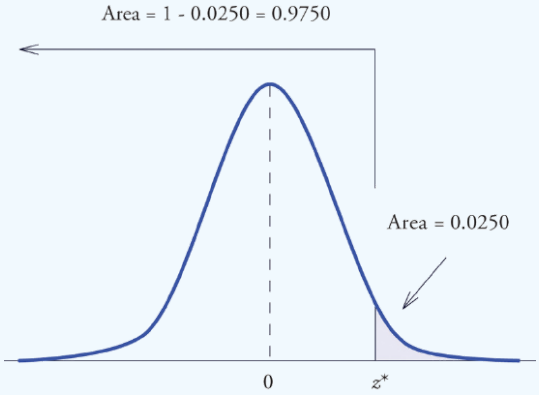


Figure 5.4.3: ZValue that Produces a Known Area

Solution

The important distinction between this example and the previous one is that here it is the area of a right tail that is known. In order to be able to use Figure 5.3.1 we must first find that area of the left tail cut off by the unknown number z^*. Since the total area under the density curve is 1, that area is $1 - 0.0250 = 0.9750$. This is the number we look for in the interior of Figure 5.3.1. It lies in the row with the heading 1.9 and in the column with the heading 0.06. Therefore $z^* = 1.96$.

Definition: standard normal random variable

The value of the standard normal random variable Z that cuts off a right tail of area c is denoted z_c. By symmetry, value of Z that cuts off a left tail of area c is $-z_c$. See Figure 5.4.4.

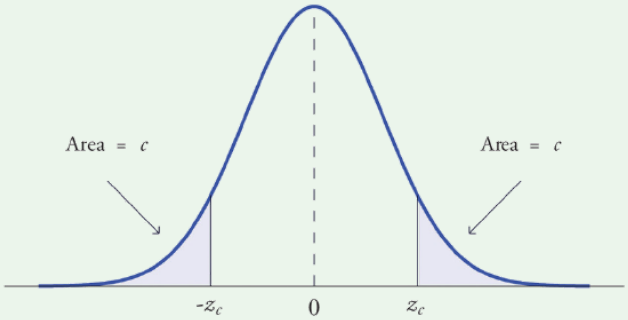


Figure 5.4.4: The Numbers z_c and $-z_c$

The previous two examples were atypical because the areas we were looking for in the interior of Figure 5.3.1 were actually there. The following example illustrates the situation that is more common.

Example 5.4.3

Find $z_{.01}$ and $-z_{.01}$, the values of Z that cut off right and left tails of area 0.01 in the standard normal distribution.

Solution

Since $-z_{.01}$ cuts off a left tail of area 0.01 and Figure 5.3.1 is a table of left tails, we look for the number 0.0100 in the interior of the table. It is not there, but falls between the two numbers 0.0102 and 0.0099 in the row with heading -2.3. The number 0.0099 is closer to 0.0100 than 0.0102 is, so for the hundredths place in $-z_{.01}$ we use the heading of the column that contains 0.0099, namely, 0.03, and write $-z_{.01} \approx -2.33$.

The answer to the second half of the problem is automatic: since $-z_{.01} = -2.33$, we conclude immediately that $z_{.01} = 2.33$.

We could just as well have solved this problem by looking for $z_{.01}$ first, and it is instructive to rework the problem this way. To begin with, we must first subtract 0.01 from 1 to find the area $1 - 0.0100 = 0.9900$ of the left tail cut off by the unknown number $z_{.01}$. See Figure 5.4.5. Then we search for the area 0.9900 in Figure 5.4.5. It is not there, but falls between the numbers 0.9898 and 0.9901 in the row with heading 2.3. Since 0.9901 is closer to 0.9900 than 0.9898 is, we use the column heading above it, 0.03, to obtain the approximation $z_{.01} \approx 2.33$. Then finally $-z_{.01} \approx -2.33$.

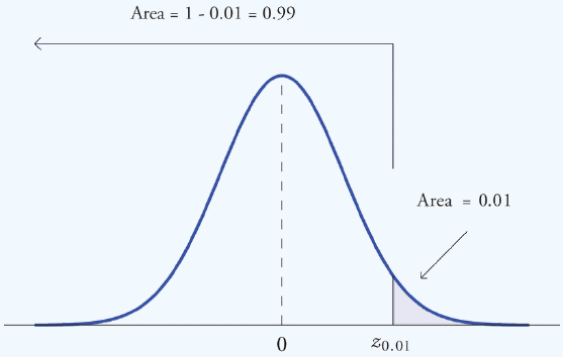


Figure 5.4.5: Computation of the Number $z_{.01}$

Tails of General Normal Distributions

The problem of finding the value x^* of a general normally distributed random variable X that cuts off a tail of a specified area also arises. This problem may be solved in two steps.

1. Suppose X is a normally distributed random variable with mean μ and standard deviation σ. To find the value x^* of X that cuts off a left or right tail of area c in the distribution of X:
2. find the value z^* of Z that cuts off a left or right tail of area c in the standard normal distribution; z^* is the z-score of x^*; compute x^* using the destandardization formula

$$x^* = \mu + z^*\sigma$$

In short, solve the corresponding problem for the standard normal distribution, thereby obtaining the z-score of x^*, then destandardize to obtain x^*.

Example 5.4.4

Find x^* such that $P(X < x^*) = 0.9332$, where X is a normal random variable with mean $\mu = 10$ and standard deviation $\sigma = 2.5$.

Solution

All the ideas for the solution are illustrated in Figure 5.4.6. Since 0.9332 is the area of a left tail, we can find z^* simply by looking for 0.9332 in the interior of Figure 5.3.1. It is in the row and column with headings 1.5 and 0.00, hence $z^* = 1.50$. Thus x^* is 1.50 standard deviations above the mean, so

$$x^* = \mu + z^*\sigma = 10 + (1.50) \cdot (0.2) = 13.75$$

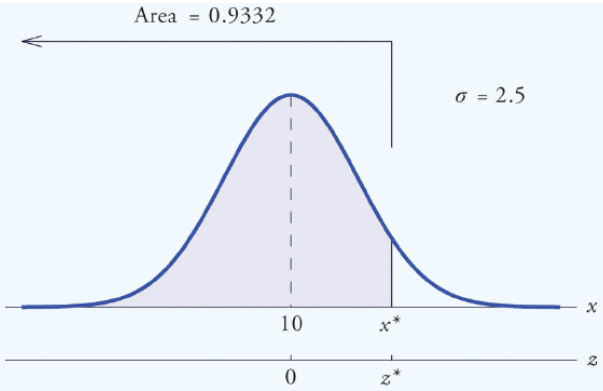


Figure 5.4.6: Tail of a Normally Distributed Random Variable

Example 5.4.5

Find x^* such that $P(X > x^*) = 0.65$, where X is a normal random variable with mean $\mu = 175$ and standard deviation $\sigma = 12$.

Solution

The situation is illustrated in Figure 5.4.7. Since 0.65 is the area of a right tail, we first subtract it from 1 to obtain $1 - 0.65 = 0.35$, the area of the complementary left tail. We find z^* by looking for 0.3500 in the interior of Figure 5.3.1. It is not present, but lies between table entries 0.3520 and 0.3483. The entry 0.3483 with row and column headings -0.3 and 0.09 is closer to 0.3500 than the other entry is, so $z^* \approx -0.39$. Thus x^* is 0.39 standard deviations below the mean, so

$$x^* = \mu + z^*\sigma = 175 + (-0.39) \cdot (12) = 170.32$$

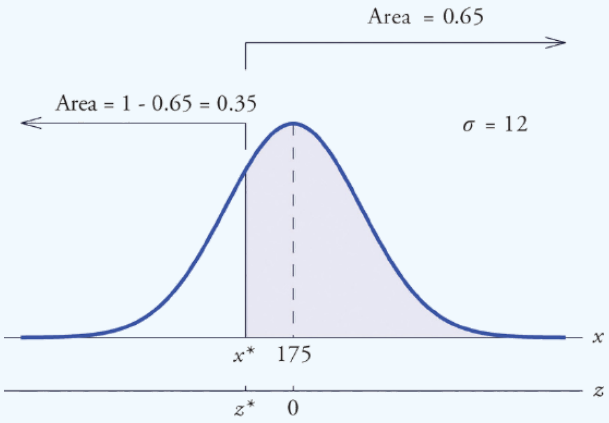


Figure 5.4.7: Tail of a Normally Distributed Random Variable

Example 5.4.6

Scores on a standardized college entrance examination (CEE) are normally distributed with mean 510 and standard deviation 60. A selective university decides to give serious consideration for admission to applicants whose CEE scores are in the top 5% of all CEE scores. Find the minimum score that meets this criterion for serious consideration for admission.

Solution

Let X denote the score made on the CEE by a randomly selected individual. Then X is normally distributed with mean 510 and standard deviation 60. The probability that X lie in a particular interval is the same as the proportion of all exam scores that lie in that interval. Thus the minimum score that is in the top 5% of all CEE is the score x^* that cuts off a right tail in the distribution of X of area 0.05 (5% expressed as a proportion). See Figure 5.4.8.

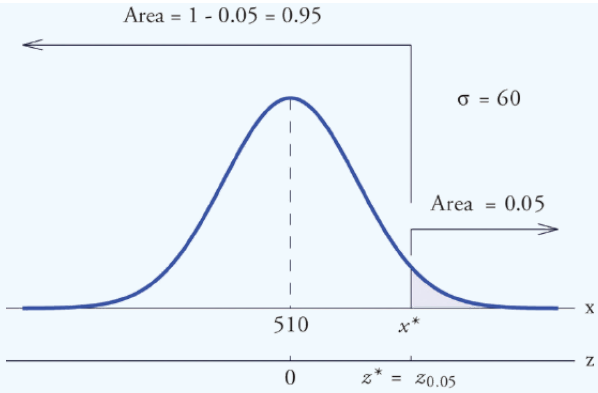


Figure 5.4.8: Tail of a Normally Distributed Random Variable

Since 0.0500 is the area of a right tail, we first subtract it from 1 to obtain $1 - 0.0500 = 0.9500$, the area of the complementary left tail. We find $z^* = z_{.05}$by looking for 0.9500 in the interior of Figure 5.3.1. It is not present, and lies exactly half-way between the two nearest entries that are, 0.9495 and 0.9505. In the case of a tie like this, we will always average the values of Z corresponding to the two table entries, obtaining here the value $z^* = 1.645$. Using this value, we conclude that x^* is 1.645 standard deviations above the mean, so

$$x^* = \mu + z^*\sigma = 510 + (1.645) \cdot (60) = 608.7$$

Example 5.4.7

All boys at a military school must run a fixed course as fast as they can as part of a physical examination. Finishing times are normally distributed with mean 29 minutes and standard deviation 2 minutes. The middle 75% of all finishing times are classified as "average." Find the range of times that are average finishing times by this definition.

Solution

Let X denote the finish time of a randomly selected boy. Then X is normally distributed with mean 29 and standard deviation 2. The probability that X lie in a particular interval is the same as the proportion of all finish times that lie in that interval. Thus the situation is as shown in Figure 5.4.9. Because the area in the middle corresponding to "average" times is 0.75, the areas of the two tails add up to $1 - 0.75 = 0.25$ in all. By the symmetry of the density curve each tail must have half of this total, or area 0.125 each. Thus the fastest time that is "average" has z-score $-z_{.125}$, which by Figure 5.3.1 is -1.15, and the slowest time that is "average" has z-score $z_{.125} = 1.15$. The fastest and slowest times that are still considered average are

$$x_{fast} = \mu + (-z_{.125})\sigma = 29 + (-1.15) \cdot (2) = 26.7$$

and

$$x_{slow} = \mu + z_{.125}\sigma = 29 + (1.15) \cdot (2) = 31.3$$

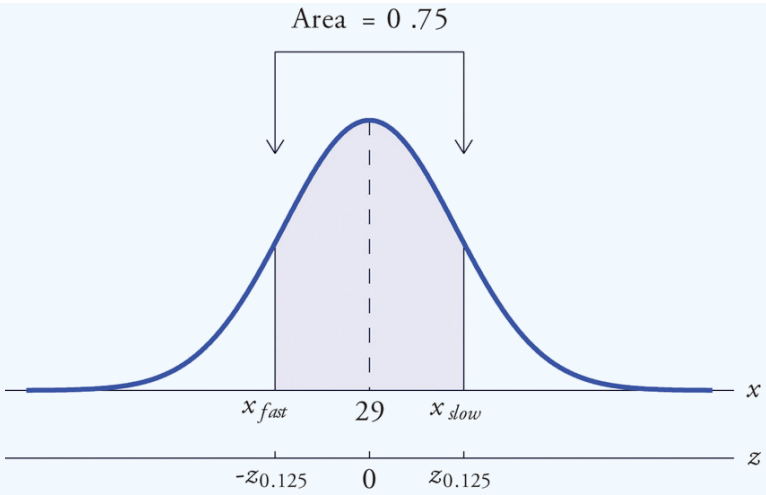


Figure 5.4.9: Distribution of Times to Run a Course

A boy has an average finishing time if he runs the course with a time between 26.7 and 31.3 minutes, or equivalently between 26 minutes 42 seconds and 31 minutes 18 seconds.

Key Takeaways

- The problem of finding the number z^* so that the probability $P(Z < z^*)$ is a specified value c is solved by looking for the number c in the interior of Figure 5.3.1 and reading z^* from the margins.
- The problem of finding the number z^* so that the probability $P(Z > z^*)$ is a specified value c is solved by looking for the complementary probability $1 - c$ in the interior of Figure 5.3.1 and reading z^* from the margins.
- For a normal random variable X with mean μ and standard deviation σ, the problem of finding the number x^* so that $P(X < x^*)$ is a specified value c (or so that $P(X > x^*)$ is a specified value c) is solved in two steps:
  - (1) solve the corresponding problem for Z with the same value of c, thereby obtaining the z-score, z^*, of x^*;
  - (2) find x^* using $x^* = \mu + z^*\sigma$.
- The value of Z that cuts off a right tail of area c in the standard normal distribution is denoted z_c.

This page titled 5.4: Areas of Tails of Distributions is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

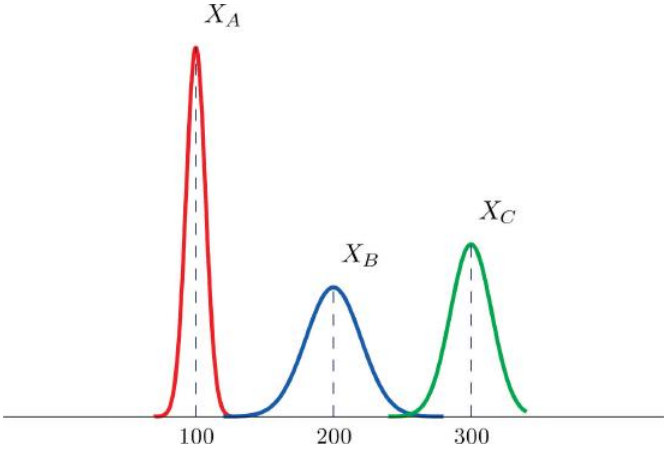
5.E: Continuous Random Variables (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

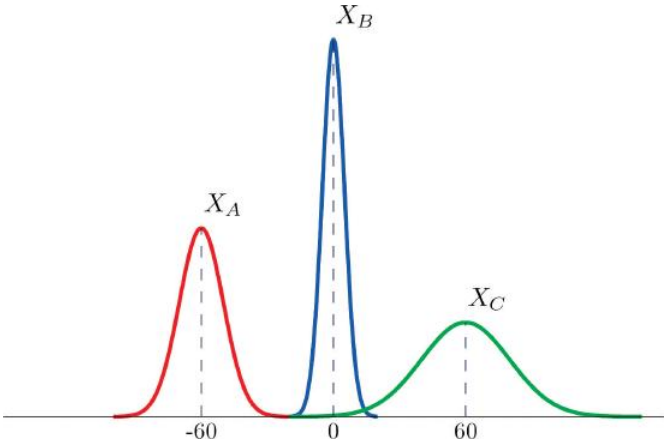
5.1: Continuous Random Variables

Basic

1. A continuous random variable X has a uniform distribution on the interval $[5, 12]$. Sketch the graph of its density function.
2. A continuous random variable X has a uniform distribution on the interval $[-3, 3]$. Sketch the graph of its density function.
3. A continuous random variable X has a normal distribution with mean 100 and standard deviation 10. Sketch a qualitatively accurate graph of its density function.
4. A continuous random variable X has a normal distribution with mean 73 and standard deviation 2.5. Sketch a qualitatively accurate graph of its density function.
5. A continuous random variable X has a normal distribution with mean 73. The probability that X takes a value greater than 80 is 0.212. Use this information and the symmetry of the density function to find the probability that X takes a value less than 66. Sketch the density curve with relevant regions shaded to illustrate the computation.
6. A continuous random variable X has a normal distribution with mean 169. The probability that X takes a value greater than 180 is 0.17. Use this information and the symmetry of the density function to find the probability that X takes a value less than 158. Sketch the density curve with relevant regions shaded to illustrate the computation.
7. A continuous random variable X has a normal distribution with mean 50.5. The probability that X takes a value less than 54 is 0.76. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 47. Sketch the density curve with relevant regions shaded to illustrate the computation.
8. A continuous random variable X has a normal distribution with mean 12.25. The probability that X takes a value less than 13 is 0.82. Use this information and the symmetry of the density function to find the probability that X takes a value greater than 11.50. Sketch the density curve with relevant regions shaded to illustrate the computation.
9. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 15, 7, and 20. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



10. The figure provided shows the density curves of three normally distributed random variables X_A, X_B and X_C. Their standard deviations (in no particular order) are 20, 5, and 10. Use the figure to identify the values of the means μ_A, μ_B, and μ_C and standard deviations σ_A, σ_B, and σ_C of the three random variables.



Applications

11. Dogberry's alarm clock is battery operated. The battery could fail with equal probability at any time of the day or night. Every day Dogberry sets his alarm for $6:30\ a.m.$ and goes to bed at $10:00\ p.m.$. Find the probability that when the clock battery finally dies, it will do so at the most inconvenient time, between $10:00\ p.m.$ and $6:30\ a.m.$.
12. Buses running a bus line near Desdemona's house run every 15 minutes. Without paying attention to the schedule she walks to the nearest stop to take the bus to town. Find the probability that she waits more than 10 minutes.
13. The amount X of orange juice in a randomly selected half-gallon container varies according to a normal distribution with mean 64 ounces and standard deviation 0.25 ounce.
    a. Sketch the graph of the density function for X.
    b. What proportion of all containers contain less than a half gallon (64 ounces)? Explain.
    c. What is the median amount of orange juice in such containers? Explain.
14. The weight X of grass seed in bags marked 50 lb varies according to a normal distribution with mean 50 lb and standard deviation 1 ounce (0.0625 lb).
    a. Sketch the graph of the density function for X.
    b. What proportion of all bags weigh less than 50 pounds? Explain.
    c. What is the median weight of such bags? Explain.

Answers

1. The graph is a horizontal line with height $1/7$ from $x = 5$ to $x = 12$
2.
3. The graph is a bell-shaped curve centered at 100 and extending from about 70 to 130.
4.
5. 0.212
6.
7. 0.76
8.
9. $\mu_A = 100,\ \mu_B = 200,\ \mu_C = 300,\ \sigma_A = 7,\ \sigma_B = 20,\ \sigma_C = 15$
10.
11. 0.3542
12.
13. a. The graph is a bell-shaped curve centered at 64 and extending from about 63.25 to 64.75.
    b. 0.5
    c. 64

5.2: The Standard Normal Distribution

Basic

1. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(Z < -1.72)$
   b. $P(Z < 2.05)$
   c. $P(Z < 0)$
   d. $P(Z > -2.11)$
   e. $P(Z > 1.63)$
   f. $P(Z > 2.36)$
2. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(Z < -1.17)$
   b. $P(Z < -0.05)$
   c. $P(Z < 0.66)$
   d. $P(Z > -2.43)$
   e. $P(Z > -1.00)$
   f. $P(Z > 2.19)$
3. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-2.15 < Z < -1.09)$
   b. $P(-0.93 < Z < 0.55)$
   c. $P(0.68 < Z < 2.11)$
4. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-1.99 < Z < -1.03)$
   b. $P(-0.87 < Z < 1.58)$
   c. $P(0.33 < Z < 0.96)$
5. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(-4.22 < Z < -1.39)$
   b. $P(-1.37 < Z < 5.11)$
   c. $P(Z < -4.31)$
   d. $P(Z < 5.02)$
6. Use Figure 7.1.5: Cumulative Normal Probability to find the probability indicated.
   a. $P(Z > -5.31)$
   b. $P(-4.08 < Z < 0.58)$
   c. $P(Z < -6.16)$
   d. $P(-0.51 < Z < 5.63)$
7. Use Figure 7.1.5: Cumulative Normal Probability to find the probability listed. Find the second probability without referring to the table, but using the symmetry of the standard normal density curve instead. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(Z < -1.08),\ P(Z > 1.08)$
   b. $P(Z < -0.36),\ P(Z > 0.36)$
   c. $P(Z < 1.25),\ P(Z > -1.25)$
   d. $P(Z < 2.03),\ P(Z > -2.03)$
8. Use Figure 7.1.5: Cumulative Normal Probability to find the probability listed. Find the second probability without referring to the table, but using the symmetry of the standard normal density curve instead. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(Z < -2.11),\ P(Z > 2.11)$
   b. $P(Z < -0.88),\ P(Z > 0.88)$
   c. $P(Z < 2.44),\ P(Z > -2.44)$
   d. $P(Z < 3.07),\ P(Z > -3.07)$
9. The probability that a standard normal random variable Z takes a value in the union of intervals $(-\infty, -\alpha] \cup [\alpha, \infty)$, which arises in applications, will be denoted $P(Z \leq -a \text{ or } Z \geq a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the following

probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the standard normal density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.

a. $P(Z < -1.29 \text{ or } Z > 1.29)$
b. $P(Z < -2.33 \text{ or } Z > 2.33)$
c. $P(Z < -1.96 \text{ or } Z > 1.96)$
d. $P(Z < -3.09 \text{ or } Z > 3.09)$

10. The probability that a standard normal random variable Z takes a value in the union of intervals $(-\infty, -\alpha] \cup [\alpha, \infty)$, which arises in applications, will be denoted $P(Z \leq -a \text{ or } Z \geq a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the standard normal density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.

a. $P(Z < -2.58 \text{ or } Z > 2.58)$
b. $P(Z < -2.81 \text{ or } Z > 2.81)$
c. $P(Z < -1.65 \text{ or } Z > 1.65)$
d. $P(Z < -2.43 \text{ or } Z > 2.43)$

Answers

1. a. 0.0427
   b. 0.9798
   c. 0.5
   d. 0.9826
   e. 0.0516
   f. 0.0091
2.
3. a. 0.1221
   b. 0.5326
   c. 0.2309
4.
5. a. 0.0823
   b. 0.9147
   c. 0.0000
   d. 1.0000
6.
7. a. $0.1401,\ 0.1401$
   b. $0.3594,\ 0.3594$
   c. $0.8944,\ 0.8944$
   d. $0.9788,\ 0.9788$
8.
9. a. 0.1970
   b. 0.01980
   c. 0.0500
   d. 0.0020

5.3: Probability Computations for General Normal Random Variables

Basic

1. X is a normally distributed random variable with mean 57 and standard deviation 6. Find the probability indicated.
   a. $P(X < 59.5)$
   b. $P(X < 46.2)$
   c. $P(X > 52.2)$

d. $P(X > 70)$

2. X is a normally distributed random variable with mean -25 and standard deviation 4. Find the probability indicated.
   a. $P(X < -27.2)$
   b. $P(X < -14.8)$
   c. $P(X > -33.1)$
   d. $P(X > -16.5)$
3. X is a normally distributed random variable with mean 112 and standard deviation 15. Find the probability indicated.
   a. $P(100 < X < 125)$
   b. $P(91 < X < 107)$
   c. $P(118 < X < 160)$
4. X is a normally distributed random variable with mean 72 and standard deviation 22. Find the probability indicated.
   a. $P(78 < X < 127)$
   b. $P(60 < X < 90)$
   c. $P(49 < X < 71)$
5. X is a normally distributed random variable with mean 500 and standard deviation 25. Find the probability indicated.
   a. $P(X < 400)$
   b. $P(466 < X < 625)$
6. X is a normally distributed random variable with mean 0 and standard deviation 0.75. Find the probability indicated.
   a. $P(-4.02 < X < 3.82)$
   b. $P(X > 4.11)$
7. X is a normally distributed random variable with mean 15 and standard deviation 1. Use Figure 7.1.5: Cumulative Normal Probability to find the first probability listed. Find the second probability using the symmetry of the density curve. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(X < 12),\ P(X > 18)$
   b. $P(X < 14),\ P(X > 16)$
   c. $P(X < 11.25),\ P(X > 18.75)$
   d. $P(X < 12.67),\ P(X > 17.33)$
8. X is a normally distributed random variable with mean 100 and standard deviation 10. Use Figure 7.1.5: Cumulative Normal Probability to find the first probability listed. Find the second probability using the symmetry of the density curve. Sketch the density curve with relevant regions shaded to illustrate the computation.
   a. $P(X < 80),\ P(X > 120)$
   b. $P(X < 75),\ P(X > 125)$
   c. $P(X < 84.55),\ P(X > 115.45)$
   d. $P(X < 77.42),\ P(X > 122.58)$
9. X is a normally distributed random variable with mean 67 and standard deviation 13. The probability that X takes a value in the union of intervals $(-\infty, 67 - a] \cup [67 + a, \infty)$ will be denoted $P(X \leq 67 - a \text{ or } X \geq 67 + a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.
   a. $P(X < 57 \text{ or } X > 77)$
   b. $P(X < 47 \text{ or } X > 87)$
   c. $P(X < 49 \text{ or } X > 85)$
   d. $P(X < 37 \text{ or } X > 97)$
10. X is a normally distributed random variable with mean 288 and standard deviation 6. The probability that X takes a value in the union of intervals $(-\infty, 288 - a] \cup [288 + a, \infty)$ will be denoted $P(X \leq 288 - a \text{ or } X \geq 288 + a)$. Use Figure 7.1.5: Cumulative Normal Probability to find the following probabilities of this type. Sketch the density curve with relevant regions shaded to illustrate the computation. Because of the symmetry of the density curve you need to use Figure 7.1.5: Cumulative Normal Probability only one time for each part.
   a. $P(X < 278 \text{ or } X > 298)$

b. $P(X < 268 \text{ or } X > 308)$
c. $P(X < 273 \text{ or } X > 303)$
d. $P(X < 280 \text{ or } X > 296)$

Applications

11. The amount X of beverage in a can labeled 12 ounces is normally distributed with mean 12.1 ounces and standard deviation 0.05 ounce. A can is selected at random.
    a. Find the probability that the can contains at least 12 ounces.
    b. Find the probability that the can contains between 11.9 and 12.1 ounces.
12. The length of gestation for swine is normally distributed with mean 114 days and standard deviation 0.75 day. Find the probability that a litter will be born within one day of the mean of 114.
13. The systolic blood pressure X of adults in a region is normally distributed with mean 112 mm Hg and standard deviation 15 mm Hg. A person is considered "prehypertensive" if his systolic blood pressure is between 120 and 130 mm Hg. Find the probability that the blood pressure of a randomly selected person is prehypertensive.
14. Heights X of adult women are normally distributed with mean 63.7 inches and standard deviation 2.71 inches. Romeo, who is 69.25 inches tall, wishes to date only women who are shorter than he but within 4 inches of his height. Find the probability that the next woman he meets will have such a height.
15. Heights X of adult men are normally distributed with mean 69.1 inches and standard deviation 2.92 inches. Juliet, who is 63.25 inches tall, wishes to date only men who are taller than she but within 6 inches of her height. Find the probability that the next man she meets will have such a height.
16. A regulation hockey puck must weigh between 5.5 and 6 ounces. The weights X of pucks made by a particular process are normally distributed with mean 5.75 ounces and standard deviation 0.11 ounce. Find the probability that a puck made by this process will meet the weight standard.
17. A regulation golf ball may not weigh more than 1.620 ounces. The weights X of golf balls made by a particular process are normally distributed with mean 1.361 ounces and standard deviation 0.09 ounce. Find the probability that a golf ball made by this process will meet the weight standard.
18. The length of time that the battery in Hippolyta's cell phone will hold enough charge to operate acceptably is normally distributed with mean 25.6 hours and standard deviation 0.32 hour. Hippolyta forgot to charge her phone yesterday, so that at the moment she first wishes to use it today it has been 26 hours 18 minutes since the phone was last fully charged. Find the probability that the phone will operate properly.
19. The amount of non-mortgage debt per household for households in a particular income bracket in one part of the country is normally distributed with mean $\$28,350$ and standard deviation $\$3,425$. Find the probability that a randomly selected such household has between $\$20,000$ and $\$30,000$ in non-mortgage debt.
20. Birth weights of full-term babies in a certain region are normally distributed with mean 7.125 lb and standard deviation 1.290 lb. Find the probability that a randomly selected newborn will weigh less than 5.5 lb, the historic definition of prematurity.
21. The distance from the seat back to the front of the knees of seated adult males is normally distributed with mean 23.8 inches and standard deviation 1.22 inches. The distance from the seat back to the back of the next seat forward in all seats on aircraft flown by a budget airline is 26 inches. Find the proportion of adult men flying with this airline whose knees will touch the back of the seat in front of them.
22. The distance from the seat to the top of the head of seated adult males is normally distributed with mean 36.5 inches and standard deviation 1.39 inches. The distance from the seat to the roof of a particular make and model car is 40.5 inches. Find the proportion of adult men who when sitting in this car will have at least one inch of headroom (distance from the top of the head to the roof).

Additional Exercises

23. The useful life of a particular make and type of automotive tire is normally distributed with mean $57,500$ miles and standard deviation 950 miles.
    a. Find the probability that such a tire will have a useful life of between $57,000$ and $58,000$ miles.
    b. Hamlet buys four such tires. Assuming that their lifetimes are independent, find the probability that all four will last between $57,000$ and $58,000$ miles. (If so, the best tire will have no more than $1,000$ miles left on it when the first tire fails.) Hint: There is a binomial random variable here, whose value of p comes from part (a).

24. A machine produces large fasteners whose length must be within 0.5 inch of 22 inches. The lengths are normally distributed with mean 22.0 inches and standard deviation 0.17 inch.
    a. Find the probability that a randomly selected fastener produced by the machine will have an acceptable length.
    b. The machine produces 20 fasteners per hour. The length of each one is inspected. Assuming lengths of fasteners are independent, find the probability that all 20 will have acceptable length. Hint: There is a binomial random variable here, whose value of p comes from part (a).
25. The lengths of time taken by students on an algebra proficiency exam (if not forced to stop before completing it) are normally distributed with mean 28 minutes and standard deviation 1.5 minutes.
    a. Find the proportion of students who will finish the exam if a 30-minute time limit is set.
    b. Six students are taking the exam today. Find the probability that all six will finish the exam within the 30-minute limit, assuming that times taken by students are independent. Hint: There is a binomial random variable here, whose value of p comes from part (a).
26. Heights of adult men between 18 and 34 years of age are normally distributed with mean 69.1 inches and standard deviation 2.92 inches. One requirement for enlistment in the military is that men must stand between 60 and 80 inches tall.
    a. Find the probability that a randomly elected man meets the height requirement for military service.
    b. Twenty-three men independently contact a recruiter this week. Find the probability that all of them meet the height requirement. Hint: There is a binomial random variable here, whose value of p comes from part (a).
27. A regulation hockey puck must weigh between 5.5 and 6 ounces. In an alternative manufacturing process the mean weight of pucks produced is 5.75 ounce. The weights of pucks have a normal distribution whose standard deviation can be decreased by increasingly stringent (and expensive) controls on the manufacturing process. Find the maximum allowable standard deviation so that at most 0.005 of all pucks will fail to meet the weight standard. (Hint: The distribution is symmetric and is centered at the middle of the interval of acceptable weights.)
28. The amount of gasoline X delivered by a metered pump when it registers 5 gallons is a normally distributed random variable. The standard deviation σ of Xmeasures the precision of the pump; the smaller σ is the smaller the variation from delivery to delivery. A typical standard for pumps is that when they show that 5 gallons of fuel has been delivered the actual amount must be between 4.97 and 5.03 gallons (which corresponds to being off by at most about half a cup). Supposing that the mean of X is 5, find the largest that σ can be so that $P(4.97 < X < 5.03)$ is 1.0000 to four decimal places when computed using Figure 7.1.5: Cumulative Normal Probability which means that the pump is sufficiently accurate. (Hint: The z-score of 5.03 will be the smallest value of Z so that Figure 7.1.5: Cumulative Normal Probability gives $P(Z < z) = 1.0000$).

Answers

1. a. 0.6628
   b. 0.0359
   c. 0.7881
   d. 0.0150
2.
3. a. 0.5959
   b. 0.2899
   c. 0.3439
4.
5. a. 0.0000
   b. 0.9131
6.
7. a. $0.0013,\ 0.0013$
   b. $0.1587,\ 0.1587$
   c. $0.0001,\ 0.0001$
   d. $0.0099,\ 0.0099$
8.
9. a. 0.4412
   b. 0.1236

c. 0.1676
d. 0.0208

10.

11. a. 0.9772
    b. 0.5000

12.

13. 0.1830

14.

15. 0.4971

16.

17. 0.9980

18.

19. 0.6771

20.

21. 0.0359

22.

23. a. 0.4038
    b. 0.0266

24.

25. a. 0.9082
    b. 0.5612

26.

27. 0.089

5.4: Areas of Tails of Distributions

Basic

1. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.0075$
   b. $P(Z < z*) = 0.9850$
   c. $P(Z > z*) = 0.8997$
   d. $P(Z > z*) = 0.0110$
2. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.3300$
   b. $P(Z < z*) = 0.9901$
   c. $P(Z > z*) = 0.0055$
   d. $P(Z > z*) = 0.7995$
3. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.1500$
   b. $P(Z < z*) = 0.7500$
   c. $P(Z > z*) = 0.3333$
   d. $P(Z > z*) = 0.8000$
4. Find the value of $z*$ that yields the probability shown.
   a. $P(Z < z*) = 0.2200$
   b. $P(Z < z*) = 0.6000$
   c. $P(Z > z*) = 0.0750$
   d. $P(Z > z*) = 0.8200$
5. Find the indicated value of Z. (It is easier to find $-z_c$and negate it.)
   a. $Z_{0.025}$
   b. $Z_{0.20}$

6. Find the indicated value of Z. (It is easier to find $-z_c$and negate it.)
    a. $Z_{0.002}$
    b. $Z_{0.02}$
7. Find the value of $x*$ that yields the probability shown, where X is a normally distributed random variable X with mean 83 and standard deviation 4.
    a. $P(X < x*) = 0.8700$
    b. $P(X > x*) = 0.0500$
8. Find the value of $x*$ that yields the probability shown, where X is a normally distributed random variable X with mean 54 and standard deviation 12.
    a. $P(X < x*) = 0.0900$
    b. $P(X > x*) = 0.6500$
9. X is a normally distributed random variable X with mean 15 and standard deviation 0.25. Find the values X_L and X_R of X that are symmetrically located with respect to the mean of X and satisfy $P(X_L < X < X_R) = 0.80$. (Hint. First solve the corresponding problem for Z).
10. X is a normally distributed random variable X with mean 28 and standard deviation 3.7. Find the values X_L and X_R of X that are symmetrically located with respect to the mean of X and satisfy $P(X_L < X < X_R) = 0.65$. (Hint. First solve the corresponding problem for Z).

Applications

11. Scores on a national exam are normally distributed with mean 382 and standard deviation 26.
    a. Find the score that is the 50^{th} percentile.
    b. Find the score that is the 90^{th} percentile.
12. Heights of women are normally distributed with mean 63.7 inches and standard deviation 2.47 inches.
    a. Find the height that is the 10^{th} percentile.
    b. Find the height that is the 80^{th} percentile.
13. The monthly amount of water used per household in a small community is normally distributed with mean $7,069$ gallons and standard deviation 58 gallons. Find the three quartiles for the amount of water used.
14. The quantity of gasoline purchased in a single sale at a chain of filling stations in a certain region is normally distributed with mean 11.6 gallons and standard deviation 2.78 gallons. Find the three quartiles for the quantity of gasoline purchased in a single sale.
15. Scores on the common final exam given in a large enrollment multiple section course were normally distributed with mean 69.35 and standard deviation 12.93. The department has the rule that in order to receive an A in the course his score must be in the top 10% of all exam scores. Find the minimum exam score that meets this requirement.
16. The average finishing time among all high school boys in a particular track event in a certain state is 5 minutes 17 seconds. Times are normally distributed with standard deviation 12 seconds.
    a. The qualifying time in this event for participation in the state meet is to be set so that only the fastest 5% of all runners qualify. Find the qualifying time. (Hint: Convert seconds to minutes.)
    b. In the western region of the state the times of all boys running in this event are normally distributed with standard deviation 12 seconds, but with mean 5 minutes 22 seconds. Find the proportion of boys from this region who qualify to run in this event in the state meet.
17. Tests of a new tire developed by a tire manufacturer led to an estimated mean tread life of $67,350$ miles and standard deviation of $1,120$ miles. The manufacturer will advertise the lifetime of the tire (for example, a "$50,000$ mile tire") using the largest value for which it is expected that 98% of the tires will last at least that long. Assuming tire life is normally distributed, find that advertised value.
18. Tests of a new light led to an estimated mean life of $1,321$ hours and standard deviation of 106 hours. The manufacturer will advertise the lifetime of the bulb using the largest value for which it is expected that 90% of the bulbs will last at least that long. Assuming bulb life is normally distributed, find that advertised value.
19. The weights X of eggs produced at a particular farm are normally distributed with mean 1.72 ounces and standard deviation 0.12 ounce. Eggs whose weights lie in the middle 75% of the distribution of weights of all eggs are classified as "medium." Find the

maximum and minimum weights of such eggs. (These weights are endpoints of an interval that is symmetric about the mean and in which the weights of 75% of the eggs produced at this farm lie.)

20. The lengths X of hardwood flooring strips are normally distributed with mean 28.9 inches and standard deviation 6.12 inches. Strips whose lengths lie in the middle 80% of the distribution of lengths of all strips are classified as “average-length strips.” Find the maximum and minimum lengths of such strips. (These lengths are endpoints of an interval that is symmetric about the mean and in which the lengths of 80% of the hardwood strips lie.)
21. All students in a large enrollment multiple section course take common in-class exams and a common final, and submit common homework assignments. Course grades are assigned based on students' final overall scores, which are approximately normally distributed. The department assigns a C to students whose scores constitute the middle $2/3$ of all scores. If scores this semester had mean 72.5 and standard deviation 6.14, find the interval of scores that will be assigned a C.
22. Researchers wish to investigate the overall health of individuals with abnormally high or low levels of glucose in the blood stream. Suppose glucose levels are normally distributed with mean 96 and standard deviation 8.5 mg/dl, and that “normal” is defined as the middle 90% of the population. Find the interval of normal glucose levels, that is, the interval centered at 96 that contains 90% of all glucose levels in the population.

Additional Exercises

23. A machine for filling 2-liter bottles of soft drink delivers an amount to each bottle that varies from bottle to bottle according to a normal distribution with standard deviation 0.002 liter and mean whatever amount the machine is set to deliver.
    a. If the machine is set to deliver 2 liters (so the mean amount delivered is 2 liters) what proportion of the bottles will contain at least 2 liters of soft drink?
    b. Find the minimum setting of the mean amount delivered by the machine so that at least 99% of all bottles will contain at least 2 liters.
24. A nursery has observed that the mean number of days it must darken the environment of a species poinsettia plant daily in order to have it ready for market is 71 days. Suppose the lengths of such periods of darkening are normally distributed with standard deviation 2 days. Find the number of days in advance of the projected delivery dates of the plants to market that the nursery must begin the daily darkening process in order that at least 95% of the plants will be ready on time. (Poinsettias are so long-lived that once ready for market the plant remains salable indefinitely.)

Answers

1. a. -2.43
   b. 2.17
   c. -1.28
   d. 2.29
2.
3. a. -1.04
   b. 0.67
   c. 0.43
   d. -0.84
4.
5. a. 1.96
   b. 0.84
6.
7. a. 87.52
   b. 89.58
8.
9. 15.32
10.
11. a. 382
    b. 415
12.
13. $7030.14,\ 7069,\ 7107.86$

14.
15. 85.90
16.
17. $65,054$
18.
19. $1.58,\ 1.86$
20.
21. $66.5,\ 78.5$
22.
23. a. 0.5
    b. 2.005

Contributor

- Anonymous

This page titled 5.E: Continuous Random Variables (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **5.E: Continuous Random Variables (Exercises)** has no license indicated.

CHAPTER OVERVIEW

6: Sampling Distributions

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

6.1: The Mean and Standard Deviation of the Sample Mean

6.2: The Sampling Distribution of the Sample Mean

6.3: The Sample Proportion

6.E: Sampling Distributions (Exercises)

This page titled 6: Sampling Distributions is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

6.1: The Mean and Standard Deviation of the Sample Mean

Learning Objectives

- 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 6.1.1

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:

$\bar{x}$	152	154	156	158	160	162	164
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Now we apply the formulas from Section 4.2 to $\bar{X}$. For $\mu_{\bar{X}}$, we obtain.

$$\begin{aligned}\mu_{\bar{X}} &= \sum \bar{x} P(\bar{x}) \\ &= 152\left(\frac{1}{16}\right) + 154\left(\frac{2}{16}\right) + 156\left(\frac{3}{16}\right) + 158\left(\frac{4}{16}\right) + 160\left(\frac{3}{16}\right) + 162\left(\frac{2}{16}\right) + 164\left(\frac{1}{16}\right) \\ &= 158\end{aligned}$$

For $\sigma_{\bar{X}}$, we first compute $\sum \bar{x}^2 P(\bar{x})$:

$$\sum \bar{x}^2 P(\bar{x})$$

$$= 152^2\left(\frac{1}{16}\right) + 154^2\left(\frac{2}{16}\right) + 156^2\left(\frac{3}{16}\right) + 158^2\left(\frac{4}{16}\right) + 160^2\left(\frac{3}{16}\right) + 162^2\left(\frac{2}{16}\right) + 164^2\left(\frac{1}{16}\right)$$

which is $24,974$, so that

$$\sigma_{\bar{x}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - \mu_{\bar{x}}^2}$$
$$= \sqrt{24,974 - 158^2}$$
$$= \sqrt{10}$$

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

$$\mu_{\bar{X}} = \mu \tag{6.1.1}$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{6.1.2}$$

Equation 6.1.1 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 6.1.2 says that averages computed from samples vary less than individual measurements on the population do, and quantifies the relationship.

Example 6.1.2

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

$$\mu_{\bar{X}} = \mu = \$13,525$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$4,180}{\sqrt{100}} = \$418$$

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.

This page titled 6.1: The Mean and Standard Deviation of the Sample Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

6.2: The Sampling Distribution of the Sample Mean

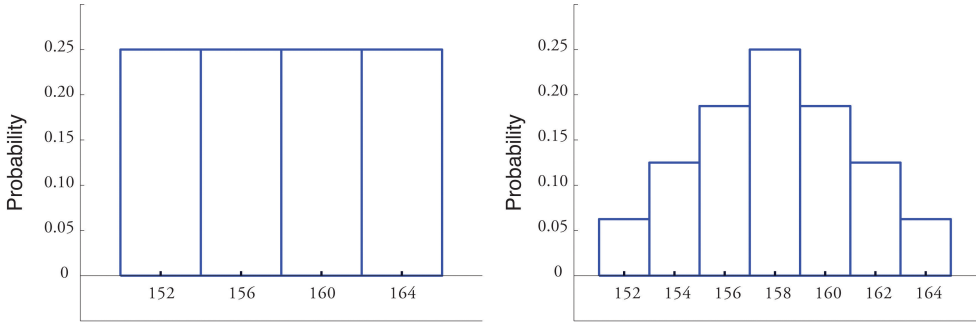
Learning Objectives

- 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
$P(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Figure 6.2.1 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.



(a) Population (b) Sample Mean

Figure 6.2.1: 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

$n = 5$:

$\bar{x}$	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

$n = 10$:

$\bar{x}$	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

$n = 20$:

$\bar{x}$	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

and

$\bar{x}$	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

Histograms illustrating these distributions are shown in Figure 6.2.2.

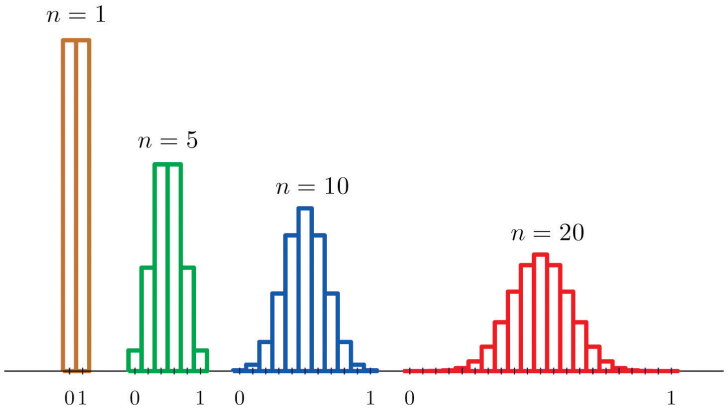


Figure 6.2.2: Distributions of the Sample Mean

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 $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$, 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 6.2.3.

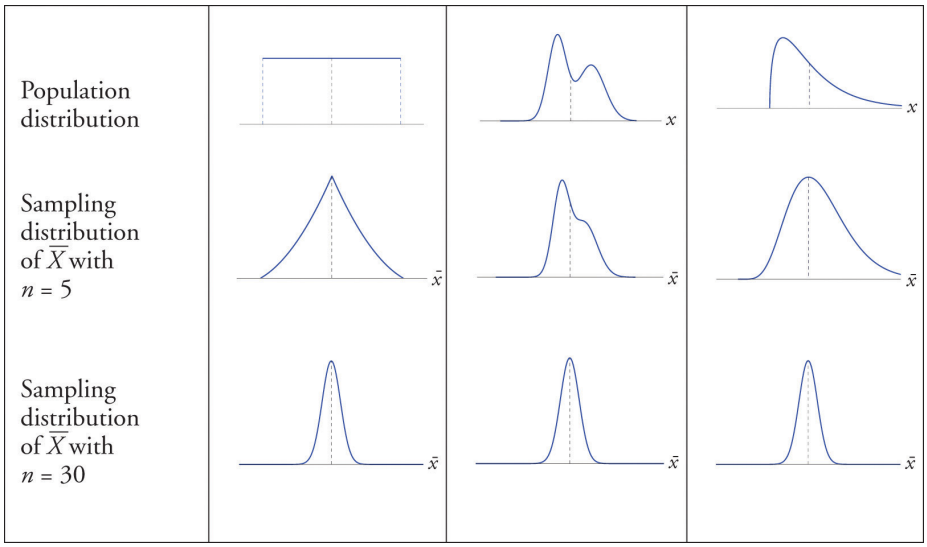


Figure 6.2.3: Distribution of Populations and Sample Means

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 $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

Example 6.2.1

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

$$\mu_{\overline{X}} = \mu = 112$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.65685$$

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 in the 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 6.2.2

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 $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.5}{10} = 0.05$, 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 $\mu_X = \mu$ and standard deviation $\sigma_X = \sigma/\sqrt{n}$, where n is the sample size.

The effect of increasing the sample size is shown in Figure 6.2.4.

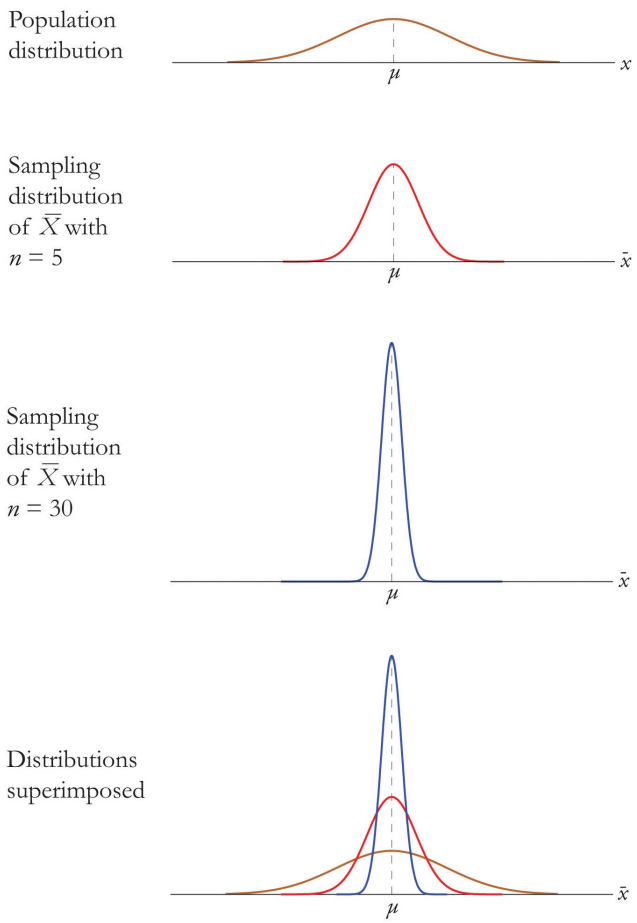


Figure 6.2.4: Distribution of Sample Means for a Normal Population

Example 6.2.3

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 6.2.4

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.

This page titled 6.2: The Sampling Distribution of the Sample Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

6.3: The Sample Proportion

Learning Objectives

- 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

$$\mu_{\hat{P}} = p$$

and

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

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,

$$p = \frac{\text{number of 1s}}{N}$$

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

$$\mu = \frac{\sum x}{N} = \frac{\text{number of 1s}}{N}$$

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 $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$.

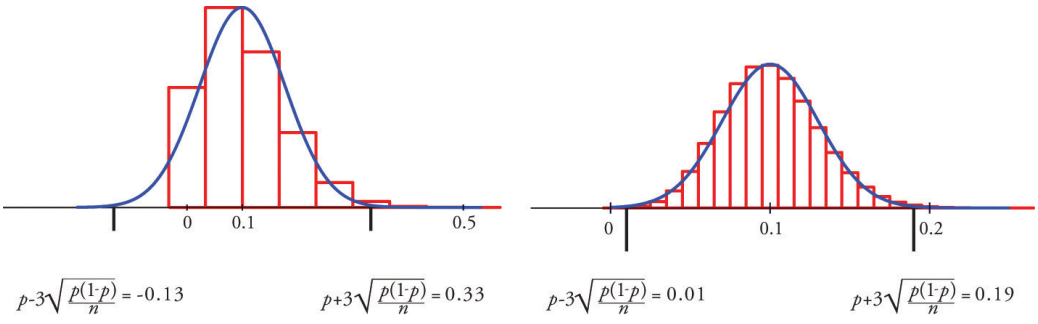
A sample is large if the interval $[p - 3\sigma_{\hat{p}},\ p + 3\sigma_{\hat{p}}]$ lies wholly within the interval $[0, 1]$.

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

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lies wholly within the interval $[0, 1]$. This is illustrated in the examples.

Figure 6.3.1 shows that when $p = 0.1$, a sample of size 15 is too small but a sample of size 100 is acceptable.



(a) $p = 0.1, \quad n = 15$ (b) $p = 0.1, \quad n = 100$

Figure 6.3.1: *Distribution of Sample Proportions*

Figure 6.3.2 shows that when $p = 0.5$ a sample of size 15 is acceptable.

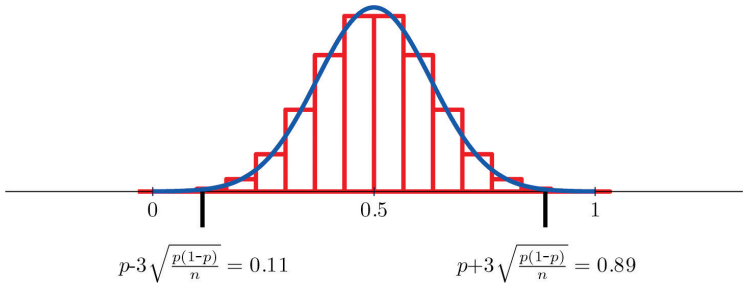


Figure 6.3.2: Distribution of Sample Proportions for $p = 0.5$ *and* $n = 15$

Example 6.3.1

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

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.38 - 0.05, 0.38 + 0.05] = [0.33, 0.43]$$

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 6.3.2

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 $n = 121$,

$$\sigma_{\hat{P}} = \sqrt{\frac{(0.90)(0.10)}{121}} = 0.0\overline{27}$$

hence

$$\left[p - 3\sigma_{\hat{P}},\, p + 3\sigma_{\hat{P}}\right] = [0.90 - 0.08, 0.90 + 0.08] = [0.82, 0.98]$$

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),

$$\begin{aligned} P(\hat{P} \le 0.84) &= P\left(Z \le \frac{0.84 - \mu_{\hat{P}}}{\sigma_{\hat{P}}}\right) \\ &= P\left(Z \le \frac{0.84 - 0.90}{0.0\overline{27}}\right) \\ &= P(Z \le -2.20) \\ &= 0.0139 \end{aligned}$$

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.

This page titled 6.3: The Sample Proportion is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

6.E: Sampling Distributions (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

6.1: The Mean and Standard Deviation of the Sample Mean

Basic

Q6.1.1

Random samples of size 225 are drawn from a population with mean 100 and standard deviation 20. Find the mean and standard deviation of the sample mean.

Q6.1.2

Random samples of size 64 are drawn from a population with mean 32 and standard deviation 5. Find the mean and standard deviation of the sample mean.

Q6.1.3

A population has mean 75 and standard deviation 12.

a. Random samples of size 121 are taken. Find the mean and standard deviation of the sample mean.
b. How would the answers to part (a) change if the size of the samples were 400 instead of 121?

Q6.1.4

A population has mean 5.75 and standard deviation 1.02.

a. Random samples of size 81 are taken. Find the mean and standard deviation of the sample mean.
b. How would the answers to part (a) change if the size of the samples were 25 instead of 81?

Answers

S6.1.1

$\mu_{\bar{X}} = 100,\ \sigma_{\bar{X}} = 1.33$

S6.1.3

a. $\mu_{\bar{X}} = 75,\ \sigma_{\bar{X}} = 1.09$
b. $\mu_{\bar{X}}$ stays the same but $\sigma_{\bar{X}}$ decreases to 0.6

6.2: The Sampling Distribution of the Sample Mean

Basic

1. A population has mean 128 and standard deviation 22.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 36.
   b. Find the probability that the mean of a sample of size 36 will be within 10 units of the population mean, that is, between 118 and 138.
2. A population has mean $1,542$ and standard deviation 246.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 100.
   b. Find the probability that the mean of a sample of size 100 will be within 100 units of the population mean, that is, between $1,442$ and $1,642$.
3. A population has mean 73.5 and standard deviation 2.5.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 30.
   b. Find the probability that the mean of a sample of size 30 will be less than 72.
4. A population has mean 48.4 and standard deviation 6.3.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 64.
   b. Find the probability that the mean of a sample of size 64 will be less than 46.7.
5. A normally distributed population has mean 25.6 and standard deviation 3.3.

a. Find the probability that a single randomly selected element X of the population exceeds 30.
b. Find the mean and standard deviation of $\overline{X}$ for samples of size 9.
c. Find the probability that the mean of a sample of size 9 drawn from this population exceeds 30.

6. A normally distributed population has mean 57.7 and standard deviation 12.1.
   a. Find the probability that a single randomly selected element X of the population is less than 45.
   b. Find the mean and standard deviation of $\overline{X}$ for samples of size 16.
   c. Find the probability that the mean of a sample of size 16 drawn from this population is less than 45.
7. A population has mean 557 and standard deviation 35.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 50.
   b. Find the probability that the mean of a sample of size 50 will be more than 570.
8. A population has mean 16 and standard deviation 1.7.
   a. Find the mean and standard deviation of $\overline{X}$ for samples of size 80.
   b. Find the probability that the mean of a sample of size 80 will be more than 16.4.
9. A normally distributed population has mean $1,214$ and standard deviation 122.
   a. Find the probability that a single randomly selected element X of the population is between $1,100$ and $1,300$.
   b. Find the mean and standard deviation of $\overline{X}$ for samples of size 25.
   c. Find the probability that the mean of a sample of size 25 drawn from this population is between $1,100$ and $1,300$.
10. A normally distributed population has mean $57,800$ and standard deviation 750.
    a. Find the probability that a single randomly selected element X of the population is between $57,000$ and $58,000$.
    b. Find the mean and standard deviation of $\overline{X}$ for samples of size 100.
    c. Find the probability that the mean of a sample of size 100 drawn from this population is between $57,000$ and $58,000$.
11. A population has mean 72 and standard deviation 6.
    a. Find the mean and standard deviation of $\overline{X}$ for samples of size 45.
    b. Find the probability that the mean of a sample of size 45 will differ from the population mean 72 by at least 2 units, that is, is either less than 70 or more than 74. (Hint: One way to solve the problem is to first find the probability of the complementary event.)
12. A population has mean 12 and standard deviation 1.5.
    a. Find the mean and standard deviation of $\overline{X}$ for samples of size 90.
    b. Find the probability that the mean of a sample of size 90 will differ from the population mean 12 by at least 0.3 unit, that is, is either less than 11.7 or more than 12.3. (Hint: One way to solve the problem is to first find the probability of the complementary event.)

Applications

13. Suppose the mean number of days to germination of a variety of seed is 22, with standard deviation 2.3 days. Find the probability that the mean germination time of a sample of 160 seeds will be within 0.5 day of the population mean.
14. Suppose the mean length of time that a caller is placed on hold when telephoning a customer service center is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a sample of $1,200$ calls will be within 0.5 second of the population mean.
15. Suppose the mean amount of cholesterol in eggs labeled “large” is 186 milligrams, with standard deviation 7 milligrams. Find the probability that the mean amount of cholesterol in a sample of 144 eggs will be within 2 milligrams of the population mean.
16. Suppose that in one region of the country the mean amount of credit card debt per household in households having credit card debt is $\$15,250$, with standard deviation $\$7,125$. Find the probability that the mean amount of credit card debt in a sample of $1,600$ such households will be within $\$300$ of the population mean.
17. Suppose speeds of vehicles on a particular stretch of roadway are normally distributed with mean 36.6 mph and standard deviation 1.7 mph.
    a. Find the probability that the speed X of a randomly selected vehicle is between 35 and 40 mph.
    b. Find the probability that the mean speed $\overline{X}$ of 20 randomly selected vehicles is between 35 and 40 mph.
18. Many sharks enter a state of tonic immobility when inverted. Suppose that in a particular species of sharks the time a shark remains in a state of tonic immobility when inverted is normally distributed with mean 11.2 minutes and standard deviation 1.1 minutes.

a. If a biologist induces a state of tonic immobility in such a shark in order to study it, find the probability that the shark will remain in this state for between 10 and 13 minutes.
b. When a biologist wishes to estimate the mean time that such sharks stay immobile by inducing tonic immobility in each of a sample of 12 sharks, find the probability that mean time of immobility in the sample will be between 10 and 13 minutes.

19. Suppose the mean cost across the country of a 30-day supply of a generic drug is $\$46.58$, with standard deviation $\$4.84$. Find the probability that the mean of a sample of 100 prices of 30-day supplies of this drug will be between $\$45$ and $\$50$.
20. Suppose the mean length of time between submission of a state tax return requesting a refund and the issuance of the refund is 47 days, with standard deviation 6 days. Find the probability that in a sample of 50 returns requesting a refund, the mean such time will be more than 50 days.
21. Scores on a common final exam in a large enrollment, multiple-section freshman course are normally distributed with mean 72.7 and standard deviation 13.1.
    a. Find the probability that the score X on a randomly selected exam paper is between 70 and 80.
    b. Find the probability that the mean score $\overline{X}$ of 38 randomly selected exam papers is between 70 and 80.
22. Suppose the mean weight of school children's bookbags is 17.4 pounds, with standard deviation 2.2 pounds. Find the probability that the mean weight of a sample of 30 bookbags will exceed 17 pounds.
23. Suppose that in a certain region of the country the mean duration of first marriages that end in divorce is 7.8 years, standard deviation 1.2 years. Find the probability that in a sample of 75 divorces, the mean age of the marriages is at most 8 years.
24. Borachio eats at the same fast food restaurant every day. Suppose the time X between the moment Borachio enters the restaurant and the moment he is served his food is normally distributed with mean 4.2 minutes and standard deviation 1.3 minutes.
    a. Find the probability that when he enters the restaurant today it will be at least 5 minutes until he is served.
    b. Find the probability that average time until he is served in eight randomly selected visits to the restaurant will be at least 5 minutes.

Additional Exercises

25. A high-speed packing machine can be set to deliver between 11 and 13 ounces of a liquid. For any delivery setting in this range the amount delivered is normally distributed with mean some amount μ and with standard deviation 0.08 ounce. To calibrate the machine it is set to deliver a particular amount, many containers are filled, and 25 containers are randomly selected and the amount they contain is measured. Find the probability that the sample mean will be within 0.05 ounce of the actual mean amount being delivered to all containers.
26. A tire manufacturer states that a certain type of tire has a mean lifetime of $60,000$ miles. Suppose lifetimes are normally distributed with standard deviation $\sigma = 3,500$ miles.
    a. Find the probability that if you buy one such tire, it will last only $57,000$ or fewer miles. If you had this experience, is it particularly strong evidence that the tire is not as good as claimed?
    b. A consumer group buys five such tires and tests them. Find the probability that average lifetime of the five tires will be $57,000$ miles or less. If the mean is so low, is that particularly strong evidence that the tire is not as good as claimed?

Answers

1. a. $\mu_{\overline{X}} = 128,\ \sigma_{\overline{X}} = 3.67$
   b. 0.9936
2.
3. a. $\mu_{\overline{X}} = 73.5,\ \sigma_{\overline{X}} = 0.456$
   b. 0.0005
4.
5. a. 0.0918
   b. $\mu_{\overline{X}} = 25.6,\ \sigma_{\overline{X}} = 1.1$
   c. 0.0000
6.
7. a. $\mu_{\overline{X}} = 557,\ \sigma_{\overline{X}} = 4.9497$
   b. 0.0043
8.

9. a. 0.5818
   b. $\mu_{\overline{X}} = 1214$ $\sigma_{\overline{X}} = 24.4$
   c. 0.9998
10.
11. a. $\mu_{\overline{X}} = 72$ $\sigma_{\overline{X}} = 0.8944$
    b. 0.0250
12.
13. 0.9940
14.
15. 0.9994
16.
17. a. 0.8036
    b. 1.0000
18.
19. 0.9994
20.
21. a. 0.2955
    b. 0.8977
22.
23. 0.9251
24.
25. 0.9982

6.3: The Sample Proportion

Basic

1. The proportion of a population with a characteristic of interest is $p = 0.37$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size $1,600$.
2. The proportion of a population with a characteristic of interest is $p = 0.82$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size 900.
3. The proportion of a population with a characteristic of interest is $p = 0.76$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size $1,200$.
4. The proportion of a population with a characteristic of interest is $p = 0.37$. Find the mean and standard deviation of the sample proportion $\widehat{P}$ obtained from random samples of size 125.
5. Random samples of size 225 are drawn from a population in which the proportion with the characteristic of interest is 0.25. Decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
6. Random samples of size $1,600$ are drawn from a population in which the proportion with the characteristic of interest is 0.05. Decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
7. Random samples of size n produced sample proportions $\hat{p}$ as shown. In each case decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
   a. $n = 50,\ \hat{p} = 0.48$
   b. $n = 50,\ \hat{p} = 0.12$
   c. $n = 100,\ \hat{p} = 0.12$
8. Samples of size n produced sample proportions $\hat{p}$ as shown. In each case decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{P}$ is normally distributed.
   a. $n = 30,\ \hat{p} = 0.72$
   b. $n = 30,\ \hat{p} = 0.84$
   c. $n = 75,\ \hat{p} = 0.84$
9. A random sample of size 121 is taken from a population in which the proportion with the characteristic of interest is $p = 0.47$. Find the indicated probabilities.
   a. $P(0.45 \leq \widehat{P} \leq 0.50)$

b. $P(\widehat{P} \geq 0.50)$

10. A random sample of size 225 is taken from a population in which the proportion with the characteristic of interest is $p = 0.34$. Find the indicated probabilities.
    a. $P(0.25 \leq \widehat{P} \leq 0.40)$
    b. $P(\widehat{P} \geq 0.35)$
11. A random sample of size 900 is taken from a population in which the proportion with the characteristic of interest is p = 0.62. Find the indicated probabilities.
    a. $P(0.60 \leq \widehat{P} \leq 0.64)$
    b. $P(0.57 \leq \widehat{P} \leq 0.67)$
12. A random sample of size 1,100 is taken from a population in which the proportion with the characteristic of interest is p = 0.28. Find the indicated probabilities.
    1. $P(0.27 \leq \widehat{P} \leq 0.29)$
    2. $P(0.23 \leq \widehat{P} \leq 0.33)$

Applications

13. Suppose that 8% of all males suffer some form of color blindness. Find the probability that in a random sample of 250 men at least 10% will suffer some form of color blindness. First verify that the sample is sufficiently large to use the normal distribution.
14. Suppose that 29% of all residents of a community favor annexation by a nearby municipality. Find the probability that in a random sample of 50 residents at least 35% will favor annexation. First verify that the sample is sufficiently large to use the normal distribution.
15. Suppose that 2% of all cell phone connections by a certain provider are dropped. Find the probability that in a random sample of $1,500$ calls at most 40 will be dropped. First verify that the sample is sufficiently large to use the normal distribution.
16. Suppose that in 20% of all traffic accidents involving an injury, driver distraction in some form (for example, changing a radio station or texting) is a factor. Find the probability that in a random sample of 275 such accidents between 15% and 25% involve driver distraction in some form. First verify that the sample is sufficiently large to use the normal distribution.
17. An airline claims that 72% of all its flights to a certain region arrive on time. In a random sample of 30 recent arrivals, 19 were on time. You may assume that the normal distribution applies.
    a. Compute the sample proportion.
    b. Assuming the airline's claim is true, find the probability of a sample of size 30 producing a sample proportion so low as was observed in this sample.
18. A humane society reports that 19% of all pet dogs were adopted from an animal shelter. Assuming the truth of this assertion, find the probability that in a random sample of 80 pet dogs, between 15% and 20% were adopted from a shelter. You may assume that the normal distribution applies.
19. In one study it was found that 86% of all homes have a functional smoke detector. Suppose this proportion is valid for all homes. Find the probability that in a random sample of 600 homes, between 80% and 90% will have a functional smoke detector. You may assume that the normal distribution applies.
20. A state insurance commission estimates that 13% of all motorists in its state are uninsured. Suppose this proportion is valid. Find the probability that in a random sample of 50 motorists, at least 5 will be uninsured. You may assume that the normal distribution applies.
21. An outside financial auditor has observed that about 4% of all documents he examines contain an error of some sort. Assuming this proportion to be accurate, find the probability that a random sample of 700 documents will contain at least 30 with some sort of error. You may assume that the normal distribution applies.
22. Suppose 7% of all households have no home telephone but depend completely on cell phones. Find the probability that in a random sample of 450 households, between 25 and 35 will have no home telephone. You may assume that the normal distribution applies.

Additional Exercises

23. Some countries allow individual packages of prepackaged goods to weigh less than what is stated on the package, subject to certain conditions, such as the average of all packages being the stated weight or greater. Suppose that one requirement is that at most 4% of all packages marked 500 grams can weigh less than 490 grams. Assuming that a product actually meets this

requirement, find the probability that in a random sample of 150 such packages the proportion weighing less than 490 grams is at least 3%. You may assume that the normal distribution applies.

24. An economist wishes to investigate whether people are keeping cars longer now than in the past. He knows that five years ago, 38% of all passenger vehicles in operation were at least ten years old. He commissions a study in which 325 automobiles are randomly sampled. Of them, 132 are ten years old or older.
    a. Find the sample proportion.
    b. Find the probability that, when a sample of size 325 is drawn from a population in which the true proportion is 0.38, the sample proportion will be as large as the value you computed in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). Is there strong evidence that people are keeping their cars longer than was the case five years ago?
25. A state public health department wishes to investigate the effectiveness of a campaign against smoking. Historically 22% of all adults in the state regularly smoked cigars or cigarettes. In a survey commissioned by the public health department, 279 of $1,500$ randomly selected adults stated that they smoke regularly.
    a. Find the sample proportion.
    b. Find the probability that, when a sample of size $1,500$ is drawn from a population in which the true proportion is 0.22, the sample proportion will be no larger than the value you computed in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). How strong is the evidence that the campaign to reduce smoking has been effective?
26. In an effort to reduce the population of unwanted cats and dogs, a group of veterinarians set up a low-cost spay/neuter clinic. At the inception of the clinic a survey of pet owners indicated that 78% of all pet dogs and cats in the community were spayed or neutered. After the low-cost clinic had been in operation for three years, that figure had risen to 86%.
    a. What information is missing that you would need to compute the probability that a sample drawn from a population in which the proportion is 78% (corresponding to the assumption that the low-cost clinic had had no effect) is as high as 86%?
    b. Knowing that the size of the original sample three years ago was 150 and that the size of the recent sample was 125, compute the probability mentioned in part (a). You may assume that the normal distribution applies.
    c. Give an interpretation of the result in part (b). How strong is the evidence that the presence of the low-cost clinic has increased the proportion of pet dogs and cats that have been spayed or neutered?
27. An ordinary die is "fair" or "balanced" if each face has an equal chance of landing on top when the die is rolled. Thus the proportion of times a three is observed in a large number of tosses is expected to be close to $1/6$ or $0.1\overline{6}$. Suppose a die is rolled 240 times and shows three on top 36 times, for a sample proportion of 0.15.
    a. Find the probability that a fair die would produce a proportion of 0.15 or less. You may assume that the normal distribution applies.
    b. Give an interpretation of the result in part (b). How strong is the evidence that the die is not fair?
    c. Suppose the sample proportion 0.15 came from rolling the die $2,400$ times instead of only 240 times. Rework part (a) under these circumstances.
    d. Give an interpretation of the result in part (c). How strong is the evidence that the die is not fair?

Answers

1. $\mu_{\widehat{P}} = 0.37,\ \sigma_{\widehat{P}} = 0.012$
3. $\mu_{\widehat{P}} = 0.76,\ \sigma_{\widehat{P}} = 0.012$
5. $p \pm 3\sqrt{\frac{pq}{n}} = 0.25 \pm 0.087,$ yes
7. a. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.48 \pm 0.21,$ yes
   b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.12 \pm 0.14,$ no
   c. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.12 \pm 0.10,$ yes
9. a. 0.4154
   b. 0.2546

11. a. 0.7850
    b. 0.9980

13. $p \pm 3\sqrt{\frac{pq}{n}} = 0.08 \pm 0.05$ and $[0.03, 0.13] \subset [0, 1], 0.1210$

15. $p \pm 3\sqrt{\frac{pq}{n}} = 0.02 \pm 0.01$ and $[0.01, 0.03] \subset [0, 1], 0.9671$

17. a. 0.63
    b. 0.1446

19. 0.9977

21. 0.3483

23. 0.7357

25. a. 0.186
    b.0.0007

In a population in which the true proportion is 22% the chance that a random sample of size $1,500$ would produce a sample proportion of 18.6% or less is only $7/100$ of 1%. This is strong evidence that currently a smaller proportion than 22% smoke.

27. a. 0.2451

We would expect a sample proportion of 0.15 or less in about 24.5% of all samples of size 240, so this is practically no evidence at all that the die is not fair.

    b. 0.0139

We would expect a sample proportion of 0.15 or less in only about 1.4% of all samples of size $2,400$, so this is strong evidence that the die is not fair.

Contributor

- Anonymous

This page titled 6.E: Sampling Distributions (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **6.E: Sampling Distributions (Exercises)** has no license indicated.

CHAPTER OVERVIEW

7: Estimation

If we wish to estimate the mean μ of a population for which a census is impractical, say the average height of all 18-year-old men in the country, a reasonable strategy is to take a sample, compute its mean $\bar{x}$, and estimate the unknown number μ by the known number $\bar{x}$. For example, if the average height of 100 randomly selected men aged 18 is 70.6 inches, then we would say that the average height of all 18-year-old men is (at least approximately) 70.6 inches.

Estimating a population parameter by a single number like this is called point estimation; in the case at hand the statistic $\bar{x}$ is a point estimate of the parameter μ. The terminology arises because a single number corresponds to a single point on the number line.

A problem with a point estimate is that it gives no indication of how reliable the estimate is. In contrast, in this chapter we learn about interval estimation. In brief, in the case of estimating a population mean μ we use a formula to compute from the data a number E, called the margin of error of the estimate, and form the interval $[\bar{x}-E,\bar{x}+E]$. We do this in such a way that a certain proportion, say 95%, of all the intervals constructed from sample data by means of this formula contain the unknown parameter μ. Such an interval is called a 95% confidence interval for μ.

Continuing with the example of the average height of 18-year-old men, suppose that the sample of 100 men mentioned above for which $\bar{x}=70.6$ inches also had sample standard deviation $s = 1.7$ inches. It then turns out that $E = 0.33$ and we would state that we are 95% confident that the average height of all 18-year-old men is in the interval formed by 70.6 ± 0.33 inches, that is, the average is between 70.27 and 70.93 inches. If the sample statistics had come from a smaller sample, say a sample of 50 men, the lower reliability would show up in the 95% confidence interval being longer, hence less precise in its estimate. In this example the 95% confidence interval for the same sample statistics but with $n = 50$ is 70.6 ± 0.47 inches, or from 70.13 to 71.07 inches.

7.1: Large Sample Estimation of a Population Mean
7.2: Small Sample Estimation of a Population Mean
7.3: Large Sample Estimation of a Population Proportion
7.4: Sample Size Considerations
7.E: Estimation (Exercises)

This page titled 7: Estimation is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.1: Large Sample Estimation of a Population Mean

Learning Objectives

- To become familiar with the concept of an interval estimate of the population mean.
- To understand how to apply formulas for a confidence interval for a population mean.

The Central Limit Theorem says that, for large samples (samples of size $n \geq 30$), when viewed as a random variable the sample mean $\overline{X}$ is normally distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. The Empirical Rule says that we must go about two standard deviations from the mean to capture 95% of the values of $\overline{X}$ generated by sample after sample. A more precise distance based on the normality of $\overline{X}$ is 1.960 standard deviations, which is $E = \frac{1.960\sigma}{\sqrt{n}}$.

The key idea in the construction of the 95% confidence interval is this, as illustrated in Figure $7.1.1$, because in sample after sample 95% of the values of $\overline{X}$ lie in the interval $[\mu - E, \mu + E]$, if we adjoin to each side of the point estimate $x - a$ "wing" of length E, 95% of the intervals formed by the winged dots contain μ. The 95% confidence interval is thus $\bar{x} \pm 1.960\frac{\sigma}{\sqrt{n}}$. For a different level of confidence, say 90% or 99%, the number 1.960 will change, but the idea is the same.

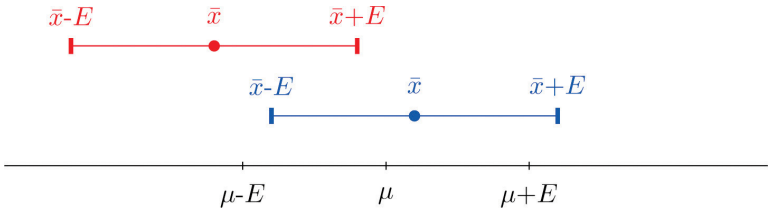


Figure $7.1.1$: When Winged Dots Capture the Population Mean

Figure $7.1.2$ shows the intervals generated by a computer simulation of drawing 40 samples from a normally distributed population and constructing the 95% confidence interval for each one. We expect that about $(0.05)(40) = 2$ of the intervals so constructed would fail to contain the population mean μ, and in this simulation two of the intervals, shown in red, do.

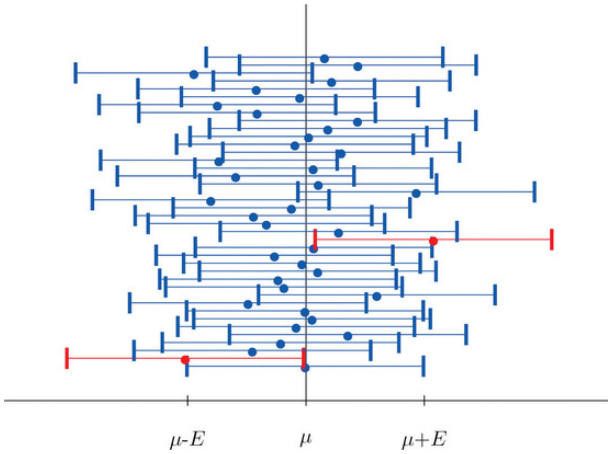


Figure $7.1.2$: Computer Simulation of 40 95% Confidence Intervals for a Mean

It is standard practice to identify the level of confidence in terms of the area α in the two tails of the distribution of $\overline{X}$ when the middle part specified by the level of confidence is taken out. This is shown in Figure $7.1.3$, drawn for the general situation, and in Figure $7.1.4$, drawn for 95% confidence.

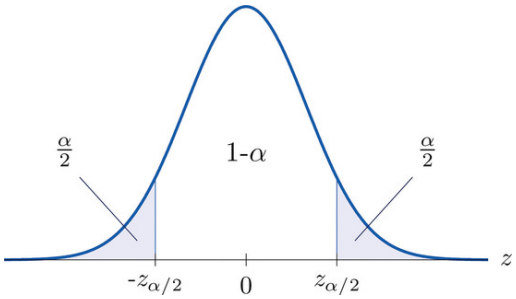


Figure 7.1.3: For $100(1-\alpha)\%$ confidence the area in each tail is $\alpha/2$.

Remember from Section 5.4 that the z-value that cuts off a right tail of area c is denoted z_c. Thus the number 1.960 in the example is $z_{.025}$, which is $z_{\frac{\alpha}{2}}$ for $\alpha = 1 - 0.95 = 0.05$.

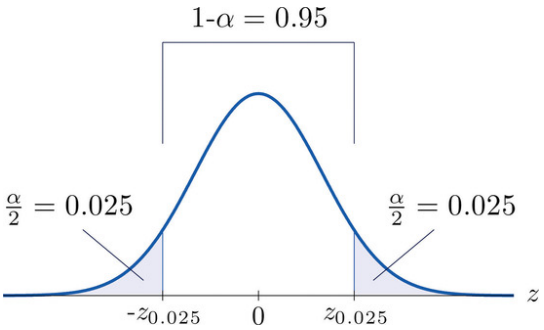


Figure 7.1.4: For 95% confidence the area in each tail is $\alpha/2 = 0.025$.

For 95% confidence the area in each tail is $\alpha/2 = 0.025$.

The level of confidence can be any number between 0 and 100%, but the most common values are probably 90% $(\alpha = 0.10)$, 95% $(\alpha = 0.05)$, and 99% $(\alpha = 0.01)$.

Thus in general for a $100(1-\alpha)\%$ confidence interval, $E = z_{\alpha/2}(\sigma/\sqrt{n})$, so the formula for the confidence interval is $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$. While sometimes the population standard deviation σ is known, typically it is not. If not, for $n \geq 30$ it is generally safe to approximate σ by the sample standard deviation s.

Large Sample $100(1-\alpha)\%$ Confidence Interval for a Population Mean

- If σ is known:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- If σ is unknown:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

A sample is considered large when $n \geq 30$.

As mentioned earlier, the number

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

or

$$E = z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

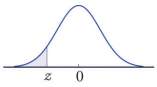
is called the *margin of error of the estimate*.

Example 7.1.1

Find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

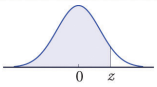
1. when the level of confidence is 90%;
2. when the level of confidence is 99%.

using the tables in Figure $7.1.5$ below.



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Cumulative Probability $P(Z \le z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8304	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 7.1.5: Cumulative Normal Probability

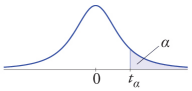
Solution:

1. For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. Since the area under the standard normal curve to the right of $z_{0.05}$ is 0.05, the area to the left of $z_{0.05}$ is 0.95. We search for the area 0.9500 in Figure 7.1.5. The closest entries in the table are 0.9495 and 0.9505, corresponding to z-values 1.64 and 1.65. Since 0.95 is halfway between 0.9495 and 0.9505 we use the average 1.645 of the z-values for $z_{0.05}$.
2. For confidence level 99%, $\alpha = 1 - 0.99 = 0.01$, so $z_{\alpha/2} = z_{0.005}$. Since the area under the standard normal curve to the right of $z_{0.005}$ is 0.005, the area to the left of $z_{0.005}$ is 0.9950. We search for the area 0.9950 in Figure 7.1.5. The closest entries in the table are 0.9949 and 0.9951, corresponding to z-values 2.57 and 2.58. Since 0.995 is halfway between 0.9949 and 0.9951 we use the average 2.575 of the z-values for $z_{0.005}$.

Example 7.1.2

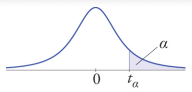
Use Figure 7.1.6 below to find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

1. when the level of confidence is 90%;
2. when the level of confidence is 99%.



Critical Values of t

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
1	1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.863	1.333	1.740	2.110	2.576	2.898	3.222	3.646	3.965
18	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.853	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.853	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.853	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.852	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.852	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.852	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.851	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.851	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.851	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.851	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
41	0.851	1.303	1.683	2.020	2.421	2.701	2.967	3.301	3.544
42	0.851	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
43	0.851	1.302	1.681	2.017	2.416	2.695	2.959	3.291	3.532
44	8.850	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
45	0.850	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
46	0.850	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
47	0.849	1.300	1.678	2.012	2.408	2.685	2.946	3.273	3.510
48	0.849	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
49	0.849	1.299	1.677	2.010	2.405	2.680	2.940	3.265	3.500
50	0.849	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496



Critical Values of t

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
51	0.849	1.298	1.675	2.008	2.402	2.676	2.934	3.258	3.492
52	0.849	1.298	1.675	2.007	2.400	2.674	2.932	3.255	3.488
53	0.849	1.298	1.674	2.006	2.399	2.672	2.929	3.251	3.484
54	0.848	1.297	1.674	2.005	2.397	2.670	2.927	3.248	3.480
55	0.848	1.297	1.673	2.004	2.396	2.668	2.925	3.245	3.476
56	0.848	1.297	1.673	2.003	2.395	2.667	2.923	3.242	3.473
57	0.848	1.297	1.672	2.002	2.394	2.665	2.920	3.239	3.470
58	0.848	1.296	1.672	2.002	2.392	2.663	2.918	3.237	3.466
59	0.848	1.296	1.671	2.001	2.391	2.662	2.916	3.234	3.463
60	0.848	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
61	0.848	1.296	1.670	2.000	2.389	2.659	2.913	3.229	3.457
62	0.848	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454

https://stats.libretexts.org/@go/page/563

63	0.847	1.295	1.669	1.998	2.387	2.656	2.909	3.225	3.452
64	0.847	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
65	0.847	1.295	1.669	1.997	2.385	2.654	2.906	3.220	3.447
66	0.847	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
67	0.847	1.294	1.668	1.996	2.383	2.651	2.903	3.216	3.442
68	0.847	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
69	0.847	1.294	1.667	1.995	2.382	2.649	2.900	3.213	3.437
70	0.847	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
71	0.847	1.294	1.667	1.994	2.380	2.647	2.897	3.209	3.433
72	0.847	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
73	0.847	1.293	1.666	1.993	2.379	2.645	2.895	3.206	3.429
74	0.847	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
75	0.846	1.293	1.665	1.992	2.377	2.643	2.892	3.202	3.425
76	0.846	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
77	0.846	1.293	1.665	1.991	2.376	2.641	2.890	3.199	3.421
78	0.846	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
79	0.846	1.292	1.664	1.990	2.374	2.640	2.888	3.197	3.418
80	0.846	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
81	0.846	1.292	1.664	1.990	2.373	2.638	2.886	3.194	3.415
82	0.846	1.292	1.664	1.989	2.373	2.637	2.885	3.193	3.413
83	0.846	1.292	1.663	1.989	2.372	2.636	2.884	3.191	3.412
84	0.846	1.292	1.663	1.989	2.372	2.636	2.883	3.190	3.410
85	0.846	1.292	1.663	1.988	2.371	2.635	2.882	3.189	3.409
86	0.846	1.291	1.663	1.988	2.370	2.634	2.881	3.188	3.407
87	0.846	1.291	1.663	1.988	2.370	2.634	2.880	3.187	3.406
88	0.846	1.291	1.662	1.987	2.369	2.633	2.880	3.185	3.405
89	0.846	1.291	1.662	1.987	2.369	2.632	2.879	3.184	3.403
90	0.846	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
91	0.846	1.291	1.662	1.986	2.368	2.631	2.877	3.182	3.401
92	0.846	1.291	1.662	1.986	2.368	2.630	2.876	3.181	3.399
93	0.846	1.291	1.661	1.986	2.367	2.630	2.876	3.180	3.398
94	0.846	1.291	1.661	1.986	2.367	2.629	2.875	3.179	3.397
95	0.845	1.291	1.661	1.985	2.366	2.629	2.874	3.178	3.396
96	0.845	1.290	1.661	1.985	2.366	2.628	2.873	3.177	3.395
97	0.845	1.290	1.661	1.985	2.365	2.627	2.873	3.176	3.394
98	0.845	1.290	1.661	1.984	2.365	2.627	2.872	3.175	3.393
99	0.845	1.290	1.660	1.984	2.365	2.626	2.871	3.175	3.392
100	0.845	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
∞ [z]	0.842	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure 7.1.6: Critical Values of t

Solution:

1. In the next section we will learn about a continuous random variable that has a probability distribution called the Student t-distribution. Figure 7.1.6 gives the value t_c that cuts off a right tail of area c for different values of c. The last line of that table, the one whose heading is the symbol ∞ for infinity and $[z]$, gives the corresponding z-value z_c that cuts off a right tail of the same area c. In particular, $z_{0.05}$ is the number in that row and in the column with the heading $t_{0.05}$. We read off directly that $z_{0.05} = 1.645$.
2. In Figure 7.1.6 $z_{0.005}$ is the number in the last row and in the column headed $t_{0.005}$, namely 2.576.

Figure 7.1.6 can be used to find z_c only for those values of c for which there is a column with the heading t_c appearing in the table; otherwise we must use Figure 7.1.5 in reverse. But when it can be done it is both faster and more accurate to use the last line of Figure 7.1.6 to find z_c than it is to do so using Figure 7.1.5 in reverse.

Example 7.1.3

A sample of size 49 has sample mean 35 and sample standard deviation 14. Construct a 98% confidence interval for the population mean using this information. Interpret its meaning.

Solution:

For confidence level 98%, $\alpha = 1 - 0.98 = 0.02$, so $z_{\alpha/2} = z_{0.01}$. From Figure 7.1.6 we read directly that $z_{0.01} = 2.326$.Thus

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 35 \pm 2.326\left(\frac{14}{\sqrt{49}}\right) = 35 \pm 4.652 \approx 35 \pm 4.7$$

We are 98% confident that the population mean μ lies in the interval $[30.3, 39.7]$, in the sense that in repeated sampling 98% of all intervals constructed from the sample data in this manner will contain μ.

Example 7.1.4

A random sample of 120 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

Solution:

For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. From Figure $7.1.6$ we read directly that $z_{0.05} = 1.645$. Since $n = 120$, $\bar{x} = 2.71$, and $s = 0.51$,

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 2.71 \pm 1.645\left(\frac{0.51}{\sqrt{120}}\right) = 2.71 \pm 0.0766$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval $(2.71 - 0.08, 2.71 + 0.08) = (2.63, 2.79)$.

- A confidence interval for a population mean is an estimate of the population mean together with an indication of reliability.
- There are different formulas for a confidence interval based on the sample size and whether or not the population standard deviation is known.
- The confidence intervals are constructed entirely from the sample data (or sample data and the population standard deviation, when it is known).

This page titled 7.1: Large Sample Estimation of a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.2: Small Sample Estimation of a Population Mean

Learning Objectives

1. To become familiar with Student's t-distribution.
2. To understand how to apply additional formulas for a confidence interval for a population mean.

The confidence interval formulas in the previous section are based on the Central Limit Theorem, the statement that for large samples $\overline{X}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. When the population mean μ is estimated with a small sample ($n < 30$), the Central Limit Theorem does not apply. In order to proceed we assume that the numerical population from which the sample is taken has a normal distribution to begin with. If this condition is satisfied then when the population standard deviation σ is known the old formula $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ can still be used to construct a $100(1-\alpha)\%$ confidence interval for μ.

If the population standard deviation is unknown and the sample size n is small then when we substitute the sample standard deviation s for σ the normal approximation is no longer valid. The solution is to use a different distribution, called Student's t-distribution with $n-1$ degrees of freedom. Student's t-distribution is very much like the standard normal distribution in that it is centered at 0 and has the same qualitative bell shape, but it has heavier tails than the standard normal distribution does, as indicated by Figure 7.2.1, in which the curve (in brown) that meets the dashed vertical line at the lowest point is the t-distribution with two degrees of freedom, the next curve (in blue) is the t-distribution with five degrees of freedom, and the thin curve (in red) is the standard normal distribution. As also indicated by the figure, as the sample size n increases, Student's t-distribution ever more closely resembles the standard normal distribution. Although there is a different t-distribution for every value of n, once the sample size is 30 or more it is typically acceptable to use the standard normal distribution instead, as we will always do in this text.

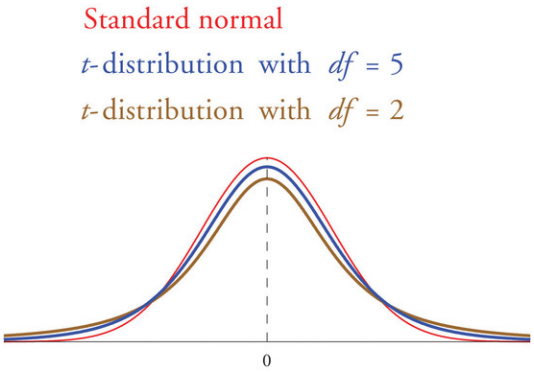


Figure 7.2.1: Student's *t-Distribution*

Just as the symbol z_c stands for the value that cuts off a right tail of area c in the standard normal distribution, so the symbol t_c stands for the value that cuts off a right tail of area c in the standard normal distribution. This gives us the following confidence interval formulas.

Small Sample $100(1-\alpha)\%$ Confidence Interval for a Population Mean

If σ is known:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

If σ is unknown:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.2.1}$$

with the degrees of freedom $df = n - 1$.

The population must be normally distributed and a sample is considered small when $n < 30$.

To use the new formula we use the line in Figure 7.1.6 that corresponds to the relevant sample size.

Example 7.2.1

A sample of size 15 drawn from a normally distributed population has sample mean 35 and sample standard deviation 14. Construct a 95% confidence interval for the population mean, and interpret its meaning.

Solution

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation 7.2.1.

Confidence level 95% means that

$$\alpha = 1 - 0.95 = 0.05$$

so $\alpha/2 = 0.025$. Since the sample size is $n = 15$, there are $n - 1 = 14$ degrees of freedom. By Figure 7.1.6 $t_{0.025} = 2.145$. Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.2.2}$$

$$= 35 \pm 2.145\left(\frac{14}{\sqrt{15}}\right) \tag{7.2.3}$$

$$= 35 \pm 7.8 \tag{7.2.4}$$

One may be 95% confident that the true value of μ is contained in the interval

$$(35 - 7.8, 35 + 7.8) = (27.2, 42.8).$$

Example 7.2.2

A random sample of 12 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution.

Solution

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is Equation 7.2.1

Confidence level 90% means that

$$\alpha = 1 - 0.90 = 0.10$$

so $\alpha/2 = 0.05$. Since the sample size is $n = 12$, there are $n - 1 = 11$ degrees of freedom. By Figure 7.1.6 $t_{0.05} = 1.796$. Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \tag{7.2.5}$$

$$= 2.71 \pm 1.796\left(\frac{0.51}{\sqrt{12}}\right) \tag{7.2.6}$$

$$= 2.71 \pm 0.26 \tag{7.2.7}$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval

$$(2.71 - 0.26, 2.71 + 0.26) = (2.45, 2.97).$$

Compare "Example 4" in Section 7.1 and "Example 6" in Section 7.1. The summary statistics in the two samples are the same, but the 90% confidence interval for the average GPA of all students at the university in "Example 4" in Section 7.1, $(2.63, 2.79)$, is shorter than the 90% confidence interval $(2.45, 2.97)$, in "Example 6" in Section 7.1. This is partly because in "Example 4" in Section 7.1 the sample size is larger; there is more information pertaining to the true value of μ in the large data set than in the small one.

Key Takeaway

- In selecting the correct formula for construction of a confidence interval for a population mean ask two questions: is the population standard deviation σ known or unknown, and is the sample large or small?

- We can construct confidence intervals with small samples only if the population is normal.

This page titled 7.2: Small Sample Estimation of a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.3: Large Sample Estimation of a Population Proportion

Learning Objectives

- To understand how to apply the formula for a confidence interval for a population proportion.

Since from Section 6.3, we know the mean, standard deviation, and sampling distribution of the sample proportion $\hat{p}$, the ideas of the previous two sections can be applied to produce a confidence interval for a population proportion. Here is the formula.

Large Sample $100(1-\alpha)\%$ Confidence Interval for a Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

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 the value of p is not known, hence neither is $\sigma_{\hat{p}}$. In that case we substitute the known quantity $\hat{p}$ for p in making the check; this means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lies wholly within the interval $[0,1]$.

Example 7.3.1

To estimate the proportion of students at a large college who are female, a random sample of 120 students is selected. There are 69 female students in the sample. Construct a 90% confidence interval for the proportion of all students at the college who are female.

Solution

The proportion of students in the sample who are female is

$$\hat{p} = 69/120 = 0.575$$

Confidence level 90% means that $\alpha = 1-0.90 = 0.10$ so $\alpha/2 = 0.05$. From the last line of Figure 7.1.6 we obtain $z_{0.05} = 1.645$.

Thus

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.575 \pm 1.645\sqrt{\frac{(0.575)(0.425)}{120}} = 0.575 \pm 0.074$$

One may be 90% confident that the true proportion of all students at the college who are female is contained in the interval $(0.575-0.074, 0.575+0.074) = (0.501, 0.649)$.

Summary

- We have a single formula for a confidence interval for a population proportion, which is valid when the sample is large.
- The condition that a sample be large is not that its size n be at least 30, but that the density function fit inside the interval $[0,1]$.

This page titled 7.3: Large Sample Estimation of a Population Proportion is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.4: Sample Size Considerations

Learning Objectives

- To learn how to apply formulas for estimating the size sample that will be needed in order to construct a confidence interval for a population mean or proportion that meets given criteria.

Sampling is typically done with a set of clear objectives in mind. For example, an economist might wish to estimate the mean yearly income of workers in a particular industry at 90% confidence and to within $\$500$. Since sampling costs time, effort, and money, it would be useful to be able to estimate the smallest size sample that is likely to meet these criteria.

Estimating μ

The confidence interval formulas for estimating a population mean μ have the form $\overline{x} \pm E$. When the population standard deviation σ is known,

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Mean

The estimated minimum sample size n needed to estimate a population mean μ to within E units at $100(1-\alpha)\%$ confidence is

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \text{ (rounded up)} \tag{7.4.1}$$

To apply Equation 7.4.1, we must have prior knowledge of the population in order to have an estimate of its standard deviation σ. In all the examples and exercises the population standard deviation will be given.

Example 7.4.1

Find the minimum sample size necessary to construct a 99% confidence interval for μ with a margin of error $E = 0.2$. Assume that the population standard deviation is $\sigma = 1.3$.

Solution

Confidence level 99% means that $\alpha = 1 - 0.99 = 0.01$ so $\alpha/2 = 0.005$. From the last line of Figure 7.1.6 we obtain $z_{0.005} = 2.576$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(2.576)^2(1.3)^2}{(0.2)^2} = 280.361536$$

which we round up to 281, since it is impossible to take a fractional observation.

Example 7.4.2

An economist wishes to estimate, with a 95% confidence interval, the yearly income of welders with at least five years experience to within $\$1,000$. He estimates that the range of incomes is no more than $\$24,000$, so using the Empirical Rule he estimates the population standard deviation to be about one-sixth as much, or about $\$4,000$. Find the estimated minimum sample size required.

Solution

Confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha/2 = 0.025$. From the last line of Figure 7.1.6 we obtain $z_{0.025} = 1.960$.

To say that the estimate is to be "to within $\$1,000$" means that $E = 1000$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.960)^2(4000)^2}{(1000)^2} = 61.4656$$

which we round up to 62.

Estimating p

The confidence interval formula for estimating a population proportion p is $\hat{p} \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the population proportion to within a certain number of percentage points means that we want the margin of error E to be no larger than that number (expressed as a proportion). Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Proportion

The estimated minimum sample size n needed to estimate a population proportion p to within E at $100(1-\alpha)\%$ confidence is

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2} (\text{rounded up})$$

There is a dilemma here: the formula for estimating how large a sample to take contains the number $\hat{p}$, which we know only after we have taken the sample. There are two ways out of this dilemma. Typically the researcher will have some idea as to the value of the population proportion p, hence of what the sample proportion $\hat{p}$ is likely to be. For example, if last month 37% of all voters thought that state taxes are too high, then it is likely that the proportion with that opinion this month will not be dramatically different, and we would use the value 0.37 for $\hat{p}$ in the formula.

The second approach to resolving the dilemma is simply to replace $\hat{p}$ in the formula by 0.5. This is because if $\hat{p}$ is large then $1-\hat{p}$ is small, and vice versa, which limits their product to a maximum value of 0.25, which occurs when $\hat{p} = 0.5$. This is called the **most conservative estimate**, since it gives the largest possible estimate of n.

Example 7.4.3

Find the necessary minimum sample size to construct a 98% confidence interval for p with a margin of error $E = 0.05$,

a. assuming that no prior knowledge about p is available; and
b. assuming that prior studies suggest that p is about 0.1.

Solution

Confidence level 98% means that $\alpha = 1 - 0.98 = 0.02$ so $\alpha/2 = 0.01$. From the last line of Figure 7.1.6 we obtain $z_{0.01} = 2.326$.

a. Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. Then

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2} = \frac{(2.326)^2(0.5)(1-0.5)}{0.05^2} = 541.0276$$

which we round up to 542.

b. Since $p \approx 0.1$ we estimate $\hat{p}$ by 0.1, and obtain

$$n = \frac{(z_{\alpha/2})^2\hat{p}(1-\hat{p})}{E^2} = \frac{(2.326)^2(0.1)(1-0.1)}{0.05^2} = 194.769936$$

which we round up to 195.

Example 7.4.4

A dermatologist wishes to estimate the proportion of young adults who apply sunscreen regularly before going out in the sun in the summer. Find the minimum sample size required to estimate the proportion to within three percentage points, at 90% confidence.

Solution

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha/2 = 0.05$. From the last line of Figure 7.1.6 we obtain $z_{0.05} = 1.645$.

Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. To estimate "to within three percentage points" means that $E = 0.03$. Then

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1-\hat{p})}{E^2} = \frac{(1.645)^2(0.5)(1-0.5)}{0.03^2} = 751.6736111$$

which we round up to 752.

Key Takeaway

- If the population standard deviation σ is known or can be estimated, then the minimum sample size needed to obtain a confidence interval for the population mean with a given maximum error of the estimate and a given level of confidence can be estimated.
- The minimum sample size needed to obtain a confidence interval for a population proportion with a given maximum error of the estimate and a given level of confidence can always be estimated. If there is prior knowledge of the population proportion p then the estimate can be sharpened.

This page titled 7.4: Sample Size Considerations is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

7.E: Estimation (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

7.1: Large Sample Estimation of a Population Mean

Basic

1. A random sample is drawn from a population of known standard deviation 11.3. Construct a 90% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 36,\ \bar{x} = 105.2,\ s = 11.2$
   b. $n = 100,\ \bar{x} = 105.2,\ s = 11.2$
2. A random sample is drawn from a population of known standard deviation 22.1. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 121,\ \bar{x} = 82.4,\ s = 21.9$
   b. $n = 81,\ \bar{x} = 82.4,\ s = 21.9$
3. A random sample is drawn from a population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.
   a. $n = 49,\ \bar{x} = 17.1,\ s = 2.1$
   b. $n = 169,\ \bar{x} = 17.1,\ s = 2.1$
4. A random sample is drawn from a population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.
   a. $n = 225,\ \bar{x} = 92.0,\ s = 8.4$
   b. $n = 64,\ \bar{x} = 92.0,\ s = 8.4$
5. A random sample of size 144 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\bar{x} = 58.2$ and $s = 2.6$.
   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.
6. A random sample of size 256 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\bar{x} = 1011$ and $s = 34$.
   a. Construct a 90% confidence interval for the population mean μ.
   b. Construct a 99% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

Applications

7. A government agency was charged by the legislature with estimating the length of time it takes citizens to fill out various forms. Two hundred randomly selected adults were timed as they filled out a particular form. The times required had mean 12.8 minutes with standard deviation 1.7 minutes. Construct a 90% confidence interval for the mean time taken for all adults to fill out this form.
8. Four hundred randomly selected working adults in a certain state, including those who worked at home, were asked the distance from their home to their workplace. The average distance was 8.84 miles with standard deviation 2.70 miles. Construct a 99% confidence interval for the mean distance from home to work for all residents of this state.
9. On every passenger vehicle that it tests an automotive magazine measures, at true speed 55 mph, the difference between the true speed of the vehicle and the speed indicated by the speedometer. For 36 vehicles tested the mean difference was -1.2 mph with standard deviation 0.2 mph. Construct a 90% confidence interval for the mean difference between true speed and indicated speed for all vehicles.
10. A corporation monitors time spent by office workers browsing the web on their computers instead of working. In a sample of computer records of 50 workers, the average amount of time spent browsing in an eight-hour work day was 27.8 minutes with standard deviation 8.2 minutes. Construct a 99.5% confidence interval for the mean time spent by all office workers in browsing the web in an eight-hour day.

11. A sample of 250 workers aged 16 and older produced an average length of time with the current employer ("job tenure") of 4.4 years with standard deviation 3.8 years. Construct a 99.9% confidence interval for the mean job tenure of all workers aged 16 or older.
12. The amount of a particular biochemical substance related to bone breakdown was measured in 30 healthy women. The sample mean and standard deviation were 3.3 nanograms per milliliter (ng/mL) and 1.4 ng/mL. Construct an 80% confidence interval for the mean level of this substance in all healthy women.
13. A corporation that owns apartment complexes wishes to estimate the average length of time residents remain in the same apartment before moving out. A sample of 150 rental contracts gave a mean length of occupancy of 3.7 years with standard deviation 1.2 years. Construct a 95% confidence interval for the mean length of occupancy of apartments owned by this corporation.
14. The designer of a garbage truck that lifts roll-out containers must estimate the mean weight the truck will lift at each collection point. A random sample of 325 containers of garbage on current collection routes yielded $\bar{x} = 75.3lb,\ s = 12.8lb$. Construct a 99.8% confidence interval for the mean weight the trucks must lift each time.
15. In order to estimate the mean amount of damage sustained by vehicles when a deer is struck, an insurance company examined the records of 50 such occurrences, and obtained a sample mean of $\$2,785$ with sample standard deviation $\$221$. Construct a 95% confidence interval for the mean amount of damage in all such accidents.
16. In order to estimate the mean FICO credit score of its members, a credit union samples the scores of 95 members, and obtains a sample mean of 738.2 with sample standard deviation 64.2. Construct a 99% confidence interval for the mean FICO score of all of its members.

Additional Exercises

17. For all settings a packing machine delivers a precise amount of liquid; the amount dispensed always has standard deviation 0.07 ounce. To calibrate the machine its setting is fixed and it is operated 50 times. The mean amount delivered is 6.02 ounces with sample standard deviation 0.04 ounce. Construct a 99.5% confidence interval for the mean amount delivered at this setting. Hint: Not all the information provided is needed.
18. A power wrench used on an assembly line applies a precise, preset amount of torque; the torque applied has standard deviation 0.73 foot-pound at every torque setting. To check that the wrench is operating within specifications it is used to tighten 100 fasteners. The mean torque applied is 36.95 foot-pounds with sample standard deviation 0.62 foot-pound. Construct a 99.9% confidence interval for the mean amount of torque applied by the wrench at this setting. Hint: Not all the information provided is needed.
19. The number of trips to a grocery store per week was recorded for a randomly selected collection of households, with the results shown in the table.

$$\begin{array}{cccccccccc} 2 & 2 & 2 & 1 & 4 & 2 & 3 & 2 & 5 & 4 \\ 2 & 3 & 5 & 0 & 3 & 2 & 3 & 1 & 4 & 3 \\ 3 & 2 & 1 & 6 & 2 & 3 & 3 & 2 & 4 & 4 \end{array} \tag{7.E.1}$$

    Construct a 95% confidence interval for the average number of trips to a grocery store per week of all households.
20. For each of 40 high school students in one county the number of days absent from school in the previous year were counted, with the results shown in the frequency table.

$$\begin{array}{c|cccccc} x & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline f & 24 & 7 & 5 & 2 & 1 & 1 \end{array} \tag{7.E.2}$$

    Construct a 90% confidence interval for the average number of days absent from school of all students in the county.
21. A town council commissioned a random sample of 85 households to estimate the number of four-wheel vehicles per household in the town. The results are shown in the following frequency table.

$$\begin{array}{c|cccccc} x & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline f & 1 & 16 & 28 & 22 & 12 & 6 \end{array} \tag{7.E.3}$$

    Construct a 98% confidence interval for the average number of four-wheel vehicles per household in the town.
22. The number of hours per day that a television set was operating was recorded for a randomly selected collection of households, with the results shown in the table.

$$\begin{array}{ccccc} 3.7 & 4.2 & 1.5 & 3.6 & 5.9 \\ 4.7 & 8.2 & 3.9 & 2.5 & 4.4 \end{array}$$

2.1	3.6	1.1	7.3	4.2
3.0	3.8	2.2	4.2	3.8
4.3	2.1	2.4	6.0	3.7
2.5	1.3	2.8	3.0	5.6

(7.E.4)

Construct a 99.8% confidence interval for the mean number of hours that a television set is in operation in all households.

Large Data Set Exercises

Large Data Set missing from the original

23. Large Data Set 1 records the SAT scores of $1,000$ students. Regarding it as a random sample of all high school students, use it to construct a 99% confidence interval for the mean SAT score of all students.
24. Large Data Set 1 records the GPAs of $1,000$ college students. Regarding it as a random sample of all college students, use it to construct a 95% confidence interval for the mean GPA of all students.
25. Large Data Set 1 lists the SAT scores of $1,000$ students.
    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 99% confidence for the mean μ of all $1,000$ SAT scores. Does it actually capture the mean μ?
26. Large Data Set 1 lists the GPAs of $1,000$ students.
    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 95% confidence for the mean μ of all $1,000$ GPAs. Does it actually capture the mean μ?

Answers

1. a. 105.2 ± 3.10
   b. 105.2 ± 1.86
2.
3. a. 17.1 ± 0.77
   b. 17.1 ± 0.42
4.
5. a. 58.2 ± 0.28
   b. 58.2 ± 0.36
   c. Asking for greater confidence requires a longer interval.
6.
7. 12.8 ± 0.20
8.
9. -1.2 ± 0.05
10.
11. 4.4 ± 0.79
12.
13. 3.7 ± 0.19
14.
15. 2785 ± 61
16.
17. 6.02 ± 0.03
18.
19. 2.8 ± 0.48
20.
21. 2.54 ± 0.30
22.

23. $(1511.43, 1546.05)$
24.
25. a. $\mu = 1528.74$
    b. $(1428.22, 1602.89)$

7.2: Small Sample Estimation of a Population Mean

Basic

1. A random sample is drawn from a normally distributed population of known standard deviation 5. Construct a 99.8% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 16,\ \bar{x} = 98,\ s = 5.6$
   b. $n = 9,\ \bar{x} = 98,\ s = 5.6$
2. A random sample is drawn from a normally distributed population of known standard deviation 10.7. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).
   a. $n = 25,\ \bar{x} = 103.3,\ s = 11.0$
   b. $n = 4,\ \bar{x} = 103.3,\ s = 11.0$
3. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.
   a. $n = 18,\ \bar{x} = 386,\ s = 24$
   b. $n = 7,\ \bar{x} = 386,\ s = 24$
4. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.
   a. $n = 8,\ \bar{x} = 58.3,\ s = 4.1$
   b. $n = 27,\ \bar{x} = 58.3,\ s = 4.1$
5. A random sample of size 14 is drawn from a normal population. The summary statistics are $\bar{x} = 933,\ and\ s = 18$.
   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.
6. A random sample of size 28 is drawn from a normal population. The summary statistics are $\bar{x} = 68.6,\ and\ s = 1.28$.
   a. Construct a 95% confidence interval for the population mean μ.
   b. Construct a 99.5% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

Application Exercises

7. City planners wish to estimate the mean lifetime of the most commonly planted trees in urban settings. A sample of 16 recently felled trees yielded mean age 32.7 years with standard deviation 3.1 years. Assuming the lifetimes of all such trees are normally distributed, construct a 99.8% confidence interval for the mean lifetime of all such trees.
8. To estimate the number of calories in a cup of diced chicken breast meat, the number of calories in a sample of four separate cups of meat is measured. The sample mean is 211.8 calories with sample standard deviation 0.9 calorie. Assuming the caloric content of all such chicken meat is normally distributed, construct a 95% confidence interval for the mean number of calories in one cup of meat.
9. A college athletic program wishes to estimate the average increase in the total weight an athlete can lift in three different lifts after following a particular training program for six weeks. Twenty-five randomly selected athletes when placed on the program exhibited a mean gain of 47.3 lb with standard deviation 6.4 lb. Construct a 90% confidence interval for the mean increase in lifting capacity all athletes would experience if placed on the training program. Assume increases among all athletes are normally distributed.
10. To test a new tread design with respect to stopping distance, a tire manufacturer manufactures a set of prototype tires and measures the stopping distance from 70 mph on a standard test car. A sample of 25 stopping distances yielded a sample mean 173 feet with sample standard deviation 8 feet. Construct a 98% confidence interval for the mean stopping distance for these tires. Assume a normal distribution of stopping distances.

11. A manufacturer of chokes for shotguns tests a choke by shooting 15 patterns at targets 40 yards away with a specified load of shot. The mean number of shot in a 30-inch circle is 53.5 with standard deviation 1.6. Construct an 80% confidence interval for the mean number of shot in a 30-inch circle at 40 yards for this choke with the specified load. Assume a normal distribution of the number of shot in a 30-inch circle at 40 yards for this choke.
12. In order to estimate the speaking vocabulary of three-year-old children in a particular socioeconomic class, a sociologist studies the speech of four children. The mean and standard deviation of the sample are $\bar{x} = 1120$ and $s = 215$ words. Assuming that speaking vocabularies are normally distributed, construct an 80% confidence interval for the mean speaking vocabulary of all three-year-old children in this socioeconomic group.
13. A thread manufacturer tests a sample of eight lengths of a certain type of thread made of blended materials and obtains a mean tensile strength of 8.2 lb with standard deviation 0.06 lb. Assuming tensile strengths are normally distributed, construct a 90% confidence interval for the mean tensile strength of this thread.
14. An airline wishes to estimate the weight of the paint on a fully painted aircraft of the type it flies. In a sample of four repaintings the average weight of the paint applied was 239 pounds, with sample standard deviation 8 pounds. Assuming that weights of paint on aircraft are normally distributed, construct a 99.8% confidence interval for the mean weight of paint on all such aircraft.
15. In a study of dummy foal syndrome, the average time between birth and onset of noticeable symptoms in a sample of six foals was 18.6 hours, with standard deviation 1.7 hours. Assuming that the time to onset of symptoms in all foals is normally distributed, construct a 90% confidence interval for the mean time between birth and onset of noticeable symptoms.
16. A sample of 26 women's size 6 dresses had mean waist measurement 25.25 inches with sample standard deviation 0.375 inch. Construct a 95% confidence interval for the mean waist measurement of all size 6 women's dresses. Assume waist measurements are normally distributed.

Additional Exercises

17. Botanists studying attrition among saplings in new growth areas of forests diligently counted stems in six plots in five-year-old new growth areas, obtaining the following counts of stems per acre:

$$\begin{array}{ccc} 9,432 & 11,026 & 10,539 \\ 8,773 & 9,868 & 10,247 \end{array} \tag{7.E.5}$$

    Construct an 80% confidence interval for the mean number of stems per acre in all five-year-old new growth areas of forests. Assume that the number of stems per acre is normally distributed.
18. Nutritionists are investigating the efficacy of a diet plan designed to increase the caloric intake of elderly people. The increase in daily caloric intake in 12 individuals who are put on the plan is (a minus sign signifies that calories consumed went down):

$$\begin{array}{cccccc} 121 & 284 & -94 & 295 & 183 & 312 \\ 188 & -102 & 259 & 226 & 152 & 167 \end{array} \tag{7.E.6}$$

    Construct a 99.8% confidence interval for the mean increase in caloric intake for all people who are put on this diet. Assume that population of differences in intake is normally distributed.
19. A machine for making precision cuts in dimension lumber produces studs with lengths that vary with standard deviation 0.003 inch. Five trial cuts are made to check the machine's calibration. The mean length of the studs produced is 104.998 inches with sample standard deviation 0.004 inch. Construct a 99.5% confidence interval for the mean lengths of all studs cut by this machine. Assume lengths are normally distributed. Hint: Not all the numbers given in the problem are used.
20. The variation in time for a baked good to go through a conveyor oven at a large scale bakery has standard deviation 0.017 minute at every time setting. To check the bake time of the oven periodically four batches of goods are carefully timed. The recent check gave a mean of 27.2 minutes with sample standard deviation 0.012 minute. Construct a 99.8% confidence interval for the mean bake time of all batches baked in this oven. Assume bake times are normally distributed. Hint: Not all the numbers given in the problem are used.
21. Wildlife researchers tranquilized and weighed three adult male polar bears. The data (in pounds) are: $926, 742, 1109$. Assume the weights of all bears are normally distributed.
    a. Construct an 80% confidence interval for the mean weight of all adult male polar bears using these data.
    b. Convert the three weights in pounds to weights in kilograms using the conversion $1\ lb = 0.453\ kg$ (so the first datum changes to $(926)(0.453) = 419$). Use the converted data to construct an 80% confidence interval for the mean weight of all adult male

polar bears expressed in kilograms.

c. Convert your answer in part (a) into kilograms directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.

22. Wildlife researchers trapped and measured six adult male collared lemmings. The data (in millimeters) are: $104, 99, 112, 115, 96, 109$. Assume the lengths of all lemmings are normally distributed.

a. Construct a 90% confidence interval for the mean length of all adult male collared lemmings using these data.

b. Convert the six lengths in millimeters to lengths in inches using the conversion $1\ mm = 0.039\ in$ (so the first datum changes to $(104)(0.039) = 4.06$). Use the converted data to construct a 90% confidence interval for the mean length of all adult male collared lemmings expressed in inches.

c. Convert your answer in part (a) into inches directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.

Answers

1. a. 98 ± 3.9
   b. 98 ± 5.2
2.
3. a. 386 ± 16.4
   b. 386 ± 33.6
4.
5. a. 933 ± 6.5
   b. 933 ± 8.5
   c. Asking for greater confidence requires a longer interval.
6.
7. 32.7 ± 2.9
8.
9. 47.3 ± 2.19
10.
11. 53.5 ± 0.56
12.
13. 8.2 ± 0.04
14.
15. 18.6 ± 1.4
16.
17. 9981 ± 486
18.
19. 104.998 ± 0.004
20.
21. a. 926 ± 200
    b. 419 ± 90
    c. 419 ± 91

7.3: Large Sample Estimation of a Population Proportion

Basic

1. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 90% confidence interval for the population proportion.

a. $n = 25, \hat{p} = 0.7$

b. $n = 50, \hat{p} = 0.7$

2. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 95% confidence interval for the population proportion.
   a. $n = 2500, \hat{p} = 0.22$
   b. $n = 1200, \hat{p} = 022$
3. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 98% confidence interval for the population proportion.
   a. $n = 80, \hat{p} = 0.4$
   b. $n = 325, \hat{p} = 0.4$
4. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 99.5% confidence interval for the population proportion.
   a. $n = 200, \hat{p} = 0.85$
   b. $n = 75, \hat{p} = 0.85$
5. In a random sample of size $1,100,338$ have the characteristic of interest.
   a. Compute the sample proportion $\hat{p}$ with the characteristic of interest.
   b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
   c. Construct an 80% confidence interval for the population proportion p.
   d. Construct a 90% confidence interval for the population proportion p.
   e. Comment on why one interval is longer than the other.
6. In a random sample of size $2,400,420$ have the characteristic of interest.
   a. Compute the sample proportion $\hat{p}$ with the characteristic of interest.
   b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
   c. Construct a 90% confidence interval for the population proportion p.
   d. Construct a 99% confidence interval for the population proportion p.
   e. Comment on why one interval is longer than the other.

Applications

Q7.3.7

A security feature on some web pages is graphic representations of words that are readable by human beings but not machines. When a certain design format was tested on 450 subjects, by having them attempt to read ten disguised words, 448 subjects could read all the words.

a. Give a point estimate of the proportion p of all people who could read words disguised in this way.
b. Show that the sample is not sufficiently large to construct a confidence interval for the proportion of all people who could read words disguised in this way.

Q7.3.8

In a random sample of 900 adults, 42 defined themselves as vegetarians.

a. Give a point estimate of the proportion of all adults who would define themselves as vegetarians.
b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct an 80% confidence interval for the proportion of all adults who would define themselves as vegetarians.

Q7.3.9

In a random sample of 250 employed people, 61 said that they bring work home with them at least occasionally.

a. Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally.
b. Construct a 99% confidence interval for that proportion.

Q7.3.10

In a random sample of $1,250$ household moves, 822 were moves to a location within the same county as the original residence.

a. Give a point estimate of the proportion of all household moves that are to a location within the same county as the original residence.

b. Construct a 98% confidence interval for that proportion.

Q7.3.11

In a random sample of $12,447$ hip replacement or revision surgery procedures nationwide, 162 patients developed a surgical site infection.

a. Give a point estimate of the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.
b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct a 95% confidence interval for the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.

Q7.3.12

In a certain region prepackaged products labeled 500 g must contain on average at least 500 grams of the product, and at least 90% of all packages must weigh at least 490 grams. In a random sample of 300 packages, 288 weighed at least 490 grams.

a. Give a point estimate of the proportion of all packages that weigh at least 490 grams.
b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
c. Construct a 99.8% confidence interval for the proportion of all packages that weigh at least 490 grams.

Q7.3.13

A survey of 50 randomly selected adults in a small town asked them if their opinion on a proposed "no cruising" restriction late at night. Responses were coded 1 for in favor, 0 for indifferent, and 2 for opposed, with the results shown in the table.

$$\begin{array}{cccccccccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 1 & 1 & 2 \\ 0 & 2 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 2 & 1 & 2 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 & 1 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 & 2 \end{array} \tag{7.E.7}$$

a. Give a point estimate of the proportion of all adults in the community who are indifferent concerning the proposed restriction.
b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all adults in the community who are indifferent concerning the proposed restriction.

Q7.3.14

To try to understand the reason for returned goods, the manager of a store examines the records on 40 products that were returned in the last year. Reasons were coded by 1 for "defective," 2 for "unsatisfactory," and 0 for all other reasons, with the results shown in the table.

$$\begin{array}{cccccccccc} 0 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array} \tag{7.E.8}$$

a. Give a point estimate of the proportion of all returns that are because of something wrong with the product, that is, either defective or performed unsatisfactorily.
b. Assuming that the sample is sufficiently large, construct an 80% confidence interval for the proportion of all returns that are because of something wrong with the product.

Q7.3.15

In order to estimate the proportion of entering students who graduate within six years, the administration at a state university examined the records of 600 randomly selected students who entered the university six years ago, and found that 312 had graduated.

a. Give a point estimate of the six-year graduation rate, the proportion of entering students who graduate within six years.
b. Assuming that the sample is sufficiently large, construct a 98% confidence interval for the six-year graduation rate.

Q7.3.16

In a random sample of $2,300$ mortgages taken out in a certain region last year, 187 were adjustable-rate mortgages.

a. Give a point estimate of the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.
b. Assuming that the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.

Q7.3.17

In a research study in cattle breeding, 159 of 273 cows in several herds that were in estrus were detected by means of an intensive once a day, one-hour observation of the herds in early morning.

a. Give a point estimate of the proportion of all cattle in estrus who are detected by this method.
b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all cattle in estrus who are detected by this method.

Q7.3.18

A survey of $21,250$ households concerning telephone service gave the results shown in the table.

	Landline	No Landline
Cell phone	12,474	5,844
No cell phone	2,529	403

a. Give a point estimate for the proportion of all households in which there is a cell phone but no landline.
b. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all households in which there is a cell phone but no landline.
c. Give a point estimate for the proportion of all households in which there is no telephone service of either kind.
d. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all all households in which there is no telephone service of either kind.

Additional Exercises

19. In a random sample of 900 adults, 42 defined themselves as vegetarians. Of these 42, 29 were women.
    a. Give a point estimate of the proportion of all self-described vegetarians who are women.
    b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
    c. Construct a 90% confidence interval for the proportion of all all self-described vegetarians who are women.
20. A random sample of 185 college soccer players who had suffered injuries that resulted in loss of playing time was made with the results shown in the table. Injuries are classified according to severity of the injury and the condition under which it was sustained.

	Minor	Moderate	Serious
Practice	48	20	6
Game	62	32	17

    a. Give a point estimate for the proportion p of all injuries to college soccer players that are sustained in practice.
    b. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are sustained in practice.
    c. Give a point estimate for the proportion p of all injuries to college soccer players that are either moderate or serious.
    d. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are either moderate or serious.
21. The body mass index (BMI) was measured in $1,200$ randomly selected adults, with the results shown in the table.

	BMI		
	Under 18.5	18.5–25	Over 25
Men	36	165	315
Women	75	274	335

a. Give a point estimate for the proportion of all men whose BMI is over 25.
b. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all men whose BMI is over 25.
c. Give a point estimate for the proportion of all adults, regardless of gender, whose BMI is over 25.
d. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all adults, regardless of gender, whose BMI is over 25.

22. Confidence intervals constructed using the formula in this section often do not do as well as expected unless n is quite large, especially when the true population proportion is close to either 0 or 1. In such cases a better result is obtained by adding two successes and two failures to the actual data and then computing the confidence interval. This is the same as using the formula

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \qquad \text{(7.E.9)}$$

where

$$\tilde{p} = \frac{x+2}{n+4} \text{ and } \tilde{n} = n+4$$

Suppose that in a random sample of 600 households, 12 had no telephone service of any kind. Use the adjusted confidence interval procedure just described to form a 99.9% confidence interval for the proportion of all households that have no telephone service of any kind.

Large Data Set Exercises

Large Data Set missing from the original

23. Large Data Sets 4 and 4A list the results of 500 tosses of a die. Let p denote the proportion of all tosses of this die that would result in a four. Use the sample data to construct a 90% confidence interval for p.
24. Large Data Set 6 records results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Use the full data set (400 observations) to construct a 98% confidence interval for the proportion p of all voters who prefer Candidate A.
25. Lines 2 through 536 in Data Set 11 is a sample of 535 real estate sales in a certain region in 2008. Those that were foreclosure sales are identified with a 1 in the second column.
    a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2008 that were foreclosure sales.
    b. Use these data to construct a 90% confidence for p.
26. Lines 537 through 1106 in Large Data Set 11 is a sample of 570 real estate sales in a certain region in 2010. Those that were foreclosure sales are identified with a 1 in the second column.
    a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2010 that were foreclosure sales.
    b. Use these data to construct a 90% confidence for p.

Answers

1. a. $(0.5492, 0.8508)$
   b. $(0.5934, 0.8066)$
2.
3. a. $(0.2726, 0.5274)$
   b. $(0.3368, 0.4632)$
4.
5. a. 0.3073
   b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.31 \pm 0.04$ and $[0.27, 0.35] \subset [0, 1]$
   c. $(0.2895, 0.3251)$
   d. $(0.2844, 0.3302)$
   e. Asking for greater confidence requires a longer interval.
6.

7. a. 0.9956
   b. $(0.9862, 1.005)$
8.
9. a. 0.244
   b. $(0.1740, 0.3140)$
10.
11. a. 0.013
    b. $(0.01, 0.016)$
    c. $(0.011, 0.015)$
12.
13. a. 0.52
    b. $(0.4038, 0.6362)$
14.
15. a. 0.52
    b. $(0.4726, 0.5674)$
16.
17. a. 0.5824
    b. $(0.5333, 0.6315)$
18.
19. a. 0.69
    b. $\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.69 \pm 0.21$ and $[0.48, 0.90] \subset [0, 1]$
    c. 0.69 ± 0.12
20.
21. a. 0.6105
    b. $(0.5552, 0.6658)$
    c. 0.5583
    d. $(0.5214, 0.5952)$
22.
23. $(0.1368, 0.1912)$
24.
25. a. $\hat{p} = 0.2280$
    b. $(0.1982, 0.2579)$

7.4: Sample Size Considerations

Basic

1. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.
   a. $\sigma = 30, 95\%$ confidence, $E = 10$
   b. $\sigma = 30, 99\%$ confidence, $E = 10$
   c. $\sigma = 30, 95\%$ confidence, $E = 5$
2. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.
   a. $\sigma = 4, 95\%$ confidence, $E = 1$
   b. $\sigma = 4, 99\%$ confidence, $E = 1$
   c. $\sigma = 4, 95\%$ confidence, $E = 0.5$
3. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. $p \approx 0.37, 80\%$ confidence, $E = 0.05$

b. $p \approx 0.37, 90\%$ confidence, $E = 0.05$
c. $p \approx 0.37, 80\%$ confidence, $E = 0.01$

4. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. $p \approx 0.81, 95\%$ confidence, $E = 0.02$
   b. $p \approx 0.81, 99\%$ confidence, $E = 0.02$
   c. $p \approx 0.81, 95\%$ confidence, $E = 0.01$
5. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. 80% confidence, $E = 0.05$
   b. 90% confidence, $E = 0.05$
   c. 80% confidence, $E = 0.01$
6. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.
   a. 95% confidence, $E = 0.02$
   b. 99% confidence, $E = 0.02$
   c. 95% confidence, $E = 0.01$

Applications

7. A software engineer wishes to estimate, to within 5 seconds, the mean time that a new application takes to start up, with 95% confidence. Estimate the minimum size sample required if the standard deviation of start up times for similar software is 12 seconds.
8. A real estate agent wishes to estimate, to within $\$2.50$, the mean retail cost per square foot of newly built homes, with 80% confidence. He estimates the standard deviation of such costs at $\$5.00$. Estimate the minimum size sample required.
9. An economist wishes to estimate, to within 2 minutes, the mean time that employed persons spend commuting each day, with 95% confidence. On the assumption that the standard deviation of commuting times is 8 minutes, estimate the minimum size sample required.
10. A motor club wishes to estimate, to within 1 cent, the mean price of 1 gallon of regular gasoline in a certain region, with 98% confidence. Historically the variability of prices is measured by $\sigma = \$0.03$Estimate the minimum size sample required.
11. A bank wishes to estimate, to within $\$25$, the mean average monthly balance in its checking accounts, with 99.8% confidence. Assuming $\sigma = \$250$, estimate the minimum size sample required.
12. A retailer wishes to estimate, to within 15 seconds, the mean duration of telephone orders taken at its call center, with 99.5% confidence. In the past the standard deviation of call length has been about 1.25 minutes. Estimate the minimum size sample required. (Be careful to express all the information in the same units.)
13. The administration at a college wishes to estimate, to within two percentage points, the proportion of all its entering freshmen who graduate within four years, with 90% confidence. Estimate the minimum size sample required.
14. A chain of automotive repair stores wishes to estimate, to within five percentage points, the proportion of all passenger vehicles in operation that are at least five years old, with 98% confidence. Estimate the minimum size sample required.
15. An internet service provider wishes to estimate, to within one percentage point, the current proportion of all email that is spam, with 99.9% confidence. Last year the proportion that was spam was 71%. Estimate the minimum size sample required.
16. An agronomist wishes to estimate, to within one percentage point, the proportion of a new variety of seed that will germinate when planted, with 95% confidence. A typical germination rate is 97%. Estimate the minimum size sample required.
17. A charitable organization wishes to estimate, to within half a percentage point, the proportion of all telephone solicitations to its donors that result in a gift, with 90% confidence. Estimate the minimum sample size required, using the information that in the past the response rate has been about 30%.
18. A government agency wishes to estimate the proportion of drivers aged $16 - 24$ who have been involved in a traffic accident in the last year. It wishes to make the estimate to within one percentage point and at 90% confidence. Find the minimum sample size required, using the information that several years ago the proportion was 0.12.

Additional Exercises

19. An economist wishes to estimate, to within six months, the mean time between sales of existing homes, with 95% confidence. Estimate the minimum size sample required. In his experience virtually all houses are re-sold within 40 months, so using the

Empirical Rule he will estimate σ by one-sixth the range, or $40/6 = 6.7$.

20. A wildlife manager wishes to estimate the mean length of fish in a large lake, to within one inch, with 80% confidence. Estimate the minimum size sample required. In his experience virtually no fish caught in the lake is over 23 inches long, so using the Empirical Rule he will estimate σ by one-sixth the range, or $23/6 = 3.8$.
21. You wish to estimate the current mean birth weight of all newborns in a certain region, to within 1 ounce ($1/16$ pound) and with 95% confidence. A sample will cost $\$400$ plus $\$1.50$ for every newborn weighed. You believe the standard deviations of weight to be no more than 1.25 pounds. You have $\$2,500$ to spend on the study.
    a. Can you afford the sample required?
    b. If not, what are your options?
22. You wish to estimate a population proportion to within three percentage points, at 95% confidence. A sample will cost $\$500$ plus 50 cents for every sample element measured. You have $\$1,000$ to spend on the study.
    a. Can you afford the sample required?
    b. If not, what are your options?

Answers

1. a. 35
   b. 60
   c. 139
2.
3. a. 154
   b. 253
   c. 3832
4.
5. a. 165
   b. 271
   c. 4109
6.
7. 23
8.
9. 62
10.
11. 955
12.
13. 1692
14.
15. $22,301$
16.
17. $22,731$
18.
19. 5
20.
21. a. no
    b. decrease the confidence level

This page titled 7.E: Estimation (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **7.E: Estimation (Exercises)** has no license indicated.

CHAPTER OVERVIEW

8: Testing Hypotheses

In the sampling that we have studied so far the goal has been to estimate a population parameter. But the sampling done by the government agency has a somewhat different objective, not so much to *estimate* the population mean μ as to *test* an assertion—or a hypothesis—about it, namely, whether it is as large as 75 or not. The agency is not necessarily interested in the actual value of μ, just whether it is as claimed. Their sampling is done to perform a test of hypotheses, the subject of this chapter.

8.1: The Elements of Hypothesis Testing
8.2: Large Sample Tests for a Population Mean
8.3: The Observed Significance of a Test
8.4: Small Sample Tests for a Population Mean
8.5: Large Sample Tests for a Population Proportion
8.E: Testing Hypotheses (Exercises)

This page titled 8: Testing Hypotheses is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.1: The Elements of Hypothesis Testing

Learning Objectives

- To understand the logical framework of tests of hypotheses.
- To learn basic terminology connected with hypothesis testing.
- To learn fundamental facts about hypothesis testing.

Types of Hypotheses

A *hypothesis* about the value of a population parameter is an assertion about its value. As in the introductory example we will be concerned with testing the truth of two competing hypotheses, only one of which can be true.

Definition: null hypothesis and alternative hypothesis

- The *null hypothesis*, denoted H_0, is the statement about the population parameter that is assumed to be true unless there is convincing evidence to the contrary.
- The *alternative hypothesis*, denoted H_a, is a statement about the population parameter that is contradictory to the null hypothesis, and is accepted as true only if there is convincing evidence in favor of it.

Definition: statistical procedure

Hypothesis testing is a *statistical procedure* in which a choice is made between a null hypothesis and an alternative hypothesis based on information in a sample.

The end result of a hypotheses testing procedure is a choice of one of the following two possible conclusions:

1. Reject H_0 (and therefore accept H_a), or
2. Fail to reject H_0 (and therefore fail to accept H_a).

The null hypothesis typically represents the status quo, or what has historically been true. In the example of the respirators, we would believe the claim of the manufacturer unless there is reason not to do so, so the null hypotheses is $H_0 : \mu = 75$. The alternative hypothesis in the example is the contradictory statement $H_a : \mu < 75$. The null hypothesis will always be an assertion containing an equals sign, but depending on the situation the alternative hypothesis can have any one of three forms: with the symbol $<$, as in the example just discussed, with the symbol $>$, or with the symbol $\neq$. The following two examples illustrate the latter two cases.

Example 8.1.1

A publisher of college textbooks claims that the average price of all hardbound college textbooks is $\$127.50$. A student group believes that the actual mean is higher and wishes to test their belief. State the relevant null and alternative hypotheses.

Solution

The default option is to accept the publisher's claim unless there is compelling evidence to the contrary. Thus the null hypothesis is $H_0 : \mu = 127.50$. Since the student group thinks that the average textbook price is greater than the publisher's figure, the alternative hypothesis in this situation is $H_a : \mu > 127.50$.

Example 8.1.2

The recipe for a bakery item is designed to result in a product that contains 8 grams of fat per serving. The quality control department samples the product periodically to insure that the production process is working as designed. State the relevant null and alternative hypotheses.

Solution

The default option is to assume that the product contains the amount of fat it was formulated to contain unless there is compelling evidence to the contrary. Thus the null hypothesis is $H_0 : \mu = 8.0$. Since to contain either more fat than desired or to

contain less fat than desired are both an indication of a faulty production process, the alternative hypothesis in this situation is that the mean is different from 8.0, so $H_a : \mu \neq 8.0$.

In Example 8.1.1, the textbook example, it might seem more natural that the publisher's claim be that the average price is at most \$127.50, not exactly \$127.50. If the claim were made this way, then the null hypothesis would be $H_0 : \mu \leq 127.50$, and the value \$127.50 given in the example would be the one that is least favorable to the publisher's claim, the null hypothesis. It is always true that if the null hypothesis is retained for its least favorable value, then it is retained for every other value.

Thus in order to make the null and alternative hypotheses easy for the student to distinguish, in every example and problem in this text we will always present one of the two competing claims about the value of a parameter with an equality. The claim expressed with an equality is the null hypothesis. This is the same as always stating the null hypothesis in the least favorable light. So in the introductory example about the respirators, we stated the manufacturer's claim as "the average is 75 minutes" instead of the perhaps more natural "the average is at least 75 minutes," essentially reducing the presentation of the null hypothesis to its worst case.

The first step in hypothesis testing is to identify the null and alternative hypotheses.

The Logic of Hypothesis Testing

Although we will study hypothesis testing in situations other than for a single population mean (for example, for a population proportion instead of a mean or in comparing the means of two different populations), in this section the discussion will always be given in terms of a single population mean μ.

The null hypothesis always has the form $H_0 : \mu = \mu_0$ for a specific number μ_0 (in the respirator example $\mu_0 = 75$, in the textbook example $\mu_0 = 127.50$, and in the baked goods example $\mu_0 = 8.0$). Since the null hypothesis is accepted unless there is strong evidence to the contrary, the test procedure is based on the initial assumption that H_0 is true. This point is so important that we will repeat it in a display:

> The test procedure is based on the initial assumption that H_0 is true.

The criterion for judging between H_0 and H_a based on the sample data is: if the value of $\overline{X}$ would be highly unlikely to occur if H_0 were true, but favors the truth of H_a, then we reject H_0 in favor of H_a. Otherwise we do not reject H_0.

Supposing for now that $\overline{X}$ follows a normal distribution, when the null hypothesis is true the density function for the sample mean $\overline{X}$ must be as in Figure 8.1.1: a bell curve centered at μ_0. Thus if H_0 is true then $\overline{X}$ is likely to take a value near μ_0 and is unlikely to take values far away. Our decision procedure therefore reduces simply to:

- if H_a has the form $H_a : \mu < \mu_0$ then reject H_0 if $\bar{x}$ is far to the left of μ_0;
- if H_a has the form $H_a : \mu > \mu_0$ then reject H_0 if $\bar{x}$ is far to the right of μ_0;
- if H_a has the form $H_a : \mu \neq \mu_0$ then reject H_0 if $\bar{x}$ is far away from μ_0 in either direction.

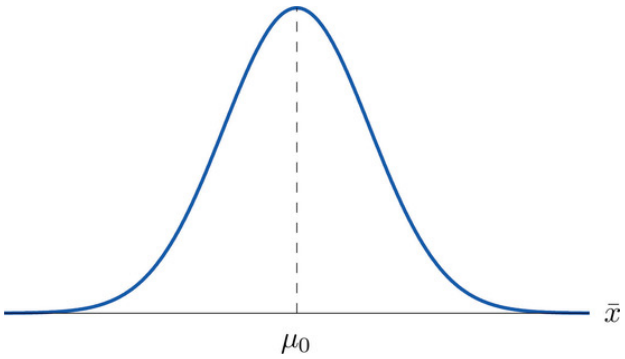


Figure 8.1.1 : The Density Curve for $\overline{X}$ *if* H_0 *Is True*

Think of the respirator example, for which the null hypothesis is $H_0 : \mu = 75$, the claim that the average time air is delivered for all respirators is 75 minutes. If the sample mean is 75 or greater then we certainly would not reject H_0 (since there is no issue with an emergency respirator delivering air even longer than claimed).

If the sample mean is slightly less than 75 then we would logically attribute the difference to sampling error and also not reject H_0 either.

Values of the sample mean that are smaller and smaller are less and less likely to come from a population for which the population mean is 75. Thus if the sample mean is far less than 75, say around 60 minutes or less, then we would certainly reject H_0, because we know that it is highly unlikely that the average of a sample would be so low if the population mean were 75. This is the rare event criterion for rejection: what we actually observed ($\overline{X} < 60$) would be so rare an event if $\mu = 75$ were true that we regard it as much more likely that the alternative hypothesis $\mu < 75$ holds.

In summary, to decide between H_0 and H_a in this example we would select a "rejection region" of values sufficiently far to the left of 75, based on the rare event criterion, and reject H_0 if the sample mean $\overline{X}$ lies in the rejection region, but not reject H_0 if it does not.

The Rejection Region

Each different form of the alternative hypothesis Ha has its own kind of rejection region:

- if (as in the respirator example) H_a has the form $H_a : \mu < \mu_0$, we reject H_0 if $\bar{x}$ is far to the left of μ_0, that is, to the left of some number C, so the rejection region has the form of an interval $(-\infty, C]$;
- if (as in the textbook example) H_a has the form $H_a : \mu > \mu_0$, we reject H_0 if $\bar{x}$ is far to the right of μ_0, that is, to the right of some number C, so the rejection region has the form of an interval $[C, \infty)$;
- if (as in the baked good example) H_a has the form $H_a : \mu \neq \mu_0$, we reject H_0 if $\bar{x}$ is far away from μ_0 in either direction, that is, either to the left of some number C or to the right of some other number C', so the rejection region has the form of the union of two intervals $(-\infty, C] \cup [C', \infty)$.

The key issue in our line of reasoning is the question of how to determine the number C or numbers C and C', called the critical value or critical values of the statistic, that determine the rejection region.

Definition: critical values

The critical value or critical values of a test of hypotheses are the number or numbers that determine the rejection region.

Suppose the rejection region is a single interval, so we need to select a single number C. Here is the procedure for doing so. We select a small probability, denoted α, say 1%, which we take as our definition of "rare event:" an event is "rare" if its probability of occurrence is less than α. (In all the examples and problems in this text the value of α will be given already.) The probability that $\overline{X}$ takes a value in an interval is the area under its density curve and above that interval, so as shown in Figure $8.1.2$ (drawn under the assumption that H_0 is true, so that the curve centers at μ_0) the critical value C is the value of $\overline{X}$ that cuts off a tail area α in the probability density curve of $\overline{X}$. When the rejection region is in two pieces, that is, composed of two intervals, the total area above both of them must be α, so the area above each one is $\alpha/2$, as also shown in Figure $8.1.2$.

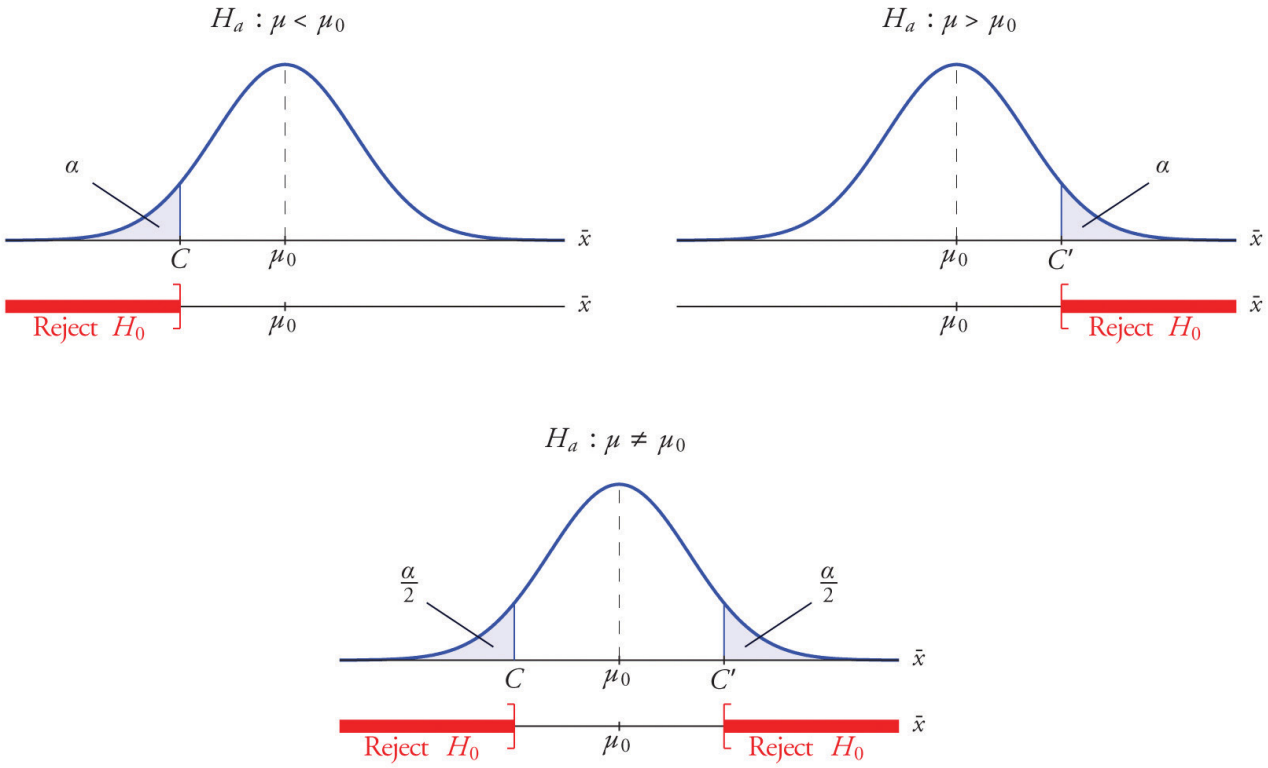


Figure 8.1.2

The number α is the total area of a tail or a pair of tails.

Example 8.1.3

In the context of Example 8.1.2, suppose that it is known that the population is normally distributed with standard deviation $\alpha = 0.15$ gram, and suppose that the test of hypotheses $H_0 : \mu = 8.0$ versus $H_a : \mu \neq 8.0$ will be performed with a sample of size 5. Construct the rejection region for the test for the choice $\alpha = 0.10$. Explain the decision procedure and interpret it.

Solution

If H_0 is true then the sample mean $\overline{X}$ is normally distributed with mean and standard deviation

$$\begin{aligned} \mu_{\overline{X}} &= \mu \\ &= 8.0 \end{aligned}$$

$$\begin{aligned} \sigma_{\overline{X}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{0.15}{\sqrt{5}} \\ &= 0.067 \end{aligned}$$

Since H_a contains the $\neq$ symbol the rejection region will be in two pieces, each one corresponding to a tail of area $\alpha/2 = 0.10/2 = 0.05$. From Figure 7.1.6, $z_{0.05} = 1.645$, so C and C' are 1.645 standard deviations of $\overline{X}$ to the right and left of its mean 8.0:

$$C = 8.0 - (1.645)(0.067) = 7.89 \text{ and } C' = 8.0 + (1.645)(0.067) = 8.11$$

The result is shown in Figure 8.1.3.

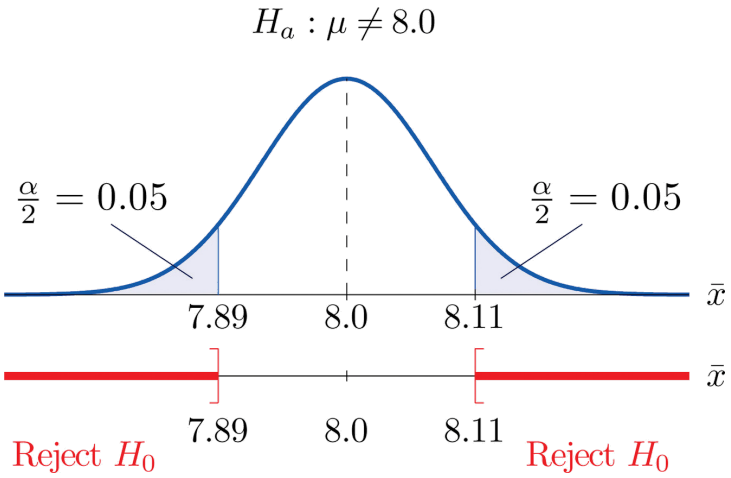


Figure 8.1.3: Rejection Region for the Choice

The decision procedure is: take a sample of size 5 and compute the sample mean $\bar{x}$. If $\bar{x}$ is either 7.89 grams or less or 8.11 grams or more then reject the hypothesis that the average amount of fat in all servings of the product is 8.0 grams in favor of the alternative that it is different from 8.0 grams. Otherwise do not reject the hypothesis that the average amount is 8.0 grams.

The reasoning is that if the true average amount of fat per serving were 8.0 grams then there would be less than a 10% chance that a sample of size 5 would produce a mean of either 7.89 grams or less or 8.11 grams or more. Hence if that happened it would be more likely that the value 8.0 is incorrect (always assuming that the population standard deviation is 0.15 gram).

Because the rejection regions are computed based on areas in tails of distributions, as shown in Figure $8.1.2$, hypothesis tests are classified according to the form of the alternative hypothesis in the following way.

Definitions: Test classifications

- If H_a has the form $\mu \neq \mu_0$ the test is called a **two-tailed test**.
- If H_a has the form $\mu < \mu_0$ the test is called a **left-tailed test**.
- If H_a has the form $\mu > \mu_0$the test is called a **right-tailed test**.

Each of the last two forms is also called a **one-tailed test**.

Two Types of Errors

The format of the testing procedure in general terms is to take a sample and use the information it contains to come to a decision about the two hypotheses. As stated before our decision will always be either

1. reject the null hypothesis H_0 in favor of the alternative H_a presented, or
2. do not reject the null hypothesis H_0 in favor of the alternative H_0 presented.

There are four possible outcomes of hypothesis testing procedure, as shown in the following table:

		True State of Nature	
	H_0 is true	H_0 is false	
Our Decision	Do not reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

As the table shows, there are two ways to be right and two ways to be wrong. Typically to reject H_0 when it is actually true is a more serious error than to fail to reject it when it is false, so the former error is labeled "**Type I**" and the latter error "**Type II**".

Definition: Type I and Type II errors

In a test of hypotheses:

- A *Type I error* is the decision to reject H_0 when it is in fact true.
- A *Type II error* is the decision not to reject H_0 when it is in fact not true.

Unless we perform a census we do not have certain knowledge, so we do not know whether our decision matches the true state of nature or if we have made an error. We reject H_0 if what we observe would be a "rare" event if H_0 were true. But rare events are not impossible: they occur with probability α. Thus when H_0 is true, a rare event will be observed in the proportion α of repeated similar tests, and H_0 will be erroneously rejected in those tests. Thus α is the probability that in following the testing procedure to decide between H_0 and H_a we will make a Type I error.

Definition: level of significance

The number α that is used to determine the rejection region is called the level of significance of the test. It is the probability that the test procedure will result in a *Type I error*.

The probability of making a Type II error is too complicated to discuss in a beginning text, so we will say no more about it than this: for a fixed sample size, choosing $alpha$ smaller in order to reduce the chance of making a Type I error has the effect of increasing the chance of making a **Type II error**. The only way to simultaneously reduce the chances of making either kind of error is to increase the sample size.

Standardizing the Test Statistic

Hypotheses testing will be considered in a number of contexts, and great unification as well as simplification results when the relevant sample statistic is *standardized* by subtracting its mean from it and then dividing by its standard deviation. The resulting statistic is called a *standardized test statistic*. In every situation treated in this and the following two chapters the standardized test statistic will have either the standard normal distribution or Student's t-distribution.

Definition: hypothesis test

A standardized test statistic for a hypothesis test is the statistic that is formed by subtracting from the statistic of interest its mean and dividing by its standard deviation.

For example, reviewing Example 8.1.3, if instead of working with the sample mean $\overline{X}$ we instead work with the test statistic

$$\frac{\overline{X} - 8.0}{0.067}$$

then the distribution involved is standard normal and the critical values are just $\pm z_{0.05}$. The extra work that was done to find that $C = 7.89$ and $C' = 8.11$ is eliminated. In every hypothesis test in this book the standardized test statistic will be governed by either the standard normal distribution or Student's t-distribution. Information about rejection regions is summarized in the following tables:

Table 8.1.1: When the test statistic has the standard normal distribution

Symbol in H_a	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -z_\alpha]$
$>$	Right-tailed test	$[z_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

Table 8.1.2: When the test statistic has Student's t-distribution

Symbol in H_a	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -t_\alpha]$

Symbol in H_a	Terminology	Rejection Region
$>$	Right-tailed test	$[t_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty)$

Every instance of hypothesis testing discussed in this and the following two chapters will have a rejection region like one of the six forms tabulated in the tables above.

No matter what the context a test of hypotheses can always be performed by applying the following systematic procedure, which will be illustrated in the examples in the succeeding sections.

Systematic Hypothesis Testing Procedure: Critical Value Approach

1. Identify the null and alternative hypotheses.
2. Identify the relevant test statistic and its distribution.
3. Compute from the data the value of the test statistic.
4. Construct the rejection region.
5. Compare the value computed in Step 3 to the rejection region constructed in Step 4 and make a decision. Formulate the decision in the context of the problem, if applicable.

The procedure that we have outlined in this section is called the "Critical Value Approach" to hypothesis testing to distinguish it from an alternative but equivalent approach that will be introduced at the end of Section 8.3.

Key Takeaway

- A test of hypotheses is a statistical process for deciding between two competing assertions about a population parameter.
- The testing procedure is formalized in a five-step procedure.

This page titled 8.1: The Elements of Hypothesis Testing is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.2: Large Sample Tests for a Population Mean

Learning Objectives

- To learn how to apply the five-step test procedure for a test of hypotheses concerning a population mean when the sample size is large.
- To learn how to interpret the result of a test of hypotheses in the context of the original narrated situation.

In this section we describe and demonstrate the procedure for conducting a test of hypotheses about the mean of a population in the case that the sample size n is at least 30. The Central Limit Theorem states that $\overline{X}$ is approximately normally distributed, and has mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, where μ and σ are the mean and the standard deviation of the population. This implies that the statistic

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution, which means that probabilities related to it are given in Figure 7.1.5 and the last line in Figure 7.1.6.

If we know σ then the statistic in the display is our test statistic. If, as is typically the case, we do not know σ, then we replace it by the sample standard deviation s. Since the sample is large the resulting test statistic still has a distribution that is approximately standard normal.

Standardized Test Statistics for Large Sample Hypothesis Tests Concerning a Single Population Mean

- If σ is known: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
- If σ is unknown: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

The test statistic has the standard normal distribution.

The distribution of the standardized test statistic and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.2.1.

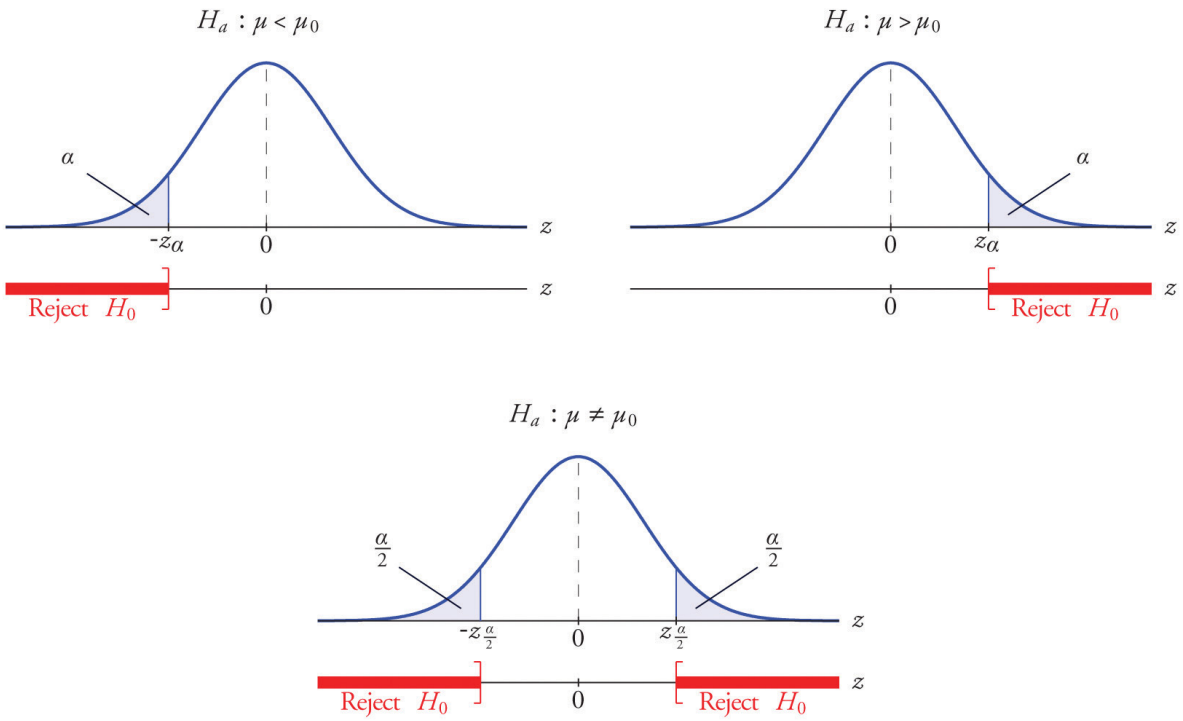


Figure 8.2.1: Distribution of the Standardized Test Statistic and the Rejection Region

Example 8.2.1

It is hoped that a newly developed pain reliever will more quickly produce perceptible reduction in pain to patients after minor surgeries than a standard pain reliever. The standard pain reliever is known to bring relief in an average of 3.5 minutes with standard deviation 2.1 minutes. To test whether the new pain reliever works more quickly than the standard one, 50 patients with minor surgeries were given the new pain reliever and their times to relief were recorded. The experiment yielded sample mean $\bar{x} = 3.1$ minutes and sample standard deviation $s = 1.5$ minutes. Is there sufficient evidence in the sample to indicate, at the 5% level of significance, that the newly developed pain reliever does deliver perceptible relief more quickly?

Solution

We perform the test of hypotheses using the five-step procedure given at the end of Section 8.1.

- **Step 1**. The natural assumption is that the new drug is no better than the old one, but must be proved to be better. Thus if μ denotes the average time until all patients who are given the new drug experience pain relief, the hypothesis test is

$$H_0 : \mu = 3.5$$
$$\text{vs}$$
$$H_a : \mu < 3.5 \ @ \ \alpha = 0.05$$

- **Step 2**. The sample is large, but the population standard deviation is unknown (the 2.1 minutes pertains to the old drug, not the new one). Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  and has the standard normal distribution.
- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.1 - 3.5}{1.5/\sqrt{50}} = -1.886$$

- **Step 4**. Since the symbol in H_a is "$<$" this is a left-tailed test, so there is a single critical value, $-z_\alpha = -z_{0.005}$, which from the last line in Figure 7.1.6 we read off as -1.645. The rejection region is $(-\infty, -1.645]$.
- **Step 5**. As shown in Figure $8.2.2$ the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the average time until patients experience perceptible relief from pain using the new pain reliever is smaller than the average time for the standard pain reliever.

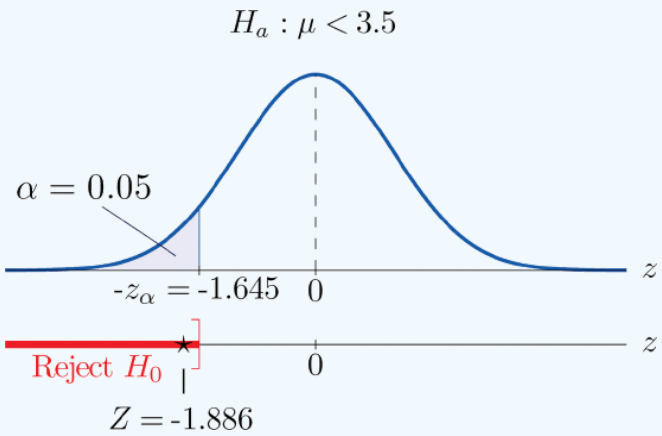


Figure $8.2.2$: Rejection Region and Test Statistic for "Example $8.2.1$"

Example 8.2.2

A cosmetics company fills its best-selling 8 ounce jars of facial cream by an automatic dispensing machine. The machine is set to dispense a mean of 8.1 ounces per jar. Uncontrollable factors in the process can shift the mean away from 8.1 and cause either underfill or overfill, both of which are undesirable. In such a case the dispensing machine is stopped and recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value 0.22 ounce. A quality control engineer routinely selects 30 jars from the assembly line to check the amounts filled. On one occasion, the sample mean is $\bar{x} = 8.2$ ounces and the sample standard deviation is $s = 0.25$ ounce. Determine if there is sufficient evidence in the sample to indicate, at the 1% level of significance, that the machine should be recalibrated.

Solution

- **Step 1**. The natural assumption is that the machine is working properly. Thus if μ denotes the mean amount of facial cream being dispensed, the hypothesis test is

$$H_0 : \mu = 8.1$$
$$\text{vs}$$
$$H_a : \mu \neq 8.1 \text{ @ } \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is known. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  and has the standard normal distribution.
- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{8.2 - 8.1}{0.22/\sqrt{30}} = 2.490$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm z_{\alpha/2} = \pm z_{0.005}$, which from the last line in Figure 7.1.6 "Critical Values of " we read off as ± 2.576. The rejection region is $(-\infty, -2.576] \cup [2.576, \infty)$.
- **Step 5**. As shown in Figure $8.2.3$ the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the average amount of product dispensed is different from 8.1 ounce. We conclude that the machine does not need to be recalibrated.

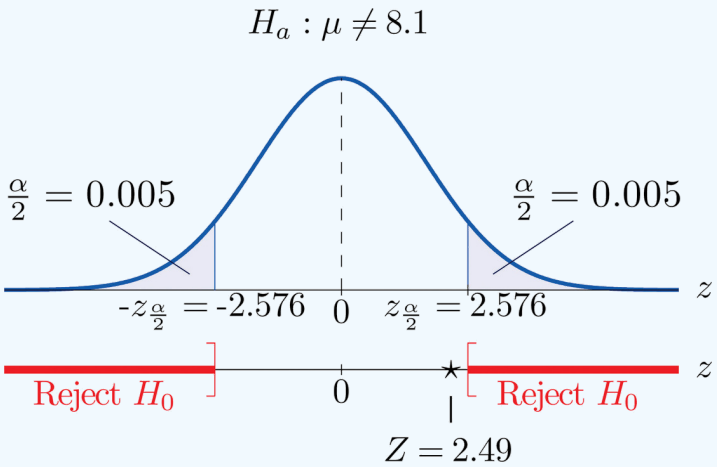


Figure $8.2.3$: Rejection Region and Test Statistic for "Example $8.2.2$"

Key Takeaway

- There are two formulas for the test statistic in testing hypotheses about a population mean with large samples. Both test statistics follow the standard normal distribution.
- The population standard deviation is used if it is known, otherwise the sample standard deviation is used.
- The same five-step procedure is used with either test statistic.

This page titled 8.2: Large Sample Tests for a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.3: The Observed Significance of a Test

Learning Objectives

- To learn what the observed significance of a test is.
- To learn how to compute the observed significance of a test.
- To learn how to apply the p-value approach to hypothesis testing.

The Observed Significance

The conceptual basis of our testing procedure is that we reject H_0 only if the data that we obtained would constitute a rare event if H_0 were actually true. The level of significance α specifies what is meant by "rare." The observed significance of the test is a measure of how rare the value of the test statistic that we have just observed would be if the null hypothesis were true. That is, the observed significance of the test just performed is the probability that, if the test were repeated with a new sample, the result of the new test would be at least as contrary to H_0 and in support of H_a as what was observed in the original test.

Definition: observed significance

The *observed significance* or p-value of a specific test of hypotheses is the probability, on the supposition that H_0 is true, of obtaining a result at least as contrary to H_0 and in favor of H_a as the result actually observed in the sample data.

Think back to "Example 8.2.1", Section 8.2 concerning the effectiveness of a new pain reliever. This was a left-tailed test in which the value of the test statistic was -1.886. To be as contrary to H_0 and in support of H_a as the result $Z = -1.886$ actually observed means to obtain a value of the test statistic in the interval $(-\infty, -1.886]$. Rounding -1.886 to -1.89, we can read directly from Figure 7.1.5 that $P(Z \leq -1.89) = 0.0294$. Thus the p-value or observed significance of the test in "Example 8.2.1", Section 8.2 is 0.0294 or about 3%. Under repeated sampling from this population, if H_0 were true then only about 3% of all samples of size 50 would give a result as contrary to H_0 and in favor of H_a as the sample we observed. Note that the probability 0.0294 is the area of the left tail cut off by the test statistic in this left-tailed test.

Analogous reasoning applies to a right-tailed or a two-tailed test, except that in the case of a two-tailed test being as far from 0 as the observed value of the test statistic but on the opposite side of 0 is just as contrary to H_0 as being the same distance away and on the same side of 0, hence the corresponding tail area is doubled.

Computational Definition of the Observed Significance of a Test of Hypotheses

The **observed significance** of a test of hypotheses is the area of the tail of the distribution cut off by the test statistic (times two in the case of a two-tailed test).

Example 8.3.1

Compute the observed significance of the test performed in "Example 8.2.2", Section 8.2.

Solution

The value of the test statistic was $z = 2.490$, which by Figure 7.1.5 cuts off a tail of area 0.0064, as shown in Figure $8.3.1$. Since the test was two-tailed, the observed significance is $2 \times 0.0064 = 0.0128$.

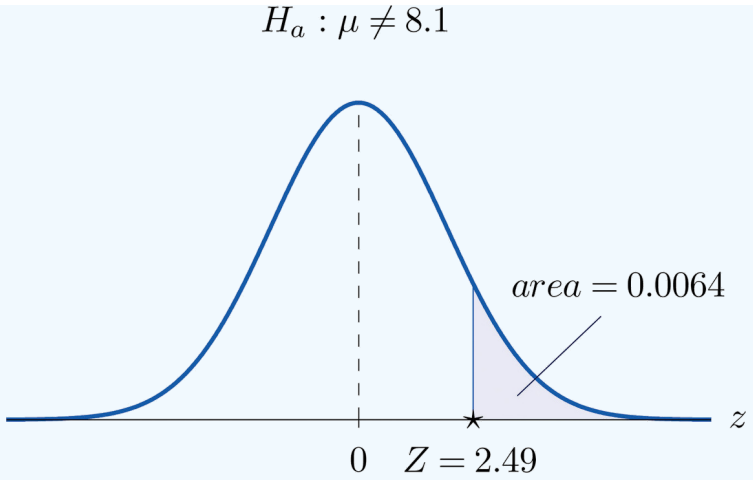


Figure 8.3.1: Area of the Tail for Example 8.3.1.

The p-value Approach to Hypothesis Testing

In "Example 8.2.1", Section 8.2 the test was performed at the 5% level of significance: the definition of “rare” event was probability $\alpha = 0.05$ or less. We saw above that the observed significance of the test was $p = 0.0294$ or about 3%. Since $p = 0.0294 < 0.05 = \alpha$ (or 3% is less than 5%), the decision turned out to be to reject: what was observed was sufficiently unlikely to qualify as an event so rare as to be regarded as (practically) incompatible with H_0.

In "Example 8.2.2", Section 8.2 the test was performed at the 1% level of significance: the definition of “rare” event was probability $\alpha = 0.01$ or less. The observed significance of the test was computed in "Example 8.3.1" as $p = 0.0128$ or about 1.3%. Since $p = 0.0128 > 0.01 = \alpha$ (or 1.3% is greater than 1%), the decision turned out to be not to reject. The event observed was unlikely, but not sufficiently unlikely to lead to rejection of the null hypothesis.

The reasoning just presented is the basis for a slightly different but equivalent formulation of the hypothesis testing process. The first three steps are the same as before, but instead of using α to compute critical values and construct a rejection region, one computes the p-value p of the test and compares it to α, rejecting H_0 if $p \leq \alpha$ and not rejecting if $p > \alpha$.

Systematic Hypothesis Testing Procedure: *p*-Value Approach

1. Identify the null and alternative hypotheses.
2. Identify the relevant test statistic and its distribution.
3. Compute from the data the value of the test statistic.
4. Compute the p-value of the test.
5. Compare the value computed in Step 4 to significance level α and make a decision: reject H_0 if $p \leq \alpha$ and do not reject H_0 if $p > \alpha$. Formulate the decision in the context of the problem, if applicable.

Example 8.3.2

The total score in a professional basketball game is the sum of the scores of the two teams. An expert commentator claims that the average total score for NBA games is 202.5. A fan suspects that this is an overstatement and that the actual average is less than 202.5. He selects a random sample of 85 games and obtains a mean total score of 199.2 with standard deviation 19.63. Determine, at the 5% level of significance, whether there is sufficient evidence in the sample to reject the expert commentator’s claim.

Solution

- **Step 1**. Let μ be the true average total game score of all NBA games. The relevant test is

$$H_0 : \mu = 202.5$$
vs
$$H_a : \mu < 202.5 \text{ @ } \alpha = 0.05$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$z = \frac{\bar{x} - \mu_0}{}$$

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{199.2 - 202.5}{19.63/\sqrt{85}} = -1.55$$

- **Step 4**. The area of the left tail cut off by $z = -1.55$ is, by Figure 7.1.5, 0.0606, as illustrated in Figure $8.3.2$. Since the test is left-tailed, the p-value is just this number, $p = 0.0606$.
- **Step 5**. Since $p = 0.0606 > 0.05 = \alpha$, the decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 5% level of significance, to conclude that the average total score of NBA games is less than 202.5.

$$H_a : \mu < 202.5$$

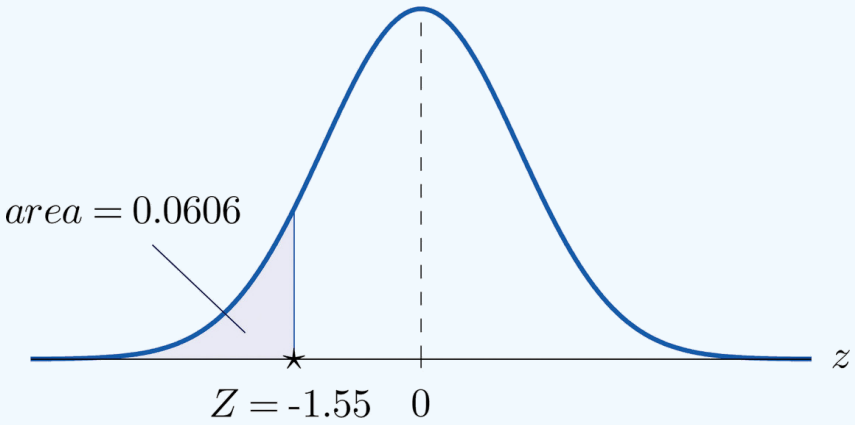


Figure $8.3.2$: Test Statistic for "Example $8.3.2$"

Example $8.3.3$

Mr. Prospero has been teaching Algebra II from a particular textbook at Remote Isle High School for many years. Over the years students in his Algebra II classes have consistently scored an average of 67 on the end of course exam (EOC). This year Mr. Prospero used a new textbook in the hope that the average score on the EOC test would be higher. The average EOC test score of the 64 students who took Algebra II from Mr. Prospero this year had mean 69.4 and sample standard deviation 6.1. Determine whether these data provide sufficient evidence, at the 1% level of significance, to conclude that the average EOC test score is higher with the new textbook.

Solution

- **Step 1**. Let μ be the true average score on the EOC exam of all Mr. Prospero's students who take the Algebra II course with the new textbook. The natural statement that would be assumed true unless there were strong evidence to the contrary is that the new book is about the same as the old one. The alternative, which it takes evidence to establish, is that the new book is better, which corresponds to a higher value of μ. Thus the relevant test is

$$H_0 : \mu = 67$$
$$\text{vs}$$
$$H_a : \mu > 67 \text{ @ } \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{x - \mu_0}{s/\sqrt{n}} = \frac{69.4 - 67}{6.1/\sqrt{64}} = 3.15$$

- **Step 4**. The area of the right tail cut off by $z = 3.15$ is, by Figure 7.1.5, $1 - 0.9992 = 0.0008$, as shown in Figure 8.3.3. Since the test is right-tailed, the p-value is just this number, $p = 0.0008$.
- **Step 5**. Since $p = 0.0008 < 0.01 = \alpha$, the decision is to reject H_0. In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 1% level of significance, to conclude that the average EOC exam score of students taking the Algebra II course from Mr. Prospero using the new book is higher than the average score of those taking the course from him but using the old book.

$$H_a : \mu > 67$$

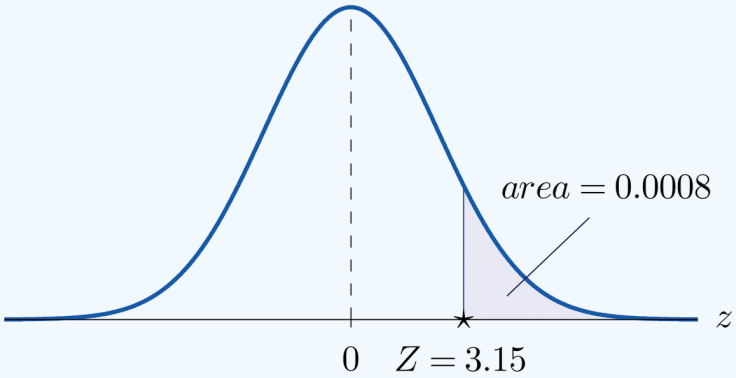


Figure 8.3.3: Test Statistic for "Example 8.3.3"

Example 8.3.4

For the surface water in a particular lake, local environmental scientists would like to maintain an average pH level at 7.4. Water samples are routinely collected to monitor the average pH level. If there is evidence of a shift in pH value, in either direction, then remedial action will be taken. On a particular day 30 water samples are taken and yield average pH reading of 7.7 with sample standard deviation 0.5. Determine, at the 1% level of significance, whether there is sufficient evidence in the sample to indicate that remedial action should be taken.

Solution

- **Step 1**. Let μ be the true average pH level at the time the samples were taken. The relevant test is

$$H_0 : \mu = 7.4$$
$$\text{vs}$$
$$H_a : \mu \neq 7.4 \ @ \ \alpha = 0.01$$

- **Step 2**. The sample is large and the population standard deviation is unknown. Thus the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the standard normal distribution.

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.7 - 7.4}{0.5/\sqrt{30}} = 3.29$$

- **Step 4**. The area of the right tail cut off by $z = 3.29$ is, by Figure 7.1.5, $1 - 0.9995 = 0.0005$, as illustrated in Figure 8.3.4. Since the test is two-tailed, the p-value is the double of this number, p=2×0.0005=0.0010.
- **Step 5**. Since $p = 0.0010 < 0.01 = \alpha$, the decision is to reject H_0. In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 1% level of significance, to conclude that the average pH of surface water in the lake is different from 7.4. That is, remedial action is indicated.

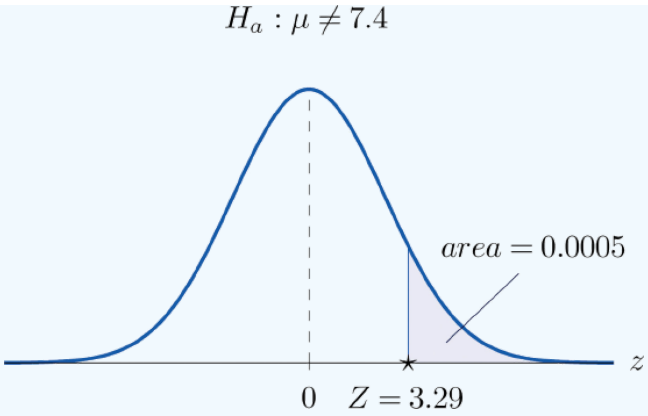


Figure 8.3.4:*Test Statistic for "Example* 8.3.4"

Key Takeaway

- The observed significance or p-value of a test is a measure of how inconsistent the sample result is with H_0 and in favor of H_a.
- The p-value approach to hypothesis testing means that one merely compares the p-value to α instead of constructing a rejection region.
- There is a systematic five-step procedure for the p-value approach to hypothesis testing.

This page titled 8.3: The Observed Significance of a Test is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.4: Small Sample Tests for a Population Mean

Learning Objectives

- To learn how to apply the five-step test procedure for test of hypotheses concerning a population mean when the sample size is small.

In the previous section hypotheses testing for population means was described in the case of large samples. The statistical validity of the tests was insured by the Central Limit Theorem, with essentially no assumptions on the distribution of the population. When sample sizes are small, as is often the case in practice, the Central Limit Theorem does not apply. One must then impose stricter assumptions on the population to give statistical validity to the test procedure. One common assumption is that the population from which the sample is taken has a normal probability distribution to begin with. Under such circumstances, if the population standard deviation is known, then the test statistic

$$\frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}}$$

still has the standard normal distribution, as in the previous two sections. If σ is unknown and is approximated by the sample standard deviation s, then the resulting test statistic

$$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$$

follows Student's t-distribution with $n-1$ degrees of freedom.

Standardized Test Statistics for Small Sample Hypothesis Tests Concerning a Single Population Mean

If σ is known:

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

If σ is unknown:

$$T=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

- The first test statistic (σ known) has the standard normal distribution.
- The second test statistic (σ unknown) has Student's t-distribution with $n-1$ degrees of freedom.
- The population must be normally distributed.

The distribution of the second standardized test statistic (the one containing s) and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure 8.4.1. This is just like Figure 8.2.1 except that now the critical values are from the t-distribution. Figure 8.2.1 still applies to the first standardized test statistic (the one containing (σ) since it follows the standard normal distribution.

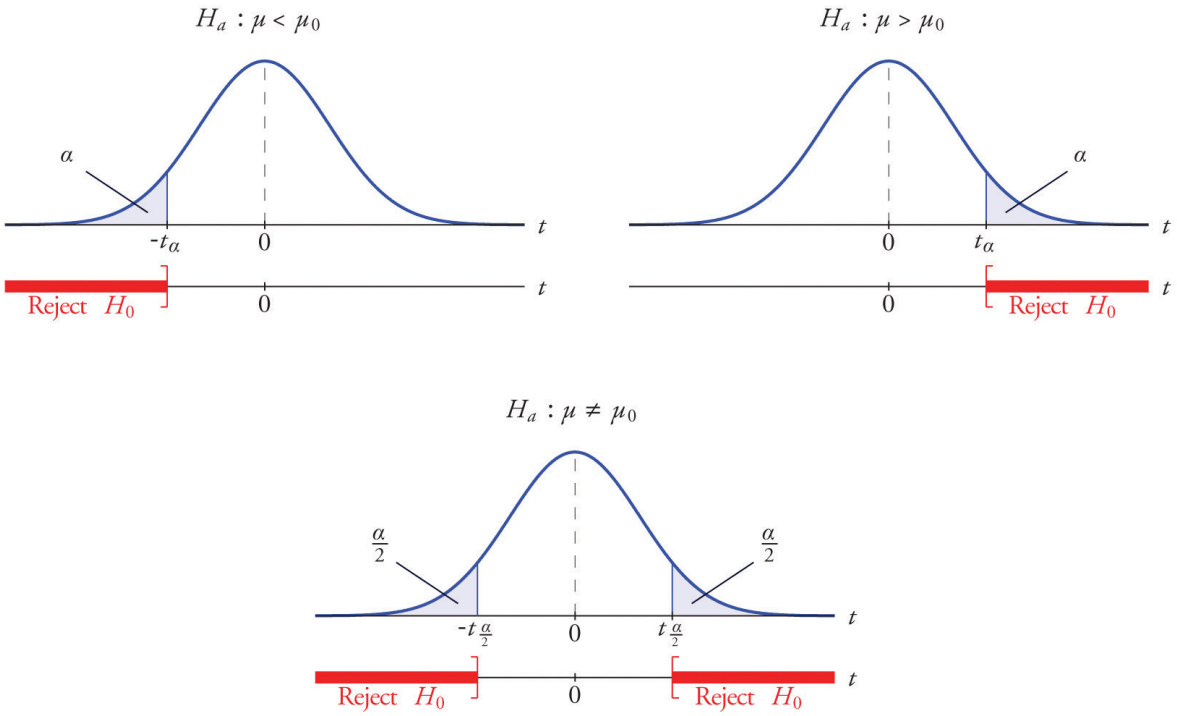


Figure 8.4.1: Distribution of the Standardized Test Statistic and the Rejection Region

The p-value of a test of hypotheses for which the test statistic has Student's t-distribution can be computed using statistical software, but it is impractical to do so using tables, since that would require 30 tables analogous to Figure 7.1.5, one for each degree of freedom from 1 to 30. Figure 7.1.6 can be used to approximate the p-value of such a test, and this is typically adequate for making a decision using the p-value approach to hypothesis testing, although not always. For this reason the tests in the two examples in this section will be made following the critical value approach to hypothesis testing summarized at the end of Section 8.1, but after each one we will show how the p-value approach could have been used.

Example 8.4.1

The price of a popular tennis racket at a national chain store is \$179. Portia bought five of the same racket at an online auction site for the following prices:

155 179 175 175 161

Assuming that the auction prices of rackets are normally distributed, determine whether there is sufficient evidence in the sample, at the 5% level of significance, to conclude that the average price of the racket is less than \$179 if purchased at an online auction.

Solution

- **Step 1**. The assertion for which evidence must be provided is that the average online price μ is less than the average price in retail stores, so the hypothesis test is

$$H_0 : \mu = 179$$
$$\text{vs}$$
$$H_a : \mu < 179 \ @ \ \alpha = 0.05$$

- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  and has the Student t-distribution with $n - 1 = 5 - 1 = 4$ degrees of freedom.
- **Step 3**. From the data we compute $\bar{x} = 169$ and $s = 10.39$. Inserting these values into the formula for the test statistic gives

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{169 - 179}{10.39/\sqrt{5}} = -2.152$$

$s/\sqrt{n}$ $10.39/\sqrt{5}$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $-t_\alpha = -t_{0.05}[df = 4]$. Reading from the row labeled $df = 4$ in Figure 7.1.6 its value is -2.132. The rejection region is $(-\infty, -2.132]$.
- **Step 5**. As shown in Figure 8.4.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the average price of such rackets purchased at online auctions is less than $179.

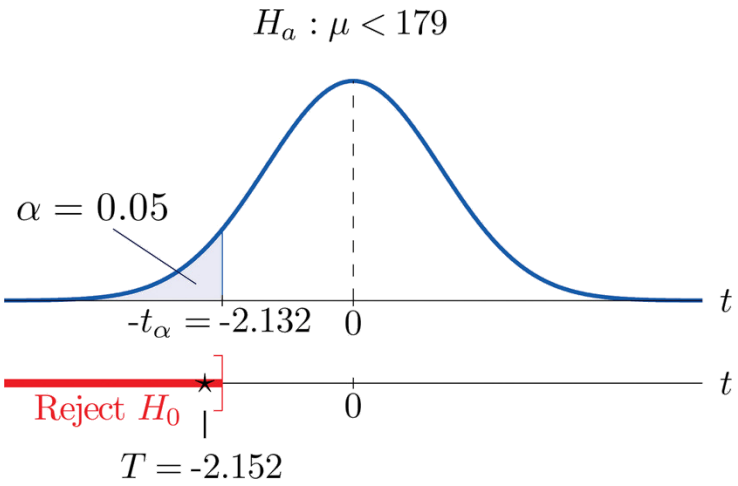


Figure 8.4.2: *Rejection Region and Test Statistic for "Example 8.4.1"*

To perform the test in Example 8.4.1 using the p-value approach, look in the row in Figure 7.1.6 with the heading $df = 4$ and search for the two t-values that bracket the unsigned value 2.152 of the test statistic. They are 2.132 and 2.776, in the columns with headings $t_{0.050}$ and $t_{0.025}$. They cut off right tails of area 0.050 and 0.025, so because 2.152 is between them it must cut off a tail of area between 0.050 and 0.025. By symmetry -2.152 cuts off a left tail of area between 0.050 and 0.025, hence the p-value corresponding to $t = -2.152$ is between 0.025 and 0.05. Although its precise value is unknown, it must be less than $\alpha = 0.05$, so the decision is to reject H_0.

Example 8.4.2

A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process the average distance between the two holes must be tightly controlled at 0.02 mm, else many units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed. Suppose at one time four units are taken and the distances are measured as

0.021 0.019 0.023 0.020

Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.

Solution

- **Step 1**. The assumption is that the process is under control unless there is strong evidence to the contrary. Since a deviation of the average distance to either side is undesirable, the relevant test is

$$H_0 : \mu = 0.02$$
vs
$$H_a : \mu \neq 0.02 \text{ @ } \alpha = 0.01$$

  where μ denotes the mean distance between the holes.
- **Step 2**. The sample is small and the population standard deviation is unknown. Thus the test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and has the Student t-distribution with $n-1=4-1=3$ degrees of freedom.

- **Step 3**. From the data we compute $\bar{x}=0.02075$ and $s=0.00171$. Inserting these values into the formula for the test statistic gives

$$T=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{0.02075-0.02}{0.00171\sqrt{4}}=0.877$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2}=-t_{0.005}[df=3]$. Reading from the row in Figure 7.1.6 labeled $df=3$ their values are ± 5.841. The rejection region is $(-\infty,-5.841]\cup[5.841,\infty)$.
- **Step 5**. As shown in Figure 8.4.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean distance between the holes in the component differs from 0.02 mm.

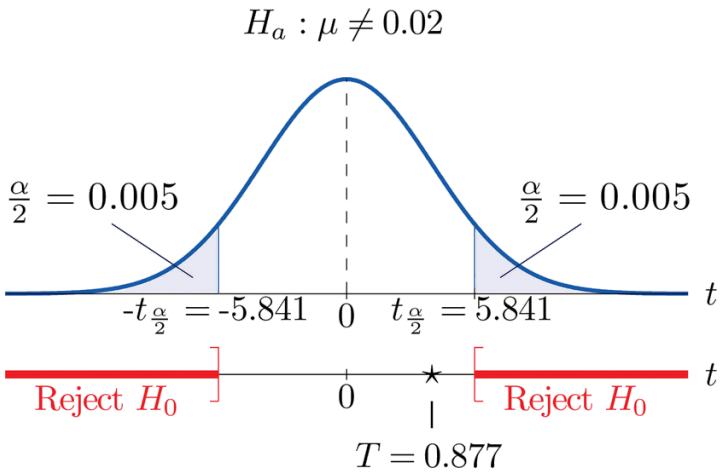


Figure 8.4.3: Rejection Region and Test Statistic for "Example 8.4.2"

To perform the test in "Example 8.4.2" using the p-value approach, look in the row in Figure 7.1.6 with the heading $df=3$ and search for the two t-values that bracket the value 0.877 of the test statistic. Actually 0.877 is smaller than the smallest number in the row, which is 0.978, in the column with heading $t_{0.200}$. The value 0.978 cuts off a right tail of area 0.200, so because 0.877 is to its left it must cut off a tail of area greater than 0.200. Thus the p-value, which is the double of the area cut off (since the test is two-tailed), is greater than 0.400. Although its precise value is unknown, it must be greater than $\alpha=0.01$, so the decision is not to reject H_0.

Key Takeaway

- There are two formulas for the test statistic in testing hypotheses about a population mean with small samples. One test statistic follows the standard normal distribution, the other Student's t-distribution.
- The population standard deviation is used if it is known, otherwise the sample standard deviation is used.
- Either five-step procedure, critical value or p-value approach, is used with either test statistic.

This page titled 8.4: Small Sample Tests for a Population Mean is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.5: Large Sample Tests for a Population Proportion

Learning Objectives

- To learn how to apply the five-step critical value test procedure for test of hypotheses concerning a population proportion.
- To learn how to apply the five-step p-value test procedure for test of hypotheses concerning a population proportion.

Both the critical value approach and the p-value approach can be applied to test hypotheses about a population proportion p. The null hypothesis will have the form $H_0 : p = p_0$ for some specific number p_0 between 0 and 1. The alternative hypothesis will be one of the three inequalities

- $p < p_0$,
- $p > p_0$, or
- $p \neq p_0$

for the same number p_0 that appears in the null hypothesis.

The information in Section 6.3 gives the following formula for the test statistic and its distribution. In the formula p_0 is the numerical value of p that appears in the two hypotheses, $q_0 = 1 - p_0, \hat{p}$ is the sample proportion, and n is the sample size. Remember that the condition that the sample be large is not that n be at least 30 but that the interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

lie wholly within the interval $[0, 1]$.

Standardized Test Statistic for Large Sample Hypothesis Tests Concerning a Single Population Proportion

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_o}{n}}} \tag{8.5.1}$$

The test statistic has the standard normal distribution.

The distribution of the standardized test statistic and the corresponding rejection region for each form of the alternative hypothesis (left-tailed, right-tailed, or two-tailed), is shown in Figure $8.5.1$.

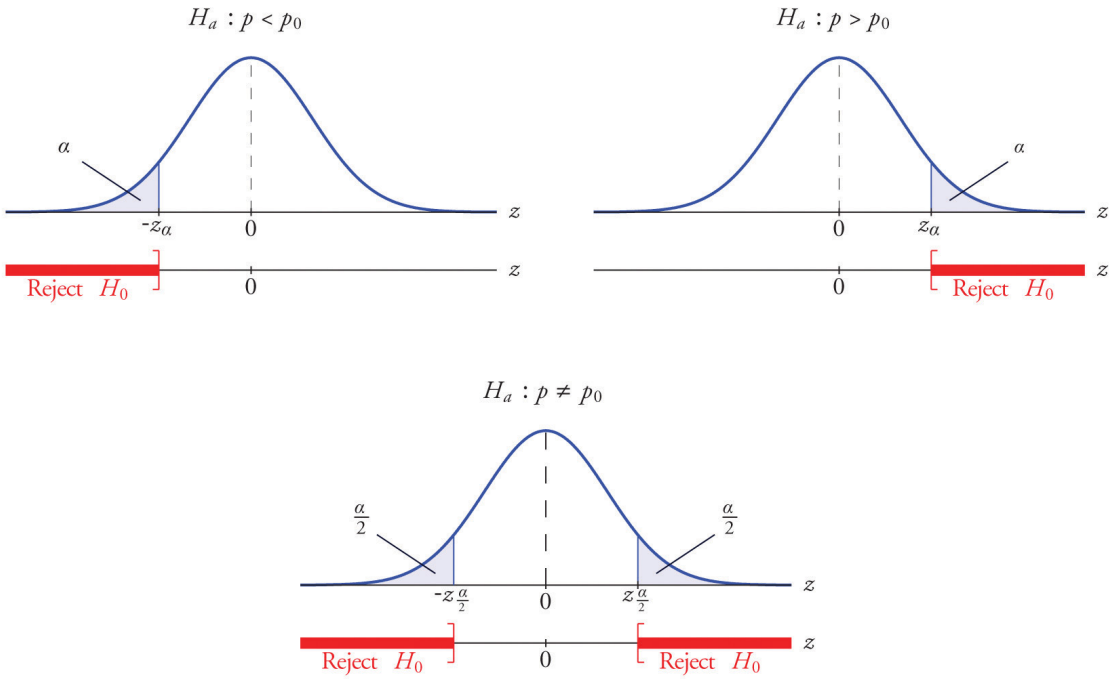


Figure 8.5.1: Distribution of the Standardized Test Statistic and the Rejection Region

Example 8.5.1

A soft drink maker claims that a majority of adults prefer its leading beverage over that of its main competitor's. To test this claim 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink maker's brand, 211 preferred the competitor's brand, and 19 could not make up their minds. Determine whether there is sufficient evidence, at the 5% level of significance, to support the soft drink maker's claim against the default that the population is evenly split in its preference.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.5.3.

We must check that the sample is sufficiently large to validly perform the test. Since $\hat{p} = 270/500 = 0.54$,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{500}} \approx 0.02$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.5.2}$$

$$= [0.54 - (3)(0.02), 0.54 + (3)(0.02)] \tag{8.5.3}$$

$$= [0.48, 0.60] \subset [0, 1] \tag{8.5.4}$$

so the sample is sufficiently large.

- **Step 1.** The relevant test is

$$H_0 : p = 0.50$$

$$vs.$$

$$H_a : p > 0.50 \,@\, \alpha = 0.05$$

where p denotes the proportion of all adults who prefer the company's beverage over that of its competitor's beverage.

- **Step 2.** The test statistic (Equation 8.5.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.5.5}$$

$$= \frac{0.54 - 0.50}{\sqrt{\frac{(0.50)(0.50)}{500}}} \tag{8.5.6}$$

$$= 1.789 \tag{8.5.7}$$

- **Step 4.** Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.05}$. Reading from the last line in Figure 7.1.6 its value is 1.645. The rejection region is $[1.645, \infty)$.
- **Step 5.** As shown in Figure 8.5.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that a majority of adults prefer the company's beverage to that of their competitor's.

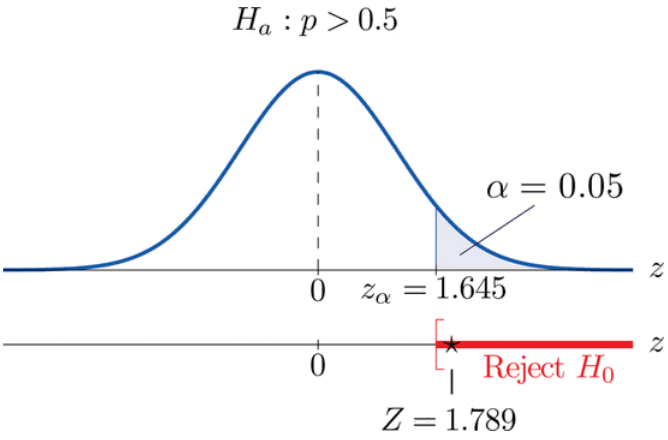


Figure 8.5.2: Rejection Region and Test Statistic for Example 8.5.1

Example 8.5.2

Globally the long-term proportion of newborns who are male is 51.46%. A researcher believes that the proportion of boys at birth changes under severe economic conditions. To test this belief randomly selected birth records of $5,000$ babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the newborns were boys. Determine whether there is sufficient evidence, at the 10% level of significance, to support the researcher's belief.

Solution

We will use the critical value approach to perform the test. The same test will be performed using the p-value approach in Example 8.5.1.

The sample is sufficiently large to validly perform the test since

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.5255)(0.4745)}{5000}} \approx 0.01$$

hence

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{8.5.8}$$

$$= [0.5255 - 0.03, 0.5255 + 0.03] \tag{8.5.9}$$

$$= [0.4955, 0.5555] \subset [0, 1] \tag{8.5.10}$$

- **Step 1**. Let p be the true proportion of boys among all newborns during the recession period. The burden of proof is to show that severe economic conditions change it from the historic long-term value of 0.5146 rather than to show that it stays the same, so the hypothesis test is

$$H_0 : p = 0.5146$$

$$vs.$$

$$H_a : p \neq 0.5146 \,@\, \alpha = 0.10$$

- **Step 2.** The test statistic (Equation 8.5.1) is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

and has the standard normal distribution.

- **Step 3.** The value of the test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{8.5.11}$$

$$= \frac{0.5255 - 0.5146}{\sqrt{\frac{(0.5146)(0.4854)}{5000}}} \tag{8.5.12}$$

$$= 1.542 \tag{8.5.13}$$

- **Step 4.** Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are a pair of critical values, $\pm z_{\alpha/2} = \pm z_{0.05} = \pm 1.645$. The rejection region is $(-\infty, -1.645] \cup [1.645, \infty)$.
- **Step 5.** As shown in Figure 8.5.3 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the proportion of newborns who are male differs from the historic proportion in times of economic recession.

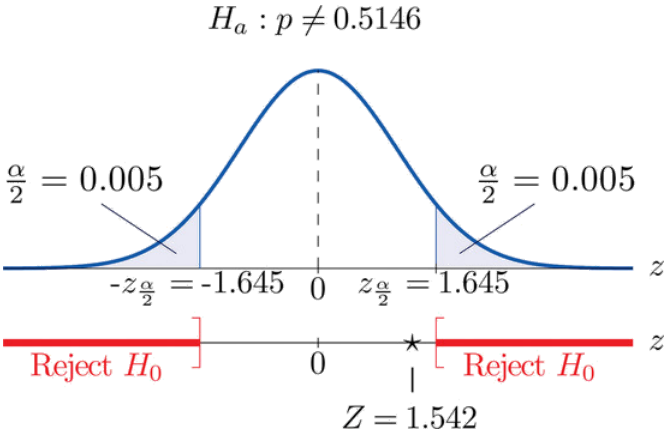


Figure 8.5.3: Rejection Region and Test Statistic for Example 8.5.2.

Example 8.5.3

Perform the test of Example 8.5.1 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.5.1 so we will not repeat them here, but only say that we know that the test is right-tailed and that value of the test statistic is $Z = 1.789$.
- **Step 4.** Since the test is right-tailed the p-value is the area under the standard normal curve cut off by the observed test statistic, $Z = 1.789$, as illustrated in Figure 8.5.4. By Figure 7.1.5 that area and therefore the p-value is $1 - 0.9633 = 0.0367$.
- **Step 5.** Since the p-value is less than $\alpha = 0.05$ the decision is to reject H_0.

$$H_a : p > 0.5$$

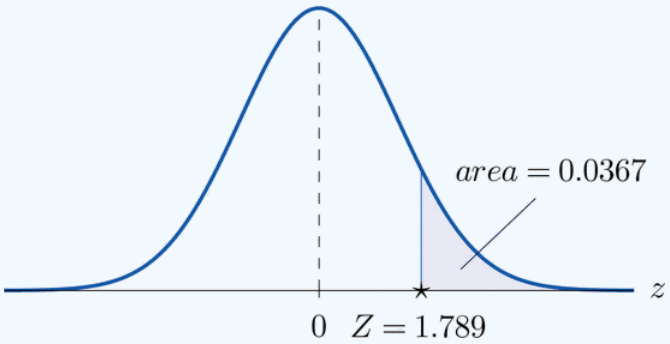


Figure 8.5.4: P-Value for Example 8.5.3

Example 8.5.4

Perform the test of Example 8.5.2 using the p-value approach.

Solution

We already know that the sample size is sufficiently large to validly perform the test.

- **Steps 1–3** of the five-step procedure described in Section 8.3 have already been done in Example 8.5.2. They tell us that the test is two-tailed and that value of the test statistic is $Z = 1.542$.
- **Step 4.** Since the test is two-tailed the p-value is the double of the area under the standard normal curve cut off by the observed test statistic, $Z = 1.542$. By Figure 7.1.5 that area is $1 - 0.9382 = 0.0618$, as illustrated in Figure 8.5.5, hence the p-value is $2 \times 0.0618 = 0.1236$.
- **Step 5.** Since the p-value is greater than $\alpha = 0.10$ the decision is not to reject H_0.

$$H_a : p \neq 0.5146$$

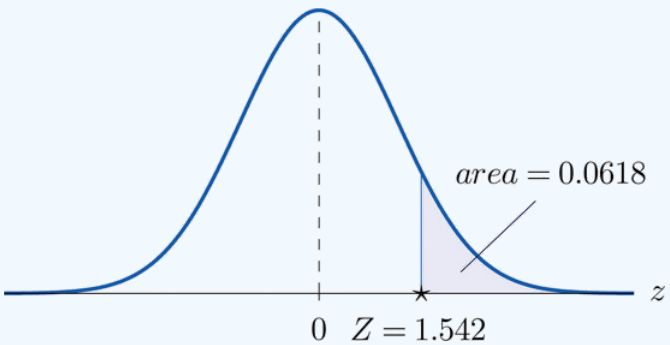


Figure 8.5.5: P-Value for Example 8.5.4

Key Takeaway

- There is one formula for the test statistic in testing hypotheses about a population proportion. The test statistic follows the standard normal distribution.
- Either five-step procedure, critical value or p-value approach, can be used.

This page titled 8.5: Large Sample Tests for a Population Proportion is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

8.E: Testing Hypotheses (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

8.1: The Elements of Hypothesis Testing

Q8.1.1

State the null and alternative hypotheses for each of the following situations. (That is, identify the correct number μ_0 and write $H_0 : \mu = \mu_0$ and the appropriate analogous expression for H_a.)

a. The average July temperature in a region historically has been $74.5^\circ F$. Perhaps it is higher now.
b. The average weight of a female airline passenger with luggage was 145 pounds ten years ago. The FAA believes it to be higher now.
c. The average stipend for doctoral students in a particular discipline at a state university is $\$14,756$. The department chairman believes that the national average is higher.
d. The average room rate in hotels in a certain region is $\$82.53$. A travel agent believes that the average in a particular resort area is different.
e. The average farm size in a predominately rural state was 69.4 acres. The secretary of agriculture of that state asserts that it is less today.

Q1.1.2

State the null and alternative hypotheses for each of the following situations. (That is, identify the correct number μ_0 and write $H_0 : \mu = \mu_0$ and the appropriate analogous expression for H_a.)

a. The average time workers spent commuting to work in Verona five years ago was 38.2 minutes. The Verona Chamber of Commerce asserts that the average is less now.
b. The mean salary for all men in a certain profession is $\$58,291$. A special interest group thinks that the mean salary for women in the same profession is different.
c. The accepted figure for the caffeine content of an 8-ounce cup of coffee is 133 mg. A dietitian believes that the average for coffee served in a local restaurants is higher.
d. The average yield per acre for all types of corn in a recent year was 161.9 bushels. An economist believes that the average yield per acre is different this year.
e. An industry association asserts that the average age of all self-described fly fishermen is 42.8 years. A sociologist suspects that it is higher.

Q1.1.3

Describe the two types of errors that can be made in a test of hypotheses.

Q1.1.4

Under what circumstance is a test of hypotheses certain to yield a correct decision?

Answers

1. a. $H_0 : \mu = 74.5 \; vs \; H_a : \mu > 74.5$
   b. $H_0 : \mu = 145 \; vs \; H_a : \mu > 145$
   c. $H_0 : \mu = 14756 \; vs \; H_a : \mu > 14756$
   d. $H_0 : \mu = 82.53 \; vs \; H_a : \mu \neq 82.53$
   e. $H_0 : \mu = 69.4 \; vs \; H_a : \mu < 69.4$
2.
3. A Type I error is made when a true H_0 is rejected. A Type II error is made when a false H_0 is not rejected.

8.2: Large Sample Tests for a Population Mean

Basic

1. Find the rejection region (for the standardized test statistic) for each hypothesis test.
   a. $H_0 : \mu = 27 \; vs \; H_a : \mu < 27 \; @ \; \alpha = 0.05$

b. $H_0 : \mu = 52 \; vs \; H_a : \mu \neq 52 \; @ \; \alpha = 0.05$
c. $H_0 : \mu = -105 \; vs \; H_a : \mu > -105 \; @ \; \alpha = 0.10$
d. $H_0 : \mu = 78.8 \; vs \; H_a : \mu \neq 78.8 \; @ \; \alpha = 0.10$

2. Find the rejection region (for the standardized test statistic) for each hypothesis test.
   a. $H_0 : \mu = 17 \; vs \; H_a : \mu < 17 \; @ \; \alpha = 0.01$
   b. $H_0 : \mu = 880 \; vs \; H_a : \mu \neq 880 \; @ \; \alpha = 0.01$
   c. $H_0 : \mu = -12 \; vs \; H_a : \mu > -12 \; @ \; \alpha = 0.05$
   d. $H_0 : \mu = 21.1 \; vs \; H_a : \mu \neq 21.1 \; @ \; \alpha = 0.05$
3. Find the rejection region (for the standardized test statistic) for each hypothesis test. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = 141 \; vs \; H_a : \mu < 141 \; @ \; \alpha = 0.20$
   b. $H_0 : \mu = -54 \; vs \; H_a : \mu < -54 \; @ \; \alpha = 0.05$
   c. $H_0 : \mu = 98.6 \; vs \; H_a : \mu \neq 98.6 \; @ \; \alpha = 0.05$
   d. $H_0 : \mu = 3.8 \; vs \; H_a : \mu > 3.8 \; @ \; \alpha = 0.001$
4. Find the rejection region (for the standardized test statistic) for each hypothesis test. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = -62 \; vs \; H_a : \mu \neq -62 \; @ \; \alpha = 0.005$
   b. $H_0 : \mu = 73 \; vs \; H_a : \mu > 73 \; @ \; \alpha = 0.001$
   c. $H_0 : \mu = 1124 \; vs \; H_a : \mu < 1124 \; @ \; \alpha = 0.001$
   d. $H_0 : \mu = 0.12 \; vs \; H_a : \mu \neq 0.12 \; @ \; \alpha = 0.001$
5. Compute the value of the test statistic for the indicated test, based on the information given.
   a. Testing $H_0 : \mu = 72.2 \; vs \; H_a : \mu > 72.2, \; \sigma$ unknown $n = 55, \; \bar{x} = 75.1, \; s = 9.25$
   b. Testing $H_0 : \mu = 58 \; vs \; H_a : \mu > 58, \; \sigma = 1.22 \; n = 40, \; \bar{x} = 58.5, \; s = 1.29$
   c. Testing $H_0 : \mu = -19.5 \; vs \; H_a : \mu < -19.5, \; \sigma$ unknown $n = 30, \; \bar{x} = -23.2, \; s = 9.55$
   d. Testing $H_0 : \mu = 805 \; vs \; H_a : \mu \neq 805, \; \sigma = 37.5 \; n = 75, \; \bar{x} = 818, \; s = 36.2$
6. Compute the value of the test statistic for the indicated test, based on the information given.
   a. Testing $H_0 : \mu = 342 \; vs \; H_a : \mu < 342, \; \sigma = 11.2 \; n = 40, \; \bar{x} = 339, \; s = 10.3$
   b. Testing $H_0 : \mu = 105 \; vs \; H_a : \mu > 105, \; \sigma = 5.3 \; n = 80, \; \bar{x} = 107, \; s = 5.1$
   c. Testing $H_0 : \mu = -13.5 \; vs \; H_a : \mu \neq -13.5, \; \sigma$ unknown $n = 32, \; \bar{x} = -13.8, \; s = 1.5$
   d. Testing $H_0 : \mu = 28 \; vs \; H_a : \mu \neq 28, \; \sigma$ unknown $n = 68, \; \bar{x} = 27.8, \; s = 1.3$
7. Perform the indicated test of hypotheses, based on the information given.
   a. Test $H_0 : \mu = 212 \; vs \; H_a : \mu < 212 \; @ \; \alpha = 0.10, \; \sigma$ unknown $n = 36, \; \bar{x} = 211.2, \; s = 2.2$
   b. Test $H_0 : \mu = -18 \; vs \; H_a : \mu > -18 \; @ \; \alpha = 0.05, \; \sigma = 3.3 \; n = 44, \; \bar{x} = -17.2, \; s = 3.1$
   c. Test $H_0 : \mu = 24 \; vs \; H_a : \mu \neq 24 \; @ \; \alpha = 0.02, \; \sigma$ unknown $n = 50, \; \bar{x} = 22.8, \; s = 1.9$
8. Perform the indicated test of hypotheses, based on the information given.
   a. Test $H_0 : \mu = 105 \; vs \; H_a : \mu > 105 \; @ \; \alpha = 0.05, \; \sigma$ unknown $n = 30, \; \bar{x} = 108, \; s = 7.2$
   b. Test $H_0 : \mu = 21.6 \; vs \; H_a : \mu < 21.6 \; @ \; \alpha = 0.01, \; \sigma$ unknown $n = 78, \; \bar{x} = 20.5, \; s = 3.9$
   c. Test $H_0 : \mu = -375 \; vs \; H_a : \mu \neq -375 \; @ \; \alpha = 0.01, \; \sigma = 18.5 \; n = 31, \; \bar{x} = -388, \; s = 18.0$

Applications

9. In the past the average length of an outgoing telephone call from a business office has been 143 seconds. A manager wishes to check whether that average has decreased after the introduction of policy changes. A sample of 100 telephone calls produced a mean of 133 seconds, with a standard deviation of 35 seconds. Perform the relevant test at the 1% level of significance.
10. The government of an impoverished country reports the mean age at death among those who have survived to adulthood as 66.2 years. A relief agency examines 30 randomly selected deaths and obtains a mean of 62.3 years with standard deviation 8.1 years. Test whether the agency's data support the alternative hypothesis, at the 1% level of significance, that the population mean is less than 66.2.
11. The average household size in a certain region several years ago was 3.14 persons. A sociologist wishes to test, at the 5% level of significance, whether it is different now. Perform the test using the information collected by the sociologist: in a random sample of 75 households, the average size was 2.98 persons, with sample standard deviation 0.82 person.

12. The recommended daily calorie intake for teenage girls is $2,200$ calories/day. A nutritionist at a state university believes the average daily caloric intake of girls in that state to be lower. Test that hypothesis, at the 5% level of significance, against the null hypothesis that the population average is $2,200$ calories/day using the following sample data: $n = 36,\ \bar{x} = 2,150,\ s = 203$
13. An automobile manufacturer recommends oil change intervals of $3,000$ miles. To compare actual intervals to the recommendation, the company randomly samples records of 50 oil changes at service facilities and obtains sample mean $3,752$ miles with sample standard deviation 638 miles. Determine whether the data provide sufficient evidence, at the 5% level of significance, that the population mean interval between oil changes exceeds $3,000$ miles.
14. A medical laboratory claims that the mean turn-around time for performance of a battery of tests on blood samples is 1.88 business days. The manager of a large medical practice believes that the actual mean is larger. A random sample of 45 blood samples yielded mean 2.09 and sample standard deviation 0.13 day. Perform the relevant test at the 10% level of significance, using these data.
15. A grocery store chain has as one standard of service that the mean time customers wait in line to begin checking out not exceed 2 minutes. To verify the performance of a store the company measures the waiting time in 30 instances, obtaining mean time 2.17 minutes with standard deviation 0.46 minute. Use these data to test the null hypothesis that the mean waiting time is 2 minutes versus the alternative that it exceeds 2 minutes, at the 10% level of significance.
16. A magazine publisher tells potential advertisers that the mean household income of its regular readership is $\$61,500$. An advertising agency wishes to test this claim against the alternative that the mean is smaller. A sample of 40 randomly selected regular readers yields mean income $\$59,800$ with standard deviation $\$5,850$. Perform the relevant test at the 1% level of significance.
17. Authors of a computer algebra system wish to compare the speed of a new computational algorithm to the currently implemented algorithm. They apply the new algorithm to 50 standard problems; it averages 8.16 seconds with standard deviation 0.17 second. The current algorithm averages 8.21 seconds on such problems. Test, at the 1% level of significance, the alternative hypothesis that the new algorithm has a lower average time than the current algorithm.
18. A random sample of the starting salaries of 35 randomly selected graduates with bachelor's degrees last year gave sample mean and standard deviation $\$41,202$ and $\$7,621$, respectively. Test whether the data provide sufficient evidence, at the 5% level of significance, to conclude that the mean starting salary of all graduates last year is less than the mean of all graduates two years before, $\$43,589$.

Additional Exercises

19. The mean household income in a region served by a chain of clothing stores is $\$48,750$. In a sample of 40 customers taken at various stores the mean income of the customers was $\$51,505$ with standard deviation $\$6,852$.
    a. Test at the 10% level of significance the null hypothesis that the mean household income of customers of the chain is $\$48,750$ against that alternative that it is different from $\$48,750$.
    b. The sample mean is greater than $\$48,750$, suggesting that the actual mean of people who patronize this store is greater than $\$48,750$. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
20. The labor charge for repairs at an automobile service center are based on a standard time specified for each type of repair. The time specified for replacement of universal joint in a drive shaft is one hour. The manager reviews a sample of 30 such repairs. The average of the actual repair times is 0.86 hour with standard deviation 0.32 hour.
    a. Test at the 1% level of significance the null hypothesis that the actual mean time for this repair differs from one hour.
    b. The sample mean is less than one hour, suggesting that the mean actual time for this repair is less than one hour. Perform this test, also at the 1% level of significance. (The computation of the test statistic done in part (a) still applies here.)

Large Data Set Exercises

Large Data Set missing from the original

21. Large $\text{Data Set } 1$ records the SAT scores of $1,000$ students. Regarding it as a random sample of all high school students, use it to test the hypothesis that the population mean exceeds $1,510$, at the 1% level of significance. (The null hypothesis is that $\mu = 1510$).
22. Large $\text{Data Set } 1$ records the GPAs of $1,000$ college students. Regarding it as a random sample of all college students, use it to test the hypothesis that the population mean is less than 2.50, at the 10% level of significance. (The null hypothesis is that $\mu = 2.50$).
23. Large $\text{Data Set } 1$ lists the SAT scores of $1,000$ students.

a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
b. Regard the first 50 students in the data set as a random sample drawn from the population of part (a) and use it to test the hypothesis that the population mean exceeds $1,510$, at the 10% level of significance. (The null hypothesis is that $\mu = 1510$).
c. Is your conclusion in part (b) in agreement with the true state of nature (which by part (a) you know), or is your decision in error? If your decision is in error, is it a Type I error or a Type II error?

24. Large $\text{Data Set } 1$ lists the GPAs of $1,000$ students.
a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
b. Regard the first 50 students in the data set as a random sample drawn from the population of part (a) and use it to test the hypothesis that the population mean is less than 2.50, at the 10% level of significance. (The null hypothesis is that $\mu = 2.50$).
c. Is your conclusion in part (b) in agreement with the true state of nature (which by part (a) you know), or is your decision in error? If your decision is in error, is it a Type I error or a Type II error?

Answers

1. a. $Z \le -1.645$
   b. $Z \le -1.645$ *or* $Z \ge 1.96$
   c. $Z \ge 1.28$
   d. $Z \le -1.645$ *or* $Z \ge 1.645$
2.
3. a. $Z \le -0.84$
   b. $Z \le -1.645$
   c. $Z \le -1.96$ *or* $Z \ge 1.96$
   d. $Z \ge 3.1$
4.
5. a. $Z = 2.235$
   b. $Z = 2.592$
   c. $Z = -2.122$
   d. $Z = 3.002$
6.
7. a. $Z = -2.18,\ -z_{0.10} = -1.28,$ reject H_0
   b. $Z = 1.61,\ z_{0.05} = 1.645,$ do not reject H_0
   c. $Z = -4.47,\ -z_{0.01} = -2.33,$ reject H_0
8.
9. $Z = -2.86,\ -z_{0.01} = -2.33,$ reject H_0
10.
11. $Z = -1.69,\ -z_{0.025} = -1.96,$ do not reject H_0
12.
13. $Z = 8.33,\ z_{0.05} = 1.645,$ reject H_0
14.
15. $Z = 2.02,\ z_{0.10} = 1.28,$ reject H_0
16.
17. $Z = -2.08,\ -z_{0.01} = -2.33,$ do not reject H_0
18.
19. a. $Z = 2.54,\ z_{0.05} = 1.645,$ reject H_0
    b. $Z = 2.54,\ z_{0.10} = 1.28,$ reject H_0
20.
21. $H_0 : \mu = 1510$ *vs* $H_a : \mu > 1510$. Test Statistic: $Z = 2.7882$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0.
22.
23. a. $\mu_0 = 1528.74$

b. $H_0 : \mu = 1510$ *vs* $H_a : \mu > 1510$. Test Statistic: $Z = -1.41$. Rejection Region: $[1.28, \infty)$. Decision: Fail to reject H_0.
c. No, it is a Type II error.

8.3: The Observed Significance of a Test

Basic

1. Compute the observed significance of each test.
   a. Testing $H_0 : \mu = 54.7$ *vs* $H_a : \mu < 54.7$, test statistic $z = -1.72$
   b. Testing $H_0 : \mu = 195$ *vs* $H_a : \mu \neq 195$, test statistic $z = -2.07$
   c. Testing $H_0 : \mu = -45$ *vs* $H_a : \mu > -45$, test statistic $z = 2.54$
2. Compute the observed significance of each test.
   a. Testing $H_0 : \mu = 0$ *vs* $H_a : \mu \neq 0$, test statistic $z = 2.82$
   b. Testing $H_0 : \mu = 18.4$ *vs* $H_a : \mu < 18.4$, test statistic $z = -1.74$
   c. Testing $H_0 : \mu = 63.85$ *vs* $H_a : \mu > 63.85$, test statistic $z = 1.93$
3. Compute the observed significance of each test. (Some of the information given might not be needed.)
   a. Testing $H_0 : \mu = 27.5$ *vs* $H_a : \mu > 27.5$, $n = 49$, $\bar{x} = 28.9$, $s = 3.14$, test statistic $z = 3.12$
   b. Testing $H_0 : \mu = 581$ *vs* $H_a : \mu < 581$, $n = 32$, $\bar{x} = 560$, $s = 47.8$, test statistic $z = -2.49$
   c. Testing $H_0 : \mu = 138.5$ *vs* $H_a : \mu \neq 138.5$, $n = 44$, $\bar{x} = 137.6$, $s = 2.45$, test statistic $z = -2.44$
4. Compute the observed significance of each test. (Some of the information given might not be needed.)
   a. Testing $H_0 : \mu = -17.9$ *vs* $H_a : \mu < -17.9$, $n = 34$, $\bar{x} = -18.2$, $s = 0.87$, test statistic $z = -2.01$
   b. Testing $H_0 : \mu = 5.5$ *vs* $H_a : \mu \neq 5.5$, $n = 56$, $\bar{x} = 7.4$, $s = 4.82$, test statistic $z = 2.95$
   c. Testing $H_0 : \mu = 1255$ *vs* $H_a : \mu > 1255$, $n = 152$, $\bar{x} = 1257$, $s = 7.5$, test statistic $z = 3.29$
5. Make the decision in each test, based on the information provided.
   a. Testing $H_0 : \mu = 82.9$ *vs* $H_a : \mu < 82.9$ @ $\alpha = 0.05$, observed significance $p = 0.038$
   b. Testing $H_0 : \mu = 213.5$ *vs* $H_a : \mu \neq 213.5$ @ $\alpha = 0.01$, observed significance $p = 0.038$
6. Make the decision in each test, based on the information provided.
   a. Testing $H_0 : \mu = 31.4$ *vs* $H_a : \mu > 31.4$ @ $\alpha = 0.10$, observed significance $p = 0.062$
   b. Testing $H_0 : \mu = -75.5$ *vs* $H_a : \mu < -75.5$ @ $\alpha = 0.05$, observed significance $p = 0.062$

Applications

7. A lawyer believes that a certain judge imposes prison sentences for property crimes that are longer than the state average 11.7 months. He randomly selects 36 of the judge's sentences and obtains mean 13.8 and standard deviation 3.9 months.
   a. Perform the test at the 1% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 1% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
8. In a recent year the fuel economy of all passenger vehicles was 19.8 mpg. A trade organization sampled 50 passenger vehicles for fuel economy and obtained a sample mean of 20.1 mpg with standard deviation 2.45 mpg. The sample mean 20.1 exceeds 19.8, but perhaps the increase is only a result of sampling error.
   a. Perform the relevant test of hypotheses at the 20% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 20% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
9. The mean score on a 25-point placement exam in mathematics used for the past two years at a large state university is 14.3. The placement coordinator wishes to test whether the mean score on a revised version of the exam differs from 14.3. She gives the revised exam to 30 entering freshmen early in the summer; the mean score is 14.6 with standard deviation 2.4.
   a. Perform the test at the 10% level of significance using the critical value approach.
   b. Compute the observed significance of the test.
   c. Perform the test at the 10% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).

10. The mean increase in word family vocabulary among students in a one-year foreign language course is 576 word families. In order to estimate the effect of a new type of class scheduling, an instructor monitors the progress of 60 students; the sample mean increase in word family vocabulary of these students is 542 word families with sample standard deviation 18 word families.
    a. Test at the 5% level of significance whether the mean increase with the new class scheduling is different from 576 word families, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
11. The mean yield for hard red winter wheat in a certain state is 44.8 bu/acre. In a pilot program a modified growing scheme was introduced on 35 independent plots. The result was a sample mean yield of 45.4 bu/acre with sample standard deviation 1.6 bu/acre, an apparent increase in yield.
    a. Test at the 5% level of significance whether the mean yield under the new scheme is greater than 44.8 bu/acre, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).
12. The average amount of time that visitors spent looking at a retail company's old home page on the world wide web was 23.6 seconds. The company commissions a new home page. On its first day in place the mean time spent at the new page by $7,628$ visitors was 23.5 seconds with standard deviation 5.1 seconds.
    a. Test at the 5% level of significance whether the mean visit time for the new page is less than the former mean of 23.6 seconds, using the critical value approach.
    b. Compute the observed significance of the test.
    c. Perform the test at the 5% level of significance using the p-value approach. You need not repeat the first three steps, already done in part (a).

Answers

1. a. $p\text{-value} = 0.0427$
   b. $p\text{-value} = 0.0384$
   c. $p\text{-value} = 0.0055$
2.
3. a. $p\text{-value} = 0.0009$
   b. $p\text{-value} = 0.0064$
   c. $p\text{-value} = 0.0146$
4.
5. a. reject H_0
   b. do not reject H_0
6.
7. a. $Z = 3.23,\ z_{0.01} = 2.33$, reject H_0
   b. $p\text{-value} = 0.0006$
   c. reject H_0
8.
9. a. $Z = 0.68,\ z_{0.05} = 1.645$, do not reject H_0
   b. $p\text{-value} = 0.4966$
   c. do not reject H_0
10.
11. a. $Z = 2.22,\ z_{0.05} = 1.645$, reject H_0
    b. $p\text{-value} = 0.0132$
    c. reject H_0

8.4: Small Sample Tests for a Population Mean

Basic

1. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed.
   a. $H_0 : \mu = 27 \; vs \; H_a : \mu < 27 \; @ \; \alpha = 0.05, \; n = 12, \; \sigma = 2.2$
   b. $H_0 : \mu = 52 \; vs \; H_a : \mu \neq 52 \; @ \; \alpha = 0.05, \; n = 6, \; \sigma$ unknown
   c. $H_0 : \mu = -105 \; vs \; H_a : \mu > -105 \; @ \; \alpha = 0.10, \; n = 24, \; \sigma$ unknown
   d. $H_0 : \mu = 78.8 \; vs \; H_a : \mu \neq 78.8 \; @ \; \alpha = 0.10, \; n = 8, \; \sigma = 1.7$
2. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed.
   a. $H_0 : \mu = 17 \; vs \; H_a : \mu < 17 \; @ \; \alpha = 0.01, \; n = 26, \; \sigma = 0.94$
   b. $H_0 : \mu = 880 \; vs \; H_a : \mu \neq 880 \; @ \; \alpha = 0.01, \; n = 4, \; \sigma$ unknown
   c. $H_0 : \mu = -12 \; vs \; H_a : \mu > -12 \; @ \; \alpha = 0.05, \; n = 18, \; \sigma = 1.1$
   d. $H_0 : \mu = 21.1 \; vs \; H_a : \mu \neq 21.1 \; @ \; \alpha = 0.05, \; n = 23, \; \sigma$ unknown
3. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = 141 \; vs \; H_a : \mu < 141 \; @ \; \alpha = 0.20, \; n = 29, \; \sigma$ unknown
   b. $H_0 : \mu = -54 \; vs \; H_a : \mu < -54 \; @ \; \alpha = 0.05, \; n = 15, \; \sigma = 1.9$
   c. $H_0 : \mu = 98.6 \; vs \; H_a : \mu \neq 98.6 \; @ \; \alpha = 0.05, \; n = 12, \; \sigma$ unknown
   d. $H_0 : \mu = 3.8 \; vs \; H_a : \mu > 3.8 \; @ \; \alpha = 0.001, \; n = 27, \; \sigma$ unknown
4. Find the rejection region (for the standardized test statistic) for each hypothesis test based on the information given. The population is normally distributed. Identify the test as left-tailed, right-tailed, or two-tailed.
   a. $H_0 : \mu = -62 \; vs \; H_a : \mu \neq -62 \; @ \; \alpha = 0.005, \; n = 8, \; \sigma$ unknown
   b. $H_0 : \mu = 73 \; vs \; H_a : \mu > 73 \; @ \; \alpha = 0.001, \; n = 22, \; \sigma$ unknown
   c. $H_0 : \mu = 1124 \; vs \; H_a : \mu < 1124 \; @ \; \alpha = 0.001, \; n = 21, \; \sigma$ unknown
   d. $H_0 : \mu = 0.12 \; vs \; H_a : \mu \neq 0.12 \; @ \; \alpha = 0.001, \; n = 14, \; \sigma = 0.026$
5. A random sample of size 20 drawn from a normal population yielded the following results: $\bar{x} = 49.2, \; s = 1.33$
   a. Test $H_0 : \mu = 50 \; vs \; H_a : \mu \neq 50 \; @ \; \alpha = 0.01$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.
6. A random sample of size 16 drawn from a normal population yielded the following results: $\bar{x} = -0.96, \; s = 1.07$
   a. Test $H_0 : \mu = 0 \; vs \; H_a : \mu < 0 \; @ \; \alpha = 0.001$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.
7. A random sample of size 8 drawn from a normal population yielded the following results: $\bar{x} = 289, \; s = 46$
   a. Test $H_0 : \mu = 250 \; vs \; H_a : \mu > 250 \; @ \; \alpha = 0.05$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.
8. A random sample of size 12 drawn from a normal population yielded the following results: $\bar{x} = 86.2, \; s = 0.63$
   a. Test $H_0 : \mu = 85.5 \; vs \; H_a : \mu \neq 85.5 \; @ \; \alpha = 0.01$.
   b. Estimate the observed significance of the test in part (a) and state a decision based on the p-value approach to hypothesis testing.

Applications

9. Researchers wish to test the efficacy of a program intended to reduce the length of labor in childbirth. The accepted mean labor time in the birth of a first child is 15.3 hours. The mean length of the labors of 13 first-time mothers in a pilot program was 8.8 hours with standard deviation 3.1 hours. Assuming a normal distribution of times of labor, test at the 10% level of significance test whether the mean labor time for all women following this program is less than 15.3 hours.

10. A dairy farm uses the somatic cell count (SCC) report on the milk it provides to a processor as one way to monitor the health of its herd. The mean SCC from five samples of raw milk was $250,000$ cells per milliliter with standard deviation $37,500$ cell/ml. Test whether these data provide sufficient evidence, at the 10% level of significance, to conclude that the mean SCC of all milk produced at the dairy exceeds that in the previous report, $210,250$ cell/ml. Assume a normal distribution of SCC.
11. Six coins of the same type are discovered at an archaeological site. If their weights on average are significantly different from 5.25 grams then it can be assumed that their provenance is not the site itself. The coins are weighed and have mean 4.73 g with sample standard deviation 0.18 g. Perform the relevant test at the 0.1% ($1/10\text{th of } 1\%$) level of significance, assuming a normal distribution of weights of all such coins.
12. An economist wishes to determine whether people are driving less than in the past. In one region of the country the number of miles driven per household per year in the past was 18.59 thousand miles. A sample of 15 households produced a sample mean of 16.23 thousand miles for the last year, with sample standard deviation 4.06 thousand miles. Assuming a normal distribution of household driving distances per year, perform the relevant test at the 5% level of significance.
13. The recommended daily allowance of iron for females aged $19-50$ is 18 mg/day. A careful measurement of the daily iron intake of 15 women yielded a mean daily intake of 16.2 mg with sample standard deviation 4.7 mg.
    a. Assuming that daily iron intake in women is normally distributed, perform the test that the actual mean daily intake for all women is different from 18 mg/day, at the 10% level of significance.
    b. The sample mean is less than 18, suggesting that the actual population mean is less than 18 mg/day. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
14. The target temperature for a hot beverage the moment it is dispensed from a vending machine is $170^\circ F$. A sample of ten randomly selected servings from a new machine undergoing a pre-shipment inspection gave mean temperature $173^\circ F$ with sample standard deviation $6.3^\circ F$.
    a. Assuming that temperature is normally distributed, perform the test that the mean temperature of dispensed beverages is different from $170^\circ F$, at the 10% level of significance.
    b. The sample mean is greater than 170, suggesting that the actual population mean is greater than $170^\circ F$. Perform this test, also at the 10% level of significance. (The computation of the test statistic done in part (a) still applies here.)
15. The average number of days to complete recovery from a particular type of knee operation is 123.7 days. From his experience a physician suspects that use of a topical pain medication might be lengthening the recovery time. He randomly selects the records of seven knee surgery patients who used the topical medication. The times to total recovery were:

$$128 \quad 135 \quad 121 \quad 142 \quad 126 \quad 151 \quad 123 \tag{8.E.1}$$

    a. Assuming a normal distribution of recovery times, perform the relevant test of hypotheses at the 10% level of significance.
    b. Would the decision be the same at the 5% level of significance? Answer either by constructing a new rejection region (critical value approach) or by estimating the p-value of the test in part (a) and comparing it to α.
16. A 24-hour advance prediction of a day's high temperature is "unbiased" if the long-term average of the error in prediction (true high temperature minus predicted high temperature) is zero. The errors in predictions made by one meteorological station for 20 randomly selected days were:

$$\begin{array}{ccccc} 2 & 0 & -3 & 1 & -2 \\ 1 & 0 & -1 & 1 & -1 \\ -4 & 1 & 1 & -4 & 0 \\ -4 & -3 & -4 & 2 & 2 \end{array} \tag{8.E.2}$$

    a. Assuming a normal distribution of errors, test the null hypothesis that the predictions are unbiased (the mean of the population of all errors is 0) versus the alternative that it is biased (the population mean is not 0), at the 1% level of significance.
    b. Would the decision be the same at the 5% level of significance? The 10% level of significance? Answer either by constructing new rejection regions (critical value approach) or by estimating the p-value of the test in part (a) and comparing it to α.
17. Pasteurized milk may not have a standardized plate count (SPC) above $20,000$ colony-forming bacteria per milliliter (cfu/ml). The mean SPC for five samples was $21,500$ cfu/ml with sample standard deviation 750 cfu/ml. Test the null hypothesis that the mean SPC for this milk is $20,000$ versus the alternative that it is greater than $20,000$, at the 10% level of significance. Assume that the SPC follows a normal distribution.

18. One water quality standard for water that is discharged into a particular type of stream or pond is that the average daily water temperature be at most $18^{\circ}F$. Six samples taken throughout the day gave the data:

$$16.8 \quad 21.5 \quad 19.1 \quad 12.8 \quad 18.0 \quad 20.7 \tag{8.E.3}$$

The sample mean exceeds $\bar{x} = 18.15$, but perhaps this is only sampling error. Determine whether the data provide sufficient evidence, at the 10% level of significance, to conclude that the mean temperature for the entire day exceeds $18^{\circ}F$.

Additional Exercises

19. A calculator has a built-in algorithm for generating a random number according to the standard normal distribution. Twenty-five numbers thus generated have mean 0.15 and sample standard deviation 0.94. Test the null hypothesis that the mean of all numbers so generated is 0 versus the alternative that it is different from 0, at the 20% level of significance. Assume that the numbers do follow a normal distribution.
20. At every setting a high-speed packing machine delivers a product in amounts that vary from container to container with a normal distribution of standard deviation 0.12 ounce. To compare the amount delivered at the current setting to the desired amount 64.1 ounce, a quality inspector randomly selects five containers and measures the contents of each, obtaining sample mean 63.9 ounces and sample standard deviation 0.10 ounce. Test whether the data provide sufficient evidence, at the 5% level of significance, to conclude that the mean of all containers at the current setting is less than 64.1 ounces.
21. A manufacturing company receives a shipment of $1,000$ bolts of nominal shear strength $4,350$ lb. A quality control inspector selects five bolts at random and measures the shear strength of each. The data are:

$$4,320 \quad 4,290 \quad 4,360 \quad 4,350 \quad 4,320 \tag{8.E.4}$$

    a. Assuming a normal distribution of shear strengths, test the null hypothesis that the mean shear strength of all bolts in the shipment is $4,350$ lb versus the alternative that it is less than $4,350$ lb, at the 10% level of significance.
    b. Estimate the p-value (observed significance) of the test of part (a).
    c. Compare the p-value found in part (b) to $\alpha = 0.10$ andmake a decision based on the p-value approach. Explain fully.
22. A literary historian examines a newly discovered document possibly written by Oberon Theseus. The mean average sentence length of the surviving undisputed works of Oberon Theseus is 48.72 words. The historian counts words in sentences between five successive 101 periods in the document in question to obtain a mean average sentence length of 39.46 words with standard deviation 7.45 words. (Thus the sample size is five.)

    a. Determine if these data provide sufficient evidence, at the 1% level of significance, to conclude that the mean average sentence length in the document is less than 48.72.
    b. Estimate the p-value of the test.
    c. Based on the answers to parts (a) and (b), state whether or not it is likely that the document was written by Oberon Theseus.

Answers

1. a. $Z \leq -1.645$
   b. $T \leq -2.571$ *or* $T \geq 2.571$
   c. $T \geq 1.319$
   d. $Z \leq -1645$ *or* $Z \geq 1.645$
2.
3. a. $T \leq -0.855$
   b. $Z \leq -1.645$
   c. $T \leq -2.201$ *or* $T \geq 2.201$
   d. $T \geq 3.435$
4.
5. a. $T = -2.690,\ df = 19,\ -t_{0.005} = -2.861,$ do not reject H_0
   b. $0.01 < p - value < 0.02,\ \alpha = 0.01,$ do not reject H_0
6.
7. a. $T = 2.398,\ df = 7,\ t_{0.05} = 1.895,$ reject H_0
   b. $0.01 < p - value < 0.025,\ \alpha = 0.05,$ reject H_0

8.
9. $T = -7.560,\ df = 12,\ -t_{0.10} = -1.356,$ reject H_0
10.
11. $T = -7.076,\ df = 5,\ -t_{0.0005} = -6.869,$ reject H_0
12.
13. a. $T = -1.483,\ df = 14,\ -t_{0.05} = -1.761,$ do not reject H_0
    b. $T = -1.483,\ df = 14,\ -t_{0.10} = -1.345,$ reject H_0
14.
15. a. $T = 2.069,\ df = 6,\ t_{0.10} = 1.44,$ reject H_0
    b. $T = 2.069,\ df = 6,\ t_{0.05} = 1.943,$ reject H_0
16.
17. $T = 4.472,\ df = 4,\ t_{0.10} = 1.533,$ reject H_0
18.
19. $T = 0.798,\ df = 24,\ t_{0.10} = 1.318,$ do not reject H_0
20.
21. a. $T = -1.773,\ df = 4,\ -t_{0.05} = -2.132,$ do not reject H_0
    b. $0.05 < p - value < 0.10$
    c. $\alpha = 0.05,$ do not reject H_0

8.5: Large Sample Tests for a Population Proportion

Basic

On all exercises for this section you may assume that the sample is sufficiently large for the relevant test to be validly performed.

1. Compute the value of the test statistic for each test using the information given.
   a. Testing $H_0 : p = 0.50\ vs\ H_a : p > 0.50,\ n = 360,\ \hat{p} = 0.56.$
   b. Testing $H_0 : p = 0.50\ vs\ H_a : p \neq 0.50,\ n = 360,\ \hat{p} = 0.56.$
   c. Testing $H_0 : p = 0.37\ vs\ H_a : p < 0.37,\ n = 1200,\ \hat{p} = 0.35.$
2. Compute the value of the test statistic for each test using the information given.
   a. Testing $H_0 : p = 0.72\ vs\ H_a : p < 0.72,\ n = 2100,\ \hat{p} = 0.71.$
   b. Testing $H_0 : p = 0.83\ vs\ H_a : p \neq 0.83,\ n = 500,\ \hat{p} = 0.86.$
   c. Testing $H_0 : p = 0.22\ vs\ H_a : p < 0.22,\ n = 750,\ \hat{p} = 0.18.$
3. For each part of Exercise 1 construct the rejection region for the test for $\alpha = 0.05$ and make the decision based on your answer to that part of the exercise.
4. For each part of Exercise 2 construct the rejection region for the test for $\alpha = 0.05$ and make the decision based on your answer to that part of the exercise.
5. For each part of Exercise 1 compute the observed significance (p-value) of the test and compare it to $\alpha = 0.05$ in order to make the decision by the p-value approach to hypothesis testing.
6. For each part of Exercise 2 compute the observed significance (p-value) of the test and compare it to $\alpha = 0.05$ in order to make the decision by the p-value approach to hypothesis testing.
7. Perform the indicated test of hypotheses using the critical value approach.
   a. Testing $H_0 : p = 0.55\ vs\ H_a : p > 0.55\ @\ \alpha = 0.05,\ n = 300,\ \hat{p} = 0.60.$
   b. Testing $H_0 : p = 0.47\ vs\ H_a : p \neq 0.47\ @\ \alpha = 0.01,\ n = 9750,\ \hat{p} = 0.46.$
8. Perform the indicated test of hypotheses using the critical value approach.
   a. Testing $H_0 : p = 0.15\ vs\ H_a : p \neq 0.15\ @\ \alpha = 0.001,\ n = 1600,\ \hat{p} = 0.18.$
   b. Testing $H_0 : p = 0.90\ vs\ H_a : p > 0.90\ @\ \alpha = 0.01,\ n = 1100,\ \hat{p} = 0.91.$
9. Perform the indicated test of hypotheses using the p-value approach.
   a. Testing $H_0 : p = 0.37\ vs\ H_a : p \neq 0.37\ @\ \alpha = 0.005,\ n = 1300,\ \hat{p} = 0.40.$
   b. Testing $H_0 : p = 0.94\ vs\ H_a : p > 0.94\ @\ \alpha = 0.05,\ n = 1200,\ \hat{p} = 0.96.$
10. Perform the indicated test of hypotheses using the p-value approach.

a. Testing $H_0 : p = 0.25 \; vs \; H_a : p < 0.25 \; @ \; \alpha = 0.10, \; n = 850, \; \hat{p} = 0.23$.
b. Testing $H_0 : p = 0.33 \; vs \; H_a : p \neq 0.33 \; @ \; \alpha = 0.05, \; n = 1100, \; \hat{p} = 0.30$.

Applications

11. Five years ago 3.9% of children in a certain region lived with someone other than a parent. A sociologist wishes to test whether the current proportion is different. Perform the relevant test at the 5% level of significance using the following data: in a random sample of $2,759$ children, 119 lived with someone other than a parent.
12. The government of a particular country reports its literacy rate as 52%. A nongovernmental organization believes it to be less. The organization takes a random sample of 600 inhabitants and obtains a literacy rate of 42%. Perform the relevant test at the 0.5% (one-half of 1%) level of significance.
13. Two years ago 72% of household in a certain county regularly participated in recycling household waste. The county government wishes to investigate whether that proportion has increased after an intensive campaign promoting recycling. In a survey of 900 households, 674 regularly participate in recycling. Perform the relevant test at the 10% level of significance.
14. Prior to a special advertising campaign, 23% of all adults recognized a particular company's logo. At the close of the campaign the marketing department commissioned a survey in which 311 of $1,200$ randomly selected adults recognized the logo. Determine, at the 1% level of significance, whether the data provide sufficient evidence to conclude that more than 23% of all adults now recognize the company's logo.
15. A report five years ago stated that 35.5% of all state-owned bridges in a particular state were "deficient." An advocacy group took a random sample of 100 state-owned bridges in the state and found 33 to be currently rated as being "deficient." Test whether the current proportion of bridges in such condition is 35.5% versus the alternative that it is different from 35.5%, at the 10% level of significance.
16. In the previous year the proportion of deposits in checking accounts at a certain bank that were made electronically was 45%. The bank wishes to determine if the proportion is higher this year. It examined $20,000$ deposit records and found that $9,217$ were electronic. Determine, at the 1% level of significance, whether the data provide sufficient evidence to conclude that more than 45% of all deposits to checking accounts are now being made electronically.
17. According to the Federal Poverty Measure 12% of the U.S. population lives in poverty. The governor of a certain state believes that the proportion there is lower. In a sample of size $1,550,163$ were impoverished according to the federal measure.
    a. Test whether the true proportion of the state's population that is impoverished is less than 12%, at the 5% level of significance.
    b. Compute the observed significance of the test.
18. An insurance company states that it settles 85% of all life insurance claims within 30 days. A consumer group asks the state insurance commission to investigate. In a sample of 250 life insurance claims, 203 were settled within 30 days.
    a. Test whether the true proportion of all life insurance claims made to this company that are settled within 30 days is less than 85%, at the 5% level of significance.
    b. Compute the observed significance of the test.
19. A special interest group asserts that 90% of all smokers began smoking before age 18. In a sample of 850 smokers, 687 began smoking before age 18.
    a. Test whether the true proportion of all smokers who began smoking before age 18 is less than 90%, at the 1% level of significance.
    b. Compute the observed significance of the test.
20. In the past, 68% of a garage's business was with former patrons. The owner of the garage samples 200 repair invoices and finds that for only 114 of them the patron was a repeat customer.
    a. Test whether the true proportion of all current business that is with repeat customers is less than 68%, at the 1% level of significance.
    b. Compute the observed significance of the test.

Additional Exercises

21. A rule of thumb is that for working individuals one-quarter of household income should be spent on housing. A financial advisor believes that the average proportion of income spent on housing is more than 0.25. In a sample of 30 households, the mean proportion of household income spent on housing was 0.285 with a standard deviation of 0.063. Perform the relevant test of hypotheses at the 1% level of significance. Hint: This exercise could have been presented in an earlier section.

22. Ice cream is legally required to contain at least 10% milk fat by weight. The manufacturer of an economy ice cream wishes to be close to the legal limit, hence produces its ice cream with a target proportion of 0.106 milk fat. A sample of five containers yielded a mean proportion of 0.094 milk fat with standard deviation 0.002. Test the null hypothesis that the mean proportion of milk fat in all containers is 0.106 against the alternative that it is less than 0.106, at the 10% level of significance. Assume that the proportion of milk fat in containers is normally distributed. Hint: This exercise could have been presented in an earlier section.

Large Data Set Exercises

Large Data Sets missing

23. Large Data Sets 4 and 4A list the results of 500 tosses of a die. Let p denote the proportion of all tosses of this die that would result in a five. Use the sample data to test the hypothesis that p is different from $1/6$, at the 20% level of significance.
24. Large Data Set 6 records results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Use the full data set (400 observations) to test the hypothesis that the proportion p of all voters who prefer Candidate A exceeds 0.35. Test at the 10% level of significance.
25. Lines 2 through 536 in Large Data Set 11 is a sample of 535 real estate sales in a certain region in 2008. Those that were foreclosure sales are identified with a 1 in the second column. Use these data to test, at the 10% level of significance, the hypothesis that the proportion p of all real estate sales in this region in 2008 that were foreclosure sales was less than 25%. (The null hypothesis is $H_0 : p = 0.25$).
26. Lines 537 through 1106 in Large Data Set 11 is a sample of 570 real estate sales in a certain region in 2010. Those that were foreclosure sales are identified with a 1 in the second column. Use these data to test, at the 5% level of significance, the hypothesis that the proportion p of all real estate sales in this region in 2010 that were foreclosure sales was greater than 23%. (The null hypothesis is $H_0 : p = 0.25$).

Answers

1. a. $Z = 2.277$
   b. $Z = 2.277$
   c. $Z = -1.435$
2.
3. a. $Z \geq 1.645$; reject H_0
   b. $Z \leq -1.96$ *or* $Z \geq 1.96$; reject H_0
   c. $Z \leq -1.645$; do not reject H_0
4.
5. a. $p - value = 0.0116,\ \alpha = 0.05$; reject H_0
   b. $p - value = 0.0232,\ \alpha = 0.05$; reject H_0
   c. $p - value = 0.0749,\ \alpha = 0.05$; do not reject H_0
6.
7. a. $Z = 1.74,\ z_{0.05} = 1.645$; reject H_0
   b. $Z = -1.98,\ -z_{0.005} = -2.576$; do not reject H_0
8.
9. a. $Z = 2.24,\ p - value = 0.025, \alpha = 0.005$; do not reject H_0
   b. $Z = 2.92,\ p - value = 0.0018, \alpha = 0.05$; reject H_0
10.
11. $Z = 1.11,\ z_{0.025} = 1.96$; do not reject H_0
12.
13. $Z = 1.93,\ z_{0.10} = 1.28$; reject H_0
14.
15. $Z = -0.523,\ \pm z_{0.05} = \pm 1.645$; do not reject H_0
16.
17. a. $Z = -1.798,\ -z_{0.05} = -1.645$; reject H_0
    b. $p - value = 0.0359$

18.
19. a. $Z = -8.92,\ -z_{0.01} = -2.33$; reject H_0
    b. $p - value \approx 0$
20.
21. $Z = 3.04,\ z_{0.01} = 2.33$; reject H_0
22.
23. $H_0 : p = 1/6\ vs\ H_a : p \neq 1/6$. Test Statistic: $Z = -0.76$. Rejection Region: $(-\infty, -1.28] \cup [1.28, \infty)$. Decision: Fail to reject H_0.
24.
25. $H_0 : p = 0.25\ vs\ H_a : p < 0.25$. Test Statistic: $Z = -1.17$. Rejection Region: $(-\infty, -1.28]$. Decision: Fail to reject H_0.

Contributor

- Anonymous

This page titled 8.E: Testing Hypotheses (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **8.E: Testing Hypotheses (Exercises)** has no license indicated.

CHAPTER OVERVIEW

9: Two-Sample Problems

The previous two chapters treated the questions of estimating and making inferences about a parameter of a single population. In this chapter we consider a comparison of parameters that belong to two different populations. For example, we might wish to compare the average income of all adults in one region of the country with the average income of those in another region, or we might wish to compare the proportion of all men who are vegetarians with the proportion of all women who are vegetarians. We will study construction of confidence intervals and tests of hypotheses in four situations, depending on the parameter of interest, the sizes of the samples drawn from each of the populations, and the method of sampling. We also examine sample size considerations.

9.1: Comparison of Two Population Means- Large, Independent Samples

9.2: Comparison of Two Population Means - Small, Independent Samples

9.3: Comparison of Two Population Means - Paired Samples

9.4: Comparison of Two Population Proportions

9.5: Sample Size Considerations

9.E: Two-Sample Problems (Exercises)

This page titled 9: Two-Sample Problems is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Comparison of Two Population Means- Large, Independent Samples

Learning Objectives

- To understand the logical framework for estimating the difference between the means of two distinct populations and performing tests of hypotheses concerning those means.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using large, independent samples.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using large, independent samples.

Suppose we wish to compare the means of two distinct populations. Figure 9.1.1 illustrates the conceptual framework of our investigation in this and the next section. Each population has a mean and a standard deviation. We arbitrarily label one population as Population 1 and the other as Population 2, and subscript the parameters with the numbers 1 and 2 to tell them apart. We draw a random sample from Population 1 and label the sample statistics it yields with the subscript 1. Without reference to the first sample we draw a sample from Population 2 and label its sample statistics with the subscript 2.

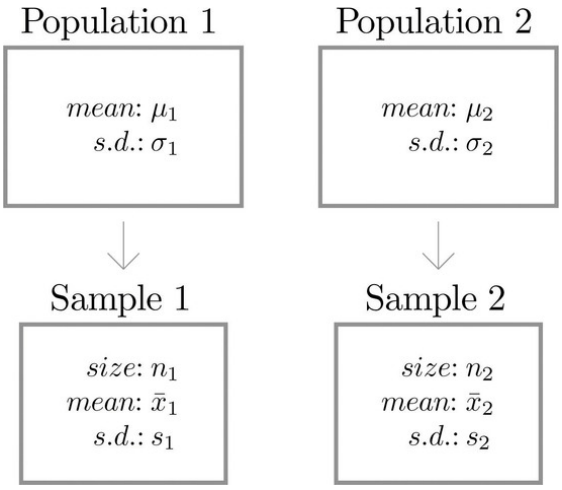


Figure 9.1.1: Independent Sampling from Two Populations

Definition: Independence

Samples from two distinct populations are *independent* if each one is drawn without reference to the other, and has no connection with the other.

Our goal is to use the information in the samples to estimate the difference $\mu_1 - \mu_2$ in the means of the two populations and to make statistically valid inferences about it.

Confidence Intervals

Since the mean $x - 1$ of the sample drawn from Population 1 is a good estimator of μ_1 and the mean $x - 2$ of the sample drawn from Population 2 is a good estimator of μ_2, a reasonable point estimate of the difference $\mu_1 - \mu_2$ is $\bar{x_1} - \bar{x_2}$. In order to widen this point estimate into a confidence interval, we first suppose that both samples are large, that is, that both $n_1 \geq 30$ and $n_2 \geq 30$. If so, then the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid. The symbols s_1^2 and s_2^2 denote the squares of s_1 and s_2. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

$100(1 - \alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Large, Independent Samples

The samples must be independent, and *each* sample must be large:

Example 9.1.1

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1 = 174$	$n_2 = 355$
$x - 1 = 3.51$	$x - 2 = 3.24$
$s_1 = 0.51$	$s_2 = 0.52$

Construct a point estimate and a 99% confidence interval for $\mu_1 - \mu_2$, the difference in average satisfaction levels of customers of the two companies as measured on this five-point scale.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x_1} - \bar{x_2} = 3.51 - 3.24 = 0.27$$

In words, we estimate that the average customer satisfaction level for Company 1 is 0.27 points higher on this five-point scale than it is for Company 2.

To apply the formula for the confidence interval, proceed exactly as was done in Chapter 7. The 99% confidence level means that $\alpha = 1 - 0.99 = 0.01$ so that $z_{\alpha/2} = z_{0.005}$. From Figure 7.1.6 "Critical Values of " we read directly that $z_{0.005} = 2.576$. Thus

$$(\bar{x_1} - \bar{x_2}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.27 \pm 2.576\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}} = 0.27 \pm 0.12$$

We are 99% confident that the difference in the population means lies in the interval $[0.15, 0.39]$, in the sense that in repeated sampling 99% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. In the context of the problem we say we are 99% confident that the average level of customer satisfaction for Company 1 is between 0.15 and 0.39 points higher, on this five-point scale, than that for Company 2.

Hypothesis Testing

Hypotheses concerning the relative sizes of the means of two populations are tested using the same critical value and p-value procedures that were used in the case of a single population. All that is needed is to know how to express the null and alternative hypotheses and to know the formula for the standardized test statistic and the distribution that it follows.

The null and alternative hypotheses will always be expressed in terms of the difference of the two population means. Thus the null hypothesis will always be written

$$H_0 : \mu_1 - \mu_2 = D_0$$

where D_0 is a number that is deduced from the statement of the situation. As was the case with a single population the alternative hypothesis can take one of the three forms, with the same terminology:

Form of H*a*	Terminology
$H_a : \mu_1 - \mu_2 < D_0$	Left-tailed
$H_a : \mu_1 - \mu_2 > D_0$	Right-tailed
$H_a : \mu_1 - \mu_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution. (In the relatively rare case that both population standard deviations σ_1 and σ_2 are known they would be used instead of the sample standard deviations.)

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Large, Independent Samples

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: $n_1 \geq 30$ and $n_2 \geq 30$.

Example 9.1.2

Refer to Example 9.1.1 concerning the mean satisfaction levels of customers of two competing cable television companies. Test at the 1% level of significance whether the data provide sufficient evidence to conclude that Company 1 has a higher mean satisfaction rating than does Company 2. Use the critical value approach.

Solution:

- **Step 1**. If the mean satisfaction levels μ_1 and μ_2 are the same then $\mu_1 = \mu_2$, but we always express the null hypothesis in terms of the difference between μ_1 and μ_2, hence H_0 is $\mu_1 - \mu_2 = 0$. To say that the mean customer satisfaction for Company 1 is higher than that for Company 2 means that $\mu_1 > \mu_2$, which in terms of their difference is $\mu_1 - \mu_2 > 0$. The test is therefore

$$H_0 : \mu_1 - \mu_2 = 0$$

$$vs.$$

$$H_a : \mu_1 - \mu_2 > 0 \ \ @ \ \ \alpha = 0.01$$

- **Step 2**. Since the samples are independent and both are large the test statistic is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Step 3**. Inserting the data into the formula for the test statistic gives

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.51 - 3.24) - 0}{\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}} = 5.684$$

- **Step 4**. Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $z_\alpha = z_{0.01}$, which from the last line in Figure 7.1.6 "Critical Values of " we read off as 2.326. The rejection region is $[2.326, \infty)$.

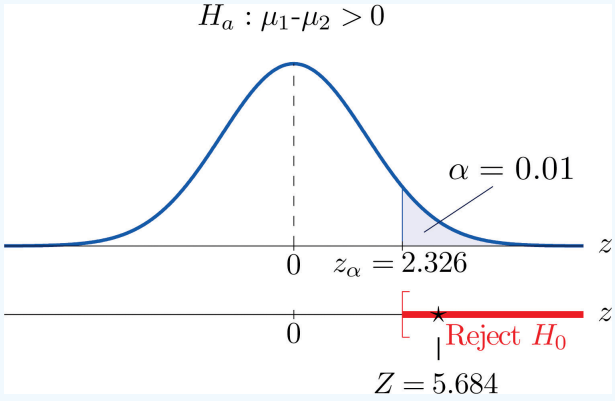


Figure 9.1.2: *Rejection Region and Test Statistic for Example* 9.1.2

- **Step 5**. As shown in Figure 9.1.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 1% level of significance, to conclude that the mean customer satisfaction for Company 1 is higher than that for Company 2.

Example 9.1.3

Perform the test of Example $9.1.2$ using the p-value approach.

Solution:

The first three steps are identical to those in Example $9.1.2$

- **Step 4**. The observed significance or p-value of the test is the area of the right tail of the standard normal distribution that is cut off by the test statistic $Z = 5.684$. The number 5.684 is too large to appear in Figure 7.1.5, which means that the area of the left tail that it cuts off is 1.0000 to four decimal places. The area that we seek, the area of the right tail, is therefore $1 - 1.0000 = 0.0000$ to four decimal places. See Figure $9.1.3$. That is, p-value=0.0000 to four decimal places. (The actual value is approximately 0.000000007.)

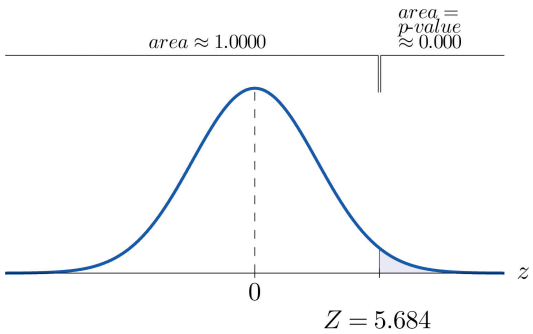


Figure $9.1.3$: *P-Value for Example* $9.1.3$

- **Step 5**. Since $0.0000 < 0.01$, $p - value < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the 1% level of significance, to conclude that the mean customer satisfaction for Company 1 is higher than that for Company 2.

Key Takeaway

- A point estimate for the difference in two population means is simply the difference in the corresponding sample means.
- In the context of estimating or testing hypotheses concerning two population means, “large” samples means that both samples are large.
- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means. The only difference is in the formula for the standardized test statistic.

This page titled 9.1: Comparison of Two Population Means- Large, Independent Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.2: Comparison of Two Population Means - Small, Independent Samples

Learning Objectives

- To learn how to construct a confidence interval for the difference in the means of two distinct populations using small, independent samples.
- To learn how to perform a test of hypotheses concerning the difference between the means of two distinct populations using small, independent samples.

When one or the other of the sample sizes is small, as is often the case in practice, the Central Limit Theorem does not apply. We must then impose conditions on the population to give statistical validity to the test procedure. We will assume that both populations from which the samples are taken have a normal probability distribution and that their standard deviations are equal.

Confidence Intervals

When the two populations are normally distributed and have equal standard deviations, the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Small, Independent Samples

$$(\bar{x_1} - \bar{x_2}) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{9.2.1}$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

The number of degrees of freedom is

$$df = n_1 + n_2 - 2.$$

The samples must be independent, the populations must be normal, and the population standard deviations must be equal. "Small" samples means that either $n_1 < 30$ or $n_2 < 30$.

The quantity s_p^2 is called the **pooled sample variance**. It is a weighted average of the two estimates s_1^2 and s_2^2 of the common variance $\sigma_1^2 = \sigma_2^2$ of the two populations.

Example 9.2.1

A software company markets a new computer game with two experimental packaging designs. Design 1 is sent to 11 stores; their average sales the first month is 52 units with sample standard deviation 12 units. Design 2 is sent to 6 stores; their average sales the first month is 46 units with sample standard deviation 10 units. Construct a point estimate and a 95% confidence interval for the difference in average monthly sales between the two package designs.

Solution

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x_1} - \bar{x_2} = 52 - 46 - 6$$

In words, we estimate that the average monthly sales for Design 1 is 6 units more per month than the average monthly sales for Design 2.

To apply the formula for the confidence interval (Equation 9.2.1), we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = 11 + 6 - 2 = 15$ we read that $t_{0.025} = 2.131$. From the formula for the pooled sample variance we compute

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(10)(12)^2 + (5)(10)^2}{15} = 129.\overline{3}$$

Thus

$$(\bar{x_1} - \bar{x_2}) \pm t_\alpha/2\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 6 \pm (2.131)\sqrt{129.\bar{3}\left(\frac{1}{11} + \frac{1}{6}\right)} \approx 6 \pm 12.3$$

We are 95% confident that the difference in the population means lies in the interval $[-6.3, 18.3]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. Because the interval contains both positive and negative values the statement in the context of the problem is that we are 95% confident that the average monthly sales for Design 1 is between 18.3 units higher and 6.3 units lower than the average monthly sales for Design 2.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means using small samples is done precisely as it is done for large samples, using the following standardized test statistic. The same conditions on the populations that were required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Small, Independent Samples

$$T = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The test statistic has Student's t-distribution with $df = n_1 + n_2 - 2$ degrees of freedom.

The samples must be independent, the populations must be normal, and the population standard deviations must be equal. "Small" samples means that either $n_1 < 30$ or $n_2 < 30$.

Example 9.2.2

Refer to Example $9.2.1$ concerning the mean sales per month for the same computer game but sold with two package designs. Test at the 1% level of significance whether the data provide sufficient evidence to conclude that the mean sales per month of the two designs are different. Use the critical value approach.

Solution

- **Step 1**. The relevant test is

$$H_0 : \mu_1 - \mu_2 = 0$$

vs.

$$H_a : \mu_1 - \mu_2 \neq 0 \ \ @ \ \ \alpha = 0.01$$

- **Step 2**. Since the samples are independent and at least one is less than 30 the test statistic is

$$T = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which has Student's t-distribution with $df = 11 + 6 - 2 = 15$ degrees of freedom.

- **Step 3**. Inserting the data and the value $D_0 = 0$ into the formula for the test statistic gives

$$T = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$= \frac{(52 - 46) - 0}{\sqrt{129.\bar{3}\left(\frac{1}{11} + \frac{1}{6}\right)}}$$
$$= 1.040$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values, $\pm t_{\alpha/2} = \pm t_{0.005}$. From the row in Figure 7.1.6 with the heading $df = 15$ we read off $t_{0.005} = 2.947$. The rejection region is $(-\infty, -2.947] \cup [2.947, \infty)$.

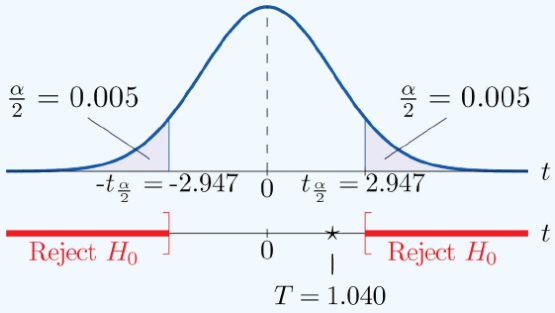


Figure 9.2.1: Rejection Region and Test Statistic for "Example 9.2.2"

- **Step 5**. As shown in Figure 9.2.1 the test statistic does not fall in the rejection region. The decision is not to reject H_0. In the context of the problem our conclusion is:

  The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean sales per month of the two designs are different.

Example 9.2.3

Perform the test of Example 9.2.2 using the p-value approach.

Solution

The first three steps are identical to those in Example 9.2.2.

- **Step 4**. Because the test is two-tailed the observed significance or p-value of the test is the double of the area of the right tail of Student'st-distribution, with 15 degrees of freedom, that is cut off by the test statistic $T = 1.040$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 15$, the number 1.040 is between the numbers 0.866 and 1.341, corresponding to $t_{0.200}$ and $t_{0.100}$. The area cut off by $t = 0.866$ is 0.200 and the area cut off by $t = 1.341$ is 0.100. Since 1.040 is between 0.866 and 1.341 the area it cuts off is between 0.200 and 0.100. Thus the p-value (since the area must be doubled) is between 0.400 and 0.200.
- **Step 5**. Since $p > 0.200 > 0.01,\ \ p > \alpha$, so the decision is not to reject the null hypothesis:

  The data do not provide sufficient evidence, at the 1% level of significance, to conclude that the mean sales per month of the two designs are different.

Key Takeaway

- In the context of estimating or testing hypotheses concerning two population means, "small" samples means that at least one sample is small. In particular, even if one sample is of size 30 or more, if the other is of size less than 30 the formulas of this section must be used.
- A confidence interval for the difference in two population means is computed using a formula in the same fashion as was done for a single population mean.

This page titled 9.2: Comparison of Two Population Means - Small, Independent Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.3: Comparison of Two Population Means - Paired Samples

Learning Objectives

- To learn the distinction between independent samples and paired samples.
- To learn how to construct a confidence interval for the difference in the means of two distinct populations using paired samples.
- To learn how to perform a test of hypotheses concerning the difference in the means of two distinct populations using paired samples

Suppose chemical engineers wish to compare the fuel economy obtained by two different formulations of gasoline. Since fuel economy varies widely from car to car, if the mean fuel economy of two independent samples of vehicles run on the two types of fuel were compared, even if one formulation were better than the other the large variability from vehicle to vehicle might make any difference arising from difference in fuel difficult to detect. Just imagine one random sample having many more large vehicles than the other. Instead of independent random samples, it would make more sense to select pairs of cars of the same make and model and driven under similar circumstances, and compare the fuel economy of the two cars in each pair. Thus the data would look something like Table 9.3.1, where the first car in each pair is operated on one formulation of the fuel (call it Type 1 gasoline) and the second car is operated on the second (call it Type 2 gasoline).

Table 9.3.1: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2
Buick LaCrosse	17.0	17.0
Dodge Viper	13.2	12.9
Honda CR-Z	35.3	35.4
Hummer H 3	13.6	13.2
Lexus RX	32.7	32.5
Mazda CX-9	18.4	18.1
Saab 9-3	22.5	22.5
Toyota Corolla	26.8	26.7
Volvo XC 90	15.1	15.0

The first column of numbers form a sample from Population 1, the population of all cars operated on Type 1 gasoline; the second column of numbers form a sample from Population 2, the population of all cars operated on Type 2 gasoline. It would be incorrect to analyze the data using the formulas from the previous section, however, since the samples were not drawn independently. What is correct is to compute the difference in the numbers in each pair (subtracting in the same order each time) to obtain the third column of numbers as shown in Table 9.3.2 and treat the differences as the data. At this point, the new sample of differences $d_1 = 0.0, \cdots, d_9 = 0.1$ in the third column of Table 9.3.2 may be considered as a random sample of size $n = 9$ selected from a population with mean $\mu_d = \mu_1 - \mu_2$. This approach essentially transforms the paired two-sample problem into a one-sample problem as discussed in the previous two chapters.

Table 9.3.2: Fuel Economy of Pairs of Vehicles

Make and Model	Car 1	Car 2	Difference
Buick LaCrosse	17.0	17.0	0.0
Dodge Viper	13.2	12.9	0.3
Honda CR-Z	35.3	35.4	−0.1
Hummer H 3	13.6	13.2	0.4

Make and Model	Car 1	Car 2	Difference
Lexus RX	32.7	32.5	0.2
Mazda CX-9	18.4	18.1	0.3
Saab 9-3	22.5	22.5	0.0
Toyota Corolla	26.8	26.7	0.1
Volvo XC 90	15.1	15.0	0.1

Note carefully that although it does not matter what order the subtraction is done, it must be done in the same order for all pairs. This is why there are both positive and negative quantities in the third column of numbers in Table 9.3.2.

Confidence Intervals

When the population of differences is normally distributed the following formula for a confidence interval for $\mu_d = \mu_1 - \mu_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Means: Paired Difference Samples

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The number of degrees of freedom is

$$df = n - 1.$$

The population of differences must be *normally distributed*.

Example 9.3.1

Using the data in Table 9.3.1 construct a point estimate and a 95% confidence interval for the difference in average fuel economy between cars operated on Type 1 gasoline and cars operated on Type 2 gasoline.

Solution

We have referred to the data in Table 9.3.1 because that is the way that the data are typically presented, but we emphasize that with paired sampling one immediately computes the differences, as given in Table 9.3.2, and uses the differences as the data.

The mean and standard deviation of the differences are

$$\bar{d} = \frac{\sum d}{n} = \frac{1.3}{9} = 0.1\bar{4}$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{1}{n}(\sum d)^2}{n-1}} = \sqrt{\frac{0.41 - \frac{1}{9}(1.3)^2}{8}} = 0.1\bar{6}$$

The point estimate of $\mu_1 - \mu_2 = \mu_d$ is

$$\bar{d} = 0.14$$

In words, we estimate that the average fuel economy of cars using Type 1 gasoline is 0.14 mpg greater than the average fuel economy of cars using Type 2 gasoline.

To apply the formula for the confidence interval, we must find $t_{\alpha/2}$. The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $t_{\alpha/2} = t_{0.025}$. From Figure 7.1.6, in the row with the heading $df = 9 - 1 = 8$ we read that $t_{0.025} = 2.306$. Thus

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}} = 0.14 \pm 2.306\left(\frac{0.1\bar{6}}{\sqrt{9}}\right) \approx 0.14 \pm 0.13$$

We are 95% confident that the difference in the population means lies in the interval $[0.01, 0.27]$, in the sense that in repeated sampling 95% of all intervals constructed from the sample data in this manner will contain $\mu_d = \mu_1 - \mu_2$. Stated differently, we are 95% confident that mean fuel economy is between 0.01 and 0.27 mpg greater with Type 1 gasoline than with Type 2 gasoline.

Hypothesis Testing

Testing hypotheses concerning the difference of two population means using paired difference samples is done precisely as it is done for independent samples, although now the null and alternative hypotheses are expressed in terms of μ_d instead of $\mu_1 - \mu_2$. Thus the null hypothesis will always be written

$$H_0 : \mu_d = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \mu_d < D_0$	Left-tailed
$H_a : \mu_d > D_0$	Right-tailed
$H_a : \mu_d \neq D_0$	Two-tailed

The same conditions on the population of differences that was required for constructing a confidence interval for the difference of the means must also be met when hypotheses are tested. Here is the standardized test statistic that is used in the test.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Means: Paired Difference Samples

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

where there are n pairs, $\bar{d}$ is the mean and s_d is the standard deviation of their differences.

The test statistic has Student's t-distribution with $df = n - 1$ degrees of freedom.

The population of differences must be normally distributed.

Example $9.3.2$: using the critical value approach

Using the data of Table $9.3.2$ test the hypothesis that mean fuel economy for Type 1 gasoline is greater than that for Type 2 gasoline against the null hypothesis that the two formulations of gasoline yield the same mean fuel economy. Test at the 5% level of significance using the critical value approach.

Solution

The only part of the table that we use is the third column, the differences.

- **Step 1**. Since the differences were computed in the order $\text{Type } 1 \text{ mpg} - \text{Type } 2 \text{ mpg}$, better fuel economy with Type 1 fuel corresponds to $\mu_d = \mu_1 - \mu_2 > 0$. Thus the test is

$$H_0 : \mu_d = 0$$
$$\text{vs.}$$
$$H_a : \mu_d > 0 \ @ \ \alpha = 0.05$$

(If the differences had been computed in the opposite order then the alternative hypotheses would have been $H_a : \mu_d < 0$.)

- **Step 2**. Since the sampling is in pairs the test statistic is

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

- **Step 3**. We have already computed $\bar{d}$ and s_d in the previous example. Inserting their values and $D_0 = 0$ into the formula for the test statistic gives

$$\frac{\bar{d} - D_0}{s_d/\sqrt{n}} = 0.14$$

$$T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{0.14}{0.16/\sqrt{3}} = 2.600$$

- **Step 4**. Since the symbol in H_a is "$>$" this is a right-tailed test, so there is a single critical value, $t_\alpha = t_{0.05}$ with 8 degrees of freedom, which from the row labeled $df = 8$ in Figure 7.1.6 we read off as 1.860. The rejection region is $[1.860, \infty)$.
- **Step 5**. As shown in Figure $9.3.1$ the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

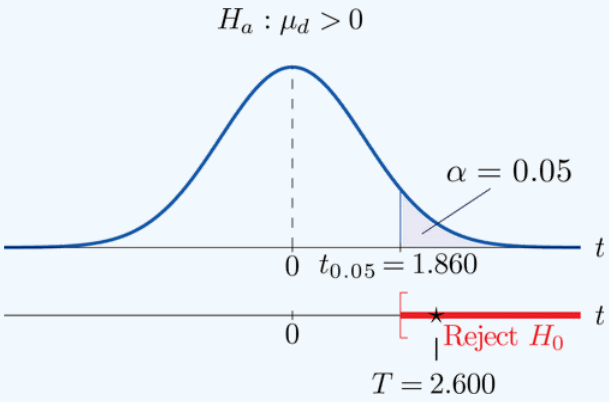


Figure $9.3.1$: Rejection Region and Test Statistic for "Example $9.3.2$"

The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

Example $9.3.3$: using the p-value approach

Perform the test in Example $9.3.2$ using the p-value approach.

Solution

The first three steps are identical to those $9.3.2$.

- **Step 4**. Because the test is one-tailed the observed significance or p-value of the test is just the area of the right tail of Student's t-distribution, with 8 degrees of freedom, that is cut off by the test statistic $T = 2.600$. We can only approximate this number. Looking in the row of Figure 7.1.6 headed $df = 8$, the number 2.600 is between the numbers 2.306 and 2.896, corresponding to $t_{0.025}$ and $t_{0.010}$. The area cut off by $t = 2.306$ is 0.025 and the area cut off by $t = 2.896$ is 0.010. Since 2.600 is between 2.306 and 2.896 the area it cuts off is between 0.025 and 0.010. Thus the p-value is between 0.025 and 0.010. In particular it is less than 0.025. See Figure $9.3.2$.

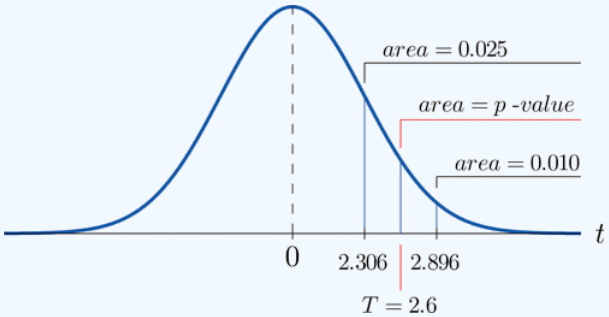


Figure $9.3.2$: P-Value for "Example $9.3.3$"

- **Step 5**. Since $0.025 < 0.05$, $p < \alpha$ so the decision is to reject the null hypothesis:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.

The paired two-sample experiment is a very powerful study design. It bypasses many unwanted sources of "statistical noise" that might otherwise influence the outcome of the experiment, and focuses on the possible difference that might arise from the one factor of interest.

If the sample is large (meaning that $n \geq 30$) then in the formula for the confidence interval we may replace $t_{\alpha/2}$ by $z_{\alpha/2}$. For hypothesis testing when the number of pairs is at least 30, we may use the same statistic as for small samples for hypothesis testing, except now it follows a standard normal distribution, so we use the last line of Figure 7.1.6 to compute critical values, and p-values can be computed exactly with Figure 7.1.5, not merely estimated using Figure 7.1.6.

Key Takeaway

- When the data are collected in pairs, the differences computed for each pair are the data that are used in the formulas.
- A confidence interval for the difference in two population means using paired sampling is computed using a formula in the same fashion as was done for a single population mean.
- The same five-step procedure used to test hypotheses concerning a single population mean is used to test hypotheses concerning the difference between two population means using pair sampling. The only difference is in the formula for the standardized test statistic.

This page titled 9.3: Comparison of Two Population Means - Paired Samples is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.4: Comparison of Two Population Proportions

Learning Objectives

- To learn how to construct a confidence interval for the difference in the proportions of two distinct populations that have a particular characteristic of interest.
- To learn how to perform a test of hypotheses concerning the difference in the proportions of two distinct populations that have a particular characteristic of interest.

Suppose we wish to compare the proportions of two populations that have a specific characteristic, such as the proportion of men who are left-handed compared to the proportion of women who are left-handed. Figure 9.4.1 illustrates the conceptual framework of our investigation. Each population is divided into two groups, the group of elements that have the characteristic of interest (for example, being left-handed) and the group of elements that do not. We arbitrarily label one population as Population 1 and the other as Population 2, and subscript the proportion of each population that possesses the characteristic with the number 1 or 2 to tell them apart. We draw a random sample from Population 1 and label the sample statistic it yields with the subscript 1. Without reference to the first sample we draw a sample from Population 2 and label its sample statistic with the subscript 2.

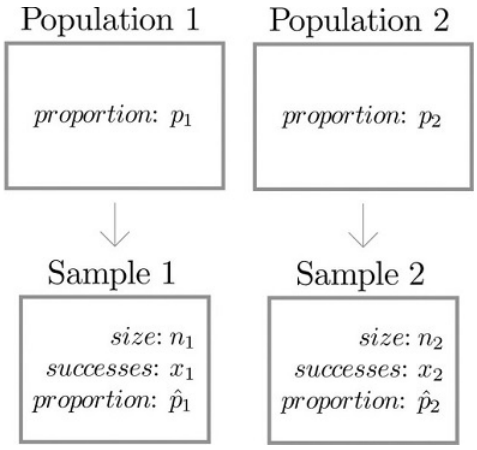


Figure 9.4.1: Independent Sampling from Two Populations In Order to Compare Proportions

Our goal is to use the information in the samples to estimate the difference $p_1 - p_2$ in the two population proportions and to make statistically valid inferences about it.

Confidence Intervals

Since the sample proportion $\hat{p}_1$ computed using the sample drawn from Population 1 is a good estimator of population proportion p_1 of Population 1 and the sample proportion $\hat{p}_2$ computed using the sample drawn from Population 2 is a good estimator of population proportion p_2 of Population 2, a reasonable point estimate of the difference $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. In order to widen this point estimate into a confidence interval we suppose that both samples are large, as described in Section 7.3 and repeated below. If so, then the following formula for a confidence interval for $p_1 - p_2$ is valid.

$100(1-\alpha)\%$ Confidence Interval for the Difference Between Two Population Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{a/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The samples must be independent, and *each* sample must be large: each of the intervals

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right]$$

and

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

$$\left[\quad \sqrt{\quad} \quad n_2 \quad , \quad \sqrt{\quad} \quad n_2 \quad \right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.4.1

The department of code enforcement of a county government issues permits to general contractors to work on residential projects. For each permit issued, the department inspects the result of the project and gives a "pass" or "fail" rating. A failed project must be re-inspected until it receives a pass rating. The department had been frustrated by the high cost of re-inspection and decided to publish the inspection records of all contractors on the web. It was hoped that public access to the records would lower the re-inspection rate. A year after the web access was made public, two samples of records were randomly selected. One sample was selected from the pool of records before the web publication and one after. The proportion of projects that passed on the first inspection was noted for each sample. The results are summarized below. Construct a point estimate and a 90% confidence interval for the difference in the passing rate on first inspection between the two time periods.

No public web access	$n_1 = 500 \;\; \hat{p_1} = 0.67$
Public web access	$n_2 = 100 \;\; \hat{p}_2 = 0.80$

Solution

The point estimate of $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 = 0.67 - 0.80 = -0.13$$

Because the "No public web access" population was labeled as Population 1 and the "Public web access" population was labeled as Population 2, in words this means that we estimate that the proportion of projects that passed on the first inspection increased by 13 percentage points after records were posted on the web.

The sample sizes are sufficiently large for constructing a confidence interval since for sample 1:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.67)(0.33)}{500}} = 0.06$$

so that

$$\left[\hat{p}_1 - 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \hat{p}_1 + 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right] = [0.67 - 0.06, 0.67 + 0.06] = [0.61, 0.73] \subset [0, 1]$$

and for sample 2:

$$3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = 3\sqrt{\frac{(0.8)(0.2)}{100}} = 0.12$$

so that

$$\left[\hat{p}_2 - 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_2 + 3\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right] = [0.8 - 0.12, 0.8 + 0.12] = [0.68, 0.92] \subset [0, 1]$$

To apply the formula for the confidence interval, we first observe that the 90% confidence level means that $\alpha = 1 - 0.90 = 0.10$ so that $z_{\alpha/2} = z_{0.05}$. From Figure 7.1.6 we read directly that $z_{0.05} = 1.645$. Thus the desired confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{9.4.1}$$

$$= 0.13 \pm 1.645\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}} \tag{9.4.2}$$

$$= -0.13 \pm 0.07 \tag{9.4.3}$$

The 90% confidence interval is $[-0.20, -0.06]$. We are 90% confident that the difference in the population proportions lies in the interval $[-0.20, -0.06]$, in the sense that in repeated sampling 90% of all intervals constructed from the sample data in this

manner will contain $p_1 - p_2$. Taking into account the labeling of the two populations, this means that we are 90% confident that the proportion of projects that pass on the first inspection is between 6 and 20 percentage points higher after public access to the records than before.

Hypothesis Testing

In hypothesis tests concerning the relative sizes of the proportions p_1 and p_2 of two populations that possess a particular characteristic, the null and alternative hypotheses will always be expressed in terms of the difference of the two population proportions. Hence the null hypothesis is always written

$$H_0 : p_1 - p_2 = D_0$$

The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : p_1 - p_2 < D_0$	Left-tailed
$H_a : p_1 - p_2 > D_0$	Right-tailed
$H_a : p_1 - p_2 \neq D_0$	Two-tailed

As long as the samples are independent and both are large the following formula for the standardized test statistic is valid, and it has the standard normal distribution.

Standardized Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Proportions

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

The test statistic has the standard normal distribution.

The samples must be independent, and each sample must be large: each of the intervals

$$\left[\hat{p_1} - 3\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}}, \hat{p_1} + 3\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}}\right]$$

and

$$\left[\hat{p_2} - 3\sqrt{\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}, \hat{p_2} + 3\sqrt{\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right]$$

must lie wholly within the interval $[0, 1]$.

Example 9.4.2

Using the data of Example $9.4.1$, test whether there is sufficient evidence to conclude that public web access to the inspection records has increased the proportion of projects that passed on the first inspection by more than 5 percentage points. Use the critical value approach at the 10% level of significance.

Solution

- **Step 1**. Taking into account the labeling of the populations an increase in passing rate at the first inspection by more than 5 percentage points after public access on the web may be expressed as $p_2 > p_1 + 0.05$, which by algebra is the same as $p_1 - p_2 < -0.05$. This is the alternative hypothesis. Since the null hypothesis is always expressed as an equality, with the same number on the right as is in the alternative hypothesis, the test is

$$H_0 : p_1 - p_2 = -0.05$$
vs.
$$H_a : p_1 - p_2 < -0.05 \quad @ \quad \alpha = 0.10$$

$H_a : p_1 - p_2 < -0.05$ @ $\alpha = 0.10$

- **Step 2**. Since the test is with respect to a difference in population proportions the test statistic is

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

- **Step 3**. Inserting the values given in Example 9.4.1 and the value $D_0 = -0.05$into the formula for the test statistic gives

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - D_0}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} = \frac{(-0.13) - (-0.05)}{\sqrt{\frac{(0.67)(0.33)}{500} + \frac{(0.8)(0.2)}{100}}} = -1.770$$

- **Step 4**. Since the symbol in H_a is "<" this is a left-tailed test, so there is a single critical value, $z_\alpha = -z_{0.10}$. From the last row in Figure 7.1.6 $z_{0.10} = 1.282$, so $-z_{0.10} = -1.282$. The rejection region is $(-\infty, -1.282]$.
- **Step 5**. As shown in Figure 9.4.2 the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 10% level of significance, to conclude that the rate of passing on the first inspection has increased by more than 5 percentage points since records were publicly posted on the web.

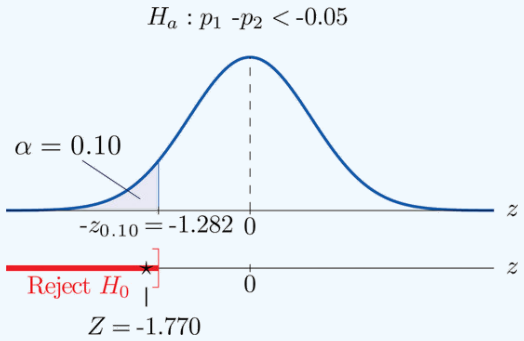


Figure 9.4.2: Rejection Region and Test Statistic for "Example 9.4.2"

Example 9.4.3

Perform the test of Example 9.4.2 using the p-value approach.

Solution

The first three steps are identical to those in Example 9.4.2

- **Step 4**. Because the test is left-tailed the observed significance or p-value of the test is just the area of the left tail of the standard normal distribution that is cut off by the test statistic $Z = -1.770$. From Figure 7.1.5 the area of the left tail determined by -1.77 is 0.0384. The p-value is 0.0384.
- **Step 5**. Since the p-value 0.0384 is less than $\alpha = 0.10$, the decision is to reject the null hypothesis: The data provide sufficient evidence, at the 10% level of significance, to conclude that the rate of passing on the first inspection has increased by more than 5 percentage points since records were publicly posted on the web.

Finally a common misuse of the formulas given in this section must be mentioned. Suppose a large pre-election survey of potential voters is conducted. Each person surveyed is asked to express a preference between, say, Candidate A and Candidate B. (Perhaps "no preference" or "other" are also choices, but that is not important.) In such a survey, estimators $\hat{p}_A$ and $\hat{p}B$ of p_A and p_B can be calculated. It is important to realize, however, that these two estimators were not calculated from two independent samples. While $\hat{p}A - \hat{p}_B$ may be a reasonable estimator of $p_A - p_B$, the formulas for confidence intervals and for the standardized test statistic given in this section are not valid for data obtained in this manner.

Key Takeaway

- A confidence interval for the difference in two population proportions is computed using a formula in the same fashion as was done for a single population mean.

- The same five-step procedure used to test hypotheses concerning a single population proportion is used to test hypotheses concerning the difference between two population proportions. The only difference is in the formula for the standardized test statistic.

This page titled 9.4: Comparison of Two Population Proportions is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.5: Sample Size Considerations

Learning Objectives

- To learn how to apply formulas for estimating the size samples that will be needed in order to construct a confidence interval for the difference in two population means or proportions that meets given criteria.

As was pointed out at the beginning of Section 7.4, sampling is typically done with definite objectives in mind. For example, a physician might wish to estimate the difference in the average amount of sleep gotten by patients suffering a certain condition with the average amount of sleep got by healthy adults, at 90% confidence and to within half an hour. Since sampling costs time, effort, and money, it would be useful to be able to estimate the smallest size samples that are likely to meet these criteria.

Estimating $\mu_1 - \mu_2$ with Independent Samples

Assuming that large samples will be required, the confidence interval formula for estimating the difference $\mu_1 - \mu_2$ between two population means using independent samples is $(\bar{x_1} - \bar{x_2}) \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. The number $z_{\alpha/2}$ is determined by the desired level of confidence.

The numbers s_1 and s_2 are estimates of the standard deviations σ_1 and σ_2 of the two populations. In analogy with what we did in Section 7.4 we will assume that we either know or can reasonably approximate σ_1 and σ_2.

We cannot solve for both n_1 and n_2, so we have to make an assumption about their relative sizes. We will specify that they be equal. With these assumptions we obtain the minimum sample sizes needed by solving the equation displayed just above for $n_1 = n_2$.

Minimum Equal Sample Sizes for Estimating the Difference in the Means of Two Populations Using Independent Samples

The estimated minimum equal sample sizes $n_1 = n_2$ needed to estimate the difference $\mu_1 - \mu_2$ in two population means to within E units at $100(1-\alpha)\%$ confidence is

$$n_1 = n_2 = \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{E^2} \text{ rounded up}$$

In all the examples and exercises the population standard deviations σ_1 and σ_2 will be given.

Example 9.5.1

A law firm wishes to estimate the difference in the mean delivery time of documents sent between two of its offices by two different courier companies, to within half an hour and with 99.5% confidence. From their records it will randomly sample the same number n of documents as delivered by each courier company. Determine how large n must be if the estimated standard deviations of the delivery times are 0.75 hour for one company and 1.15 hours for the other.

Solution

Confidence level 99.5% means that $\alpha = 1 - 0.995 = 0.005$ so $\alpha/2 = 0.0025$. From the last line of Figure 7.1.6 we obtain $z_{0.0025} = 2.807$.

To say that the estimate is to be "to within half an hour" means that $E = 0.5$. Thus

$$n = \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{E^2} = \frac{(2.807)^2(0.75^2 + 1.15^2)}{0.5^2} = 59.40953746$$

which we round up to 60, since it is impossible to take a fractional observation. The law firm must sample 60 document deliveries by each company.

Estimating $\mu_1 - \mu_2$ with Paired Samples

As we mentioned at the end of Section 9.3, if the sample is large (meaning that $n \geq 30$) then in the formula for the confidence interval we may replace $t_{\alpha/2}$ by $z_{\alpha/2}$, so that the confidence interval formula becomes $\bar{d} \pm E$ for

$$E = z_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

The number s_d is an estimate of the standard deviations σ_d of the population of differences. We must assume that we either know or can reasonably approximate σ_d. Thus, assuming that large samples will be required to meet the criteria given, we can solve the displayed equation for n to obtain an estimate of the number of pairs needed in the sample.

Minimum Sample Size for Estimating the Difference in the Means of Two Populations Using Paired Difference Samples

The estimated minimum number of pairs n needed to estimate the difference $\mu_d = \mu_1 - \mu_2$ in two population means to within E units at $100(1-\alpha)\%$ confidence using paired difference samples is

$$n = \frac{(z_{\alpha/2})^2\sigma_d^2}{E^2} \text{ rounded up}$$

In all the examples and exercises the population standard deviation of the differences σ_d will be given.

Example 9.5.2

A automotive tire manufacturer wishes to compare the mean lifetime of two tread designs under actual driving conditions. They will mount one of each type of tire on n vehicles (both on the front or both on the back) and measure the difference in remaining tread after $20,000$ miles of driving. If the standard deviation of the differences is assumed to be 0.025 inch, find the minimum samples size needed to estimate the difference in mean depth (at $20,000$ miles use) to within 0.01 inch at 99.9% confidence.

Solution

Confidence level 99.9% means that $\alpha = 1 - 0.999 = 0.001$ so $\alpha/2 = 0.0005$. From the last line of Figure 7.1.6 we obtain $z_{0.0005} = 3.291$.

To say that the estimate is to be "to within 0.01 inch" means that $E = 0.01$. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma_d^2}{E^2} = \frac{(3.291)^2(0.025)^2}{(0.01)^2} = 67.69175625$$

which we round up to 68. The manufacturer must test 68 pairs of tires.

Estimating $p_1 - p_2$

The confidence interval formula for estimating the difference $p_1 - p_2$ between two population proportions is $\hat{p_1} - \hat{p_2} \pm E$, where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. The number $z_{\alpha/2}$ is determined by the desired level of confidence.

We cannot solve for both n_1 and n_2, so we have to make an assumption about their relative sizes. We will specify that they be equal. With these assumptions we obtain the minimum sample sizes needed by solving the displayed equation for $n_1 = n_2$.

Minimum Equal Sample Sizes for Estimating the Difference in Two Population Proportions

The estimated minimum equal sample sizes $n_1 = n_2$ needed to estimate the difference $p_1 - p_2$ in two population proportions to within E percentage points at $100(1-\alpha)\%$ confidence is

$$n_1 = n_2 = \frac{(z_{\alpha/2})^2(\hat{p_1}(1-\hat{p_1} + \hat{p_2}(1-\hat{p_2}))}{E^2} \text{ rounded up}$$

E^2

Here we face the same dilemma that we encountered in the case of a single population proportion: the formula for estimating how large a sample to take contains the numbers $\hat{p}_1$ and $\hat{p}_2$, which we know only after we have taken the sample. There are two ways out of this dilemma. Typically the researcher will have some idea as to the values of the population proportions p_1 and p_2, hence of what the sample proportions $\hat{p}_1$ and $\hat{p}_2$ are likely to be. If so, those estimates can be used in the formula.

The second approach to resolving the dilemma is simply to replace each of $\hat{p}_1$ and $\hat{p}_2$ in the formula by 0.5. As in the one-population case, this is the most conservative estimate, since it gives the largest possible estimate of n. If we have an estimate of only one of p_1 and p_2 we can use that estimate for it, and use the conservative estimate 0.5 for the other.

Example 9.5.3

Find the minimum equal sample sizes necessary to construct a 98% confidence interval for the difference $p_1 - p_2$ with a margin of error $E = 0.05$,

1. assuming that no prior knowledge about p_1 or p_2 is available; and
2. assuming that prior studies suggest that $p_1 \approx 0.2$ and $p_2 \approx 0.3$.

Solution

Confidence level 98% means that $\alpha = 1 - 0.98 = 0.02$ so $\alpha/2 = 0.01$. From the last line of Figure 7.1.6 we obtain $z_{0.01} = 2.326$.

1. Since there is no prior knowledge of p_1 or p_2 we make the most conservative estimate that $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$. Then

$$\begin{aligned} n_1 = n_2 &= \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1 + \hat{p}_2(1-\hat{p}_2))}{E^2} \\ &= \frac{(2.326)^2((0.5)(0.5) + (0.5)(0.5))}{0.05^2} \\ &= 1082.0552 \end{aligned}$$

which we round up to $1,083$. We must take a sample of size $1,083$ from each population.

2. Since $p_1 \approx 0.2$ we estimate $\hat{p}_1$ by 0.2, and since $p_2 \approx 0.3$ we estimate $\hat{p}_2$ by 0.3. Thus we obtain

$$\begin{aligned} n_1 = n_2 &= \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1 + \hat{p}_2(1-\hat{p}_2))}{E^2} \\ &= \frac{(2.326)^2((0.2)(0.8) + (0.3)(0.7))}{0.05^2} \\ &= 800.720848 \end{aligned}$$

which we round up to 801. We must take a sample of size 801 from each population.

Key Takeaway

- If the population standard deviations σ_1 and σ_2 are known or can be estimated, then the minimum equal sizes of independent samples needed to obtain a confidence interval for the difference $\mu_1 - \mu_2$ in two population means with a given maximum error of the estimate E and a given level of confidence can be estimated.
- If the standard deviation σ_d of the population of differences in pairs drawn from two populations is known or can be estimated, then the minimum number of sample pairs needed under paired difference sampling to obtain a confidence interval for the difference $\mu_d = \mu_1 - \mu_2$ in two population means with a given maximum error of the estimate E and a given level of confidence can be estimated.
- The minimum equal sample sizes needed to obtain a confidence interval for the difference in two population proportions with a given maximum error of the estimate and a given level of confidence can always be estimated. If there is prior knowledge of the population proportions p_1 and p_2 then the estimate can be sharpened.

This page titled 9.5: Sample Size Considerations is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

9.E: Two-Sample Problems (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

9.1: Comparison of Two Population Means: Large, Independent Samples

Basic

Q9.1.1

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 90% confidence,

$$\begin{aligned} n_1 = 45, \bar{x_1} = 27, s_1 = 2 \\ n_2 = 60, \bar{x_2} = 22, s_2 = 3 \end{aligned} \tag{9.E.1}$$

b. 99% confidence,

$$\begin{aligned} n_1 = 30, \bar{x_1} = -112, s_1 = 9 \\ n_2 = 40, \bar{x_2} = -98, s_2 = 4 \end{aligned} \tag{9.E.2}$$

Q9.1.2

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 95% confidence,

$$\begin{aligned} n_1 = 110, \bar{x_1} = 77, s_1 = 15 \\ n_2 = 85, \bar{x_2} = 79, s_2 = 21 \end{aligned} \tag{9.E.3}$$

b. 90% confidence,

$$\begin{aligned} n_1 = 65, \bar{x_1} = -83, s_1 = 12 \\ n_2 = 65, \bar{x_2} = -74, s_2 = 8 \end{aligned} \tag{9.E.4}$$

Q9.1.3

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.5% confidence,

$$\begin{aligned} n_1 = 130, \bar{x_1} = 27.2, s_1 = 2.5 \\ n_2 = 155, \bar{x_2} = 38.8, s_2 = 4.6 \end{aligned} \tag{9.E.5}$$

b. 95% confidence,

$$\begin{aligned} n_1 = 68, \bar{x_1} = 215.5, s_1 = 12.3 \\ n_2 = 84, \bar{x_2} = 287.8, s_2 = 14.1 \end{aligned} \tag{9.E.6}$$

Q9.1.4

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.9% confidence,

$$\begin{aligned} n_1 = 275, \bar{x_1} = 70.2, s_1 = 1.5 \\ n_2 = 325, \bar{x_2} = 63.4, s_2 = 1.1 \end{aligned} \tag{9.E.7}$$

b. 90% confidence,

$$\begin{aligned} n_1 = 120, \bar{x_1} = 35.5, s_1 = 0.75 \\ n_2 = 146, \bar{x_2} = 29.6, s_2 = 0.80 \end{aligned} \tag{9.E.8}$$

Q9.1.5

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 3 \; vs \; H_a : \mu_1 - \mu_2 \neq 3 \; @ \; \alpha = 0.05$

$$n_1 = 35, \bar{x_1} = 25, s_1 = 1 \\ n_2 = 45, \bar{x_2} = 19, s_2 = 2 \tag{9.E.9}$$

b. Test $H_0 : \mu_1 - \mu_2 = -25$ *vs* $H_a : \mu_1 - \mu_2 < -25$ @ $\alpha = 0.10$

$$n_1 = 85, \bar{x_1} = 188, s_1 = 15 \\ n_2 = 62, \bar{x_2} = 215, s_2 = 19 \tag{9.E.10}$$

Q9.1.6

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 45$ *vs* $H_a : \mu_1 - \mu_2 > 45$ @ $\alpha = 0.001$

$$n_1 = 200, \bar{x_1} = 1312, s_1 = 35 \\ n_2 = 225, \bar{x_2} = 1256, s_2 = 28 \tag{9.E.11}$$

b. Test $H_0 : \mu_1 - \mu_2 = -12$ *vs* $H_a : \mu_1 - \mu_2 \neq -12$ @ $\alpha = 0.10$

$$n_1 = 35, \bar{x_1} = 121, s_1 = 6 \\ n_2 = 40, \bar{x_2} = 135 s_2 = 7 \tag{9.E.12}$$

Q9.1.7

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 0$ *vs* $H_a : \mu_1 - \mu_2 \neq 0$ @ $\alpha = 0.01$

$$n_1 = 125, \bar{x_1} = -46, s_1 = 10 \\ n_2 = 90, \bar{x_2} = -50, s_2 = 13 \tag{9.E.13}$$

b. Test $H_0 : \mu_1 - \mu_2 = 20$ *vs* $H_a : \mu_1 - \mu_2 > 20$ @ $\alpha = 0.05$

$$n_1 = 40, \bar{x_1} = 142, s_1 = 11 \\ n_2 = 40, \bar{x_2} = 118 s_2 = 10 \tag{9.E.14}$$

Q9.1.8

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach. Compute the p-value of the test as well.

a. Test $H_0 : \mu_1 - \mu_2 = 13$ *vs* $H_a : \mu_1 - \mu_2 < 13$ @ $\alpha = 0.01$

$$n_1 = 35, \bar{x_1} = 100, s_1 = 2 \\ n_2 = 35, \bar{x_2} = 88, s_2 = 2 \tag{9.E.15}$$

b. Test $H_0 : \mu_1 - \mu_2 = -10$ *vs* $H_a : \mu_1 - \mu_2 \neq -10$ @ $\alpha = 0.10$

$$n_1 = 146, \bar{x_1} = 62, s_1 = 4 \\ n_2 = 120, \bar{x_2} = 73 s_2 = 7 \tag{9.E.16}$$

Q9.1.9

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 57$ *vs* $H_a : \mu_1 - \mu_2 < 57$ @ $\alpha = 0.10$

$$n_1 = 117, \bar{x_1} = 1309, s_1 = 42 \\ n_2 = 133, \bar{x_2} = 1258, s_2 = 37 \tag{9.E.17}$$

b. Test $H_0 : \mu_1 - \mu_2 = -1.5$ *vs* $H_a : \mu_1 - \mu_2 \neq -1.5$ @ $\alpha = 0.20$

$$n_1 = 65, \bar{x_1} = 16.9, s_1 = 1.3 \\ n_2 = 57, \bar{x_2} = 18.6 s_2 = 1.1 \tag{9.E.18}$$

Q9.1.10

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = -10.5$ *vs* $H_a : \mu_1 - \mu_2 > -10.5$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 64, \bar{x_1} = 85.6, s_1 = 2.4 \\ n_2 = 50, \bar{x_2} = 95.3, s_2 = 3.1 \end{aligned} \tag{9.E.19}$$

b. Test $H_0 : \mu_1 - \mu_2 = 110$ *vs* $H_a : \mu_1 - \mu_2 \neq 110$ @ $\alpha = 0.02$

$$\begin{aligned} n_1 = 176, \bar{x_1} = 1918, s_1 = 68 \\ n_2 = 241, \bar{x_2} = 1782 s_2 = 146 \end{aligned} \tag{9.E.20}$$

Q9.1.11

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 50$ *vs* $H_a : \mu_1 - \mu_2 > 50$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 = 72, \bar{x_1} = 272, s_1 = 26 \\ n_2 = 103, \bar{x_2} = 213, s_2 = 14 \end{aligned} \tag{9.E.21}$$

b. Test $H_0 : \mu_1 - \mu_2 = 7.5$ *vs* $H_a : \mu_1 - \mu_2 \neq 7.5$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 52, \bar{x_1} = 94.3, s_1 = 2.6 \\ n_2 = 38, \bar{x_2} = 88.6 s_2 = 8.0 \end{aligned} \tag{9.E.22}$$

Q9.1.12

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 23$ *vs* $H_a : \mu_1 - \mu_2 < 23$ @ $\alpha = 0.20$

$$\begin{aligned} n_1 = 314, \bar{x_1} = 198, s_1 = 12.2 \\ n_2 = 220, \bar{x_2} = 176, s_2 = 11.5 \end{aligned} \tag{9.E.23}$$

b. Test $H_0 : \mu_1 - \mu_2 = 4.4$ *vs* $H_a : \mu_1 - \mu_2 \neq 4.4$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 32, \bar{x_1} = 40.3, s_1 = 0.5 \\ n_2 = 30, \bar{x_2} = 35.5 s_2 = 0.7 \end{aligned} \tag{9.E.24}$$

Applications

Q9.1.13

In order to investigate the relationship between mean job tenure in years among workers who have a bachelor's degree or higher and those who do not, random samples of each type of worker were taken, with the following results.

	n	$\bar{x}$	s
Bachelor's degree or higher	155	5.2	1.3
No degree	210	5.0	1.5

a. Construct the 99% confidence interval for the difference in the population means based on these data.
b. Test, at the 1% level of significance, the claim that mean job tenure among those with higher education is greater than among those without, against the default that there is no difference in the means.
c. Compute the observed significance of the test.

Q9.1.14

Records of 40 used passenger cars and 40 used pickup trucks (none used commercially) were randomly selected to investigate whether there was any difference in the mean time in years that they were kept by the original owner before being sold. For cars the mean was 5.3 years with standard deviation 2.2 years. For pickup trucks the mean was 7.1 years with standard deviation 3.0 years.

a. Construct the 95% confidence interval for the difference in the means based on these data.

b. Test the hypothesis that there is a difference in the means against the null hypothesis that there is no difference. Use the 1% level of significance.
c. Compute the observed significance of the test in part (b).

Q9.1.15

In previous years the average number of patients per hour at a hospital emergency room on weekends exceeded the average on weekdays by 6.3 visits per hour. A hospital administrator believes that the current weekend mean exceeds the weekday mean by fewer than 6.3 hours.

a. Construct the 99% confidence interval for the difference in the population means based on the following data, derived from a study in which 30 weekend and 30 weekday one-hour periods were randomly selected and the number of new patients in each recorded.

	n	$\bar{x}$	s
Weekends	30	13.8	3.1
Weekdays	30	8.6	2.7

a. Test at the 5% level of significance whether the current weekend mean exceeds the weekday mean by fewer than 6.3 patients per hour.
b. Compute the observed significance of the test.

Q9.1.16

A sociologist surveys 50 randomly selected citizens in each of two countries to compare the mean number of hours of volunteer work done by adults in each. Among the 50 inhabitants of Lilliput, the mean hours of volunteer work per year was 52, with standard deviation 11.8. Among the 50 inhabitants of Blefuscu, the mean number of hours of volunteer work per year was 37, with standard deviation 7.2.

a. Construct the 99% confidence interval for the difference in mean number of hours volunteered by all residents of Lilliput and the mean number of hours volunteered by all residents of Blefuscu.
b. Test, at the 1% level of significance, the claim that the mean number of hours volunteered by all residents of Lilliput is more than ten hours greater than the mean number of hours volunteered by all residents of Blefuscu.
c. Compute the observed significance of the test in part (b).

Q9.1.17

A university administrator asserted that upperclassmen spend more time studying than underclassmen.

a. Test this claim against the default that the average number of hours of study per week by the two groups is the same, using the following information based on random samples from each group of students. Test at the 1% level of significance.

	n	$\bar{x}$	s
Upperclassmen	35	15.6	2.9
Underclassmen	35	12.3	4.1

b. Compute the observed significance of the test.

Q9.1.18

An kinesiologist claims that the resting heart rate of men aged 18 to 25 who exercise regularly is more than five beats per minute less than that of men who do not exercise regularly. Men in each category were selected at random and their resting heart rates were measured, with the results shown.

	n	$\bar{x}$	s
Regular exercise	40	63	1.0

	n	$\bar{x}$	s
No regular exercise	30	71	1.2

a. Perform the relevant test of hypotheses at the 1% level of significance.
b. Compute the observed significance of the test.

Q9.1.19

Children in two elementary school classrooms were given two versions of the same test, but with the order of questions arranged from easier to more difficult in Version A and in reverse order in Version B. Randomly selected students from each class were given Version A and the rest Version B. The results are shown in the table.

	n	$\bar{x}$	s
Version A	31	83	4.6
Version B	32	78	4.3

a. Construct the 90% confidence interval for the difference in the means of the populations of all children taking Version A of such a test and of all children taking Version B of such a test.
b. Test at the 1% level of significance the hypothesis that the A version of the test is easier than the B version (even though the questions are the same).
c. Compute the observed significance of the test.

Q9.1.20

The Municipal Transit Authority wants to know if, on weekdays, more passengers ride the northbound blue line train towards the city center that departs at $8:15\ a.m.$ or the one that departs at $8:30\ a.m.$. The following sample statistics are assembled by the Transit Authority.

	n	$\bar{x}$	s
8:15 a.m. train	30	323	41
8:30 a.m. train	45	356	45

a. Construct the 90% confidence interval for the difference in the mean number of daily travelers on the $8:15\ a.m.$ train and the mean number of daily travelers on the $8:30\ a.m.$ train.
b. Test at the 5% level of significance whether the data provide sufficient evidence to conclude that more passengers ride the $8:30\ a.m.$ train.
c. Compute the observed significance of the test.

Q9.1.21

In comparing the academic performance of college students who are affiliated with fraternities and those male students who are unaffiliated, a random sample of students was drawn from each of the two populations on a university campus. Summary statistics on the student GPAs are given below.

	n	$\bar{x}$	s
Fraternity	645	2.90	0.47
Unaffiliated	450	2.88	0.42

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that there is a difference in average GPA between the population of fraternity students and the population of unaffiliated male students on this university campus.

Q9.1.22

In comparing the academic performance of college students who are affiliated with sororities and those female students who are unaffiliated, a random sample of students was drawn from each of the two populations on a university campus. Summary statistics on the student GPAs are given below.

	n	$\bar{x}$	s
Sorority	330	3.18	0.37
Unaffiliated	550	3.12	0.41

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that there is a difference in average GPA between the population of sorority students and the population of unaffiliated female students on this university campus.

Q9.1.23

The owner of a professional football team believes that the league has become more offense oriented since five years ago. To check his belief, 32 randomly selected games from one year's schedule were compared to 32 randomly selected games from the schedule five years later. Since more offense produces more points per game, the owner analyzed the following information on points per game (ppg).

	n	$\bar{x}$	s
ppg previously	32	20.62	4.17
ppg recently	32	22.05	4.01

Test, at the 10% level of significance, whether the data on points per game provide sufficient evidence to conclude that the game has become more offense oriented.

Q9.1.24

The owner of a professional football team believes that the league has become more offense oriented since five years ago. To check his belief, 32 randomly selected games from one year's schedule were compared to 32 randomly selected games from the schedule five years later. Since more offense produces more offensive yards per game, the owner analyzed the following information on offensive yards per game (oypg).

	n	$\bar{x}$	s
oypg previously	32	316	40
oypg recently	32	336	35

Test, at the 10% level of significance, whether the data on offensive yards per game provide sufficient evidence to conclude that the game has become more offense oriented.

Large Data Set Exercises

Large Data Sets are absent

25. Large Data Sets 1A and 1B list the SAT scores for $1,000$ randomly selected students. Denote the population of all male students as Population 1 and the population of all female students as Population 2.
    a. Restricting attention to just the males, find n_1, $\bar{x_1}$ and s_1. Restricting attention to just the females, find n_2, $\bar{x_2}$ and s_2.
    b. Let μ_1 denote the mean SAT score for all males and μ_2 the mean SAT score for all females. Use the results of part (a) to construct a 90% confidence interval for the difference $\mu_1 - \mu_2$.
    c. Test, at the 5% level of significance, the hypothesis that the mean SAT scores among males exceeds that of females.
26. Large Data Sets 1A and 1B list the SAT scores for $1,000$ randomly selected students. Denote the population of all male students as Population 1 and the population of all female students as Population 2.
    a. Restricting attention to just the males, find n_1, $\bar{x_1}$ and s_1. Restricting attention to just the females, find n_2, $\bar{x_2}$ and s_2.

b. Let μ_1 denote the mean SAT score for all males and μ_2 the mean SAT score for all females. Use the results of part (a) to construct a 95% confidence interval for the difference $\mu_1 - \mu_2$.
c. Test, at the 10% level of significance, the hypothesis that the mean SAT scores among males exceeds that of females.

27. Large Data Sets 7A and 7B list the survival times for 65 male and 75 female laboratory mice with thymic leukemia. Denote the population of all such male mice as Population 1 and the population of all such female mice as Population 2.
a. Restricting attention to just the males, find n_1, $\bar{x_1}$ and s_1. Restricting attention to just the females, find n_2, $\bar{x_2}$ and s_2.
b. Let μ_1 denote the mean survival for all males and μ_2 the mean survival time for all females. Use the results of part (a) to construct a 99% confidence interval for the difference $\mu_1 - \mu_2$.
c. Test, at the 1% level of significance, the hypothesis that the mean survival time for males exceeds that for females by more than 182 days (half a year).
d. Compute the observed significance of the test in part (c).

Answers

1. a. $(4.20, 5.80)$
b. $(-18.54, -9.46)$
2.
3. a. $(-12.81, -10.39)$
b. $(-76.50, -68.10)$
4.
5. a. $Z = 8.753, \pm z_{0.025} = \pm 1.960$, reject H_0, p-value=0.0000
b. $Z = -0.687, -z_{0.10} = -1.282$, do not reject H_0, p-value=0.2451
6.
7. a. $Z = 2.444, \pm z_{0.005} = \pm 2.576$, do not reject H_0, p-value=0.0146
b. $Z = 1.702, z_{0.05} = -1.645$, reject H_0, p-value=0.0446
8.
9. a. $Z = -1.19$, p-value=0.1170, do not reject H_0
b. $Z = -0.92$, p-value=0.3576, do not reject H_0
10.
11. a. $Z = 2.68$, p-value=0.0037, reject H_0
b. $Z = -1.34$, p-value=0.1802, do not reject H_0
12.
13. a. 0.2 ± 0.4
b. $Z = -1.466, -z_{0.050} = -1.645$, do not reject H_0 (exceeds by 6.3 or more)
c. p-value=0.0869
14.
15. a. 5.2 ± 1.9
b. $Z = -1.466, -z_{0.050} = -1.645$, do not reject H_0 (exceeds by 6.3 or more)
c. p-value=0.0708
16.
17. a. $Z = 3.888, z_{0.01} = 2.326$, reject H_0 (upperclassmen study more)
b. p-value=0.0001
18.
19. a. 5 ± 1.8
b. $Z = 4.454, z_{0.01} = 2.326$, reject H_0 (Test A is easier)
c. p-value=0.0000
20.
21. $Z = 0.738, \pm z_{0.025} = \pm 1.960$, do not reject H_0 (no difference)
22.
23. $Z = -1.398, -z_{0.10} = -1.282$, reject H_0 (more offense oriented)

24.

25. a. $n_1 = 419,\ \bar{x_1} = 1540.33,\ s_1 = 205.40,\ n_2 = 581,\ \bar{x_2} = 1520.38,\ s_2 = 217.34$
    b. $(-2.24, 42.15)$
    c. $H_0 : \mu_1 - \mu_2 = 0\ vs\ H_a : \mu_1 - \mu_2 > 0$. Test Statistic: $Z = 1.48$. Rejection Region: $[1.645, \infty)$. Decision: Fail to reject H_0.

26.

27. a. $n_1 = 65,\ \bar{x_1} = 665.97,\ s_1 = 41.60,\ n_2 = 75,\ \bar{x_2} = 455.89,\ s_2 = 63.22$
    b. $(187.06, 233.09)$
    c. $H_0 : \mu_1 - \mu_2 = 182\ vs\ H_a : \mu_1 - \mu_2 > 182$. Test Statistic: $Z = 3.14$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0.
    d. p-value=0.0008

9.2: Comparison of Two Population Means: Small, Independent Samples

Basic

In all exercises for this section assume that the populations are normal and have equal standard deviations.

Q9.2.1

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 95% confidence,

$$\begin{aligned} n_1 = 10,\ \bar{x_1} = 120,\ s_1 = 2 \\ n_2 = 15,\ \bar{x_2} = 101,\ s_1 = 4 \end{aligned} \tag{9.E.25}$$

b. 99% confidence,

$$\begin{aligned} n_1 = 6,\ \bar{x_1} = 25,\ s_1 = 1 \\ n_2 = 12,\ \bar{x_2} = 17,\ s_1 = 3 \end{aligned} \tag{9.E.26}$$

Q9.2.2

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 90% confidence,

$$\begin{aligned} n_1 = 28,\ \bar{x_1} = 212,\ s_1 = 6 \\ n_2 = 23,\ \bar{x_2} = 198,\ s_1 = 5 \end{aligned} \tag{9.E.27}$$

b. 99% confidence,

$$\begin{aligned} n_1 = 14,\ \bar{x_1} = 68,\ s_1 = 8 \\ n_2 = 20,\ \bar{x_2} = 43,\ s_1 = 3 \end{aligned} \tag{9.E.28}$$

Q9.2.3

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.9% confidence,

$$\begin{aligned} n_1 = 35,\ \bar{x_1} = 6.5,\ s_1 = 0.2 \\ n_2 = 20,\ \bar{x_2} = 6.2,\ s_1 = 0.1 \end{aligned} \tag{9.E.29}$$

b. 99% confidence,

$$\begin{aligned} n_1 = 18,\ \bar{x_1} = 77.3,\ s_1 = 1.2 \\ n_2 = 32,\ \bar{x_2} = 75.0,\ s_1 = 1.6 \end{aligned} \tag{9.E.30}$$

Q9.2.4

Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given.

a. 99.5% confidence,

$$\begin{aligned} n_1 = 40,\ \bar{x_1} = 85.6,\ s_1 = 2.8 \\ n_2 = 20,\ \bar{x_2} = 73.1,\ s_1 = 2.1 \end{aligned} \tag{9.E.31}$$

b. 99.9% confidence,

$$\begin{aligned} n_1 = 25,\ \bar{x_1} = 215,\ s_1 = 7 \\ n_2 = 35,\ \bar{x_2} = 185,\ s_1 = 12 \end{aligned} \tag{9.E.32}$$

Q9.2.5

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 11$ *vs* $H_a : \mu_1 - \mu_2 > 11$ @ $\alpha = 0.025$

$$\begin{aligned} n_1 = 6,\ \bar{x_1} = 32,\ s_1 = 2 \\ n_2 = 11,\ \bar{x_2} = 19,\ s_1 = 1 \end{aligned} \tag{9.E.33}$$

b. Test $H_0 : \mu_1 - \mu_2 = 26$ *vs* $H_a : \mu_1 - \mu_2 \neq 26$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 17,\ \bar{x_1} = 166,\ s_1 = 4 \\ n_2 = 24,\ \bar{x_2} = 138,\ s_1 = 3 \end{aligned} \tag{9.E.34}$$

Q9.2.6

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 40$ *vs* $H_a : \mu_1 - \mu_2 < 40$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 14,\ \bar{x_1} = 289,\ s_1 = 11 \\ n_2 = 12,\ \bar{x_2} = 254,\ s_1 = 9 \end{aligned} \tag{9.E.35}$$

b. Test $H_0 : \mu_1 - \mu_2 = 21$ *vs* $H_a : \mu_1 - \mu_2 \neq 21$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 23,\ \bar{x_1} = 130,\ s_1 = 6 \\ n_2 = 27,\ \bar{x_2} = 113,\ s_1 = 8 \end{aligned} \tag{9.E.36}$$

Q9.2.7

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = -15$ *vs* $H_a : \mu_1 - \mu_2 < -15$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 30,\ \bar{x_1} = 42,\ s_1 = 7 \\ n_2 = 12,\ \bar{x_2} = 60,\ s_1 = 5 \end{aligned} \tag{9.E.37}$$

b. Test $H_0 : \mu_1 - \mu_2 = 103$ *vs* $H_a : \mu_1 - \mu_2 \neq 103$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 17,\ \bar{x_1} = 711,\ s_1 = 28 \\ n_2 = 32,\ \bar{x_2} = 598,\ s_1 = 21 \end{aligned} \tag{9.E.38}$$

Q9.2.8

Perform the test of hypotheses indicated, using the data from independent samples given. Use the critical value approach.

a. Test $H_0 : \mu_1 - \mu_2 = 75$ *vs* $H_a : \mu_1 - \mu_2 > 75$ @ $\alpha = 0.025$

$$\begin{aligned} n_1 = 45,\ \bar{x_1} = 674,\ s_1 = 18 \\ n_2 = 29,\ \bar{x_2} = 591,\ s_1 = 13 \end{aligned} \tag{9.E.39}$$

b. Test $H_0 : \mu_1 - \mu_2 = -20$ *vs* $H_a : \mu_1 - \mu_2 \neq -20$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 = 30,\ \bar{x_1} = 137,\ s_1 = 8 \\ n_2 = 19,\ \bar{x_2} = 166,\ s_1 = 11 \end{aligned} \tag{9.E.40}$$

Q9.2.9

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 12$ *vs* $H_a : \mu_1 - \mu_2 > 12$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 20,\ \bar{x_1} = 133,\ s_1 = 7 \\ n_2 = 10,\ \bar{x_2} = 115,\ s_1 = 5 \end{aligned} \tag{9.E.41}$$

b. Test $H_0 : \mu_1 - \mu_2 = 46$ *vs* $H_a : \mu_1 - \mu_2 \neq 46$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 24,\ \bar{x_1} = 586,\ s_1 = 11 \\ n_2 = 27,\ \bar{x_2} = 535,\ s_1 = 13 \end{aligned} \tag{9.E.42}$$

Q9.2.10

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 38$ *vs* $H_a : \mu_1 - \mu_2 < 38$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 12,\ \bar{x_1} = 464,\ s_1 = 5 \\ n_2 = 10,\ \bar{x_2} = 432,\ s_1 = 6 \end{aligned} \tag{9.E.43}$$

b. Test $H_0 : \mu_1 - \mu_2 = 4$ *vs* $H_a : \mu_1 - \mu_2 \neq 4$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 = 14,\ \bar{x_1} = 68,\ s_1 = 2 \\ n_2 = 17,\ \bar{x_2} = 67,\ s_1 = 3 \end{aligned} \tag{9.E.44}$$

Q9.2.11

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = 50$ *vs* $H_a : \mu_1 - \mu_2 > 50$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 30,\ \bar{x_1} = 681,\ s_1 = 8 \\ n_2 = 27,\ \bar{x_2} = 625,\ s_1 = 8 \end{aligned} \tag{9.E.45}$$

b. Test $H_0 : \mu_1 - \mu_2 = 35$ *vs* $H_a : \mu_1 - \mu_2 \neq 35$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 = 36,\ \bar{x_1} = 325,\ s_1 = 11 \\ n_2 = 29,\ \bar{x_2} = 286,\ s_1 = 7 \end{aligned} \tag{9.E.46}$$

Q9.2.12

Perform the test of hypotheses indicated, using the data from independent samples given. Use the p-value approach. (The p-value can be only approximated.)

a. Test $H_0 : \mu_1 - \mu_2 = -4$ *vs* $H_a : \mu_1 - \mu_2 < -4$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 = 40,\ \bar{x_1} = 80,\ s_1 = 5 \\ n_2 = 25,\ \bar{x_2} = 87,\ s_1 = 5 \end{aligned} \tag{9.E.47}$$

b. Test $H_0 : \mu_1 - \mu_2 = 21$ *vs* $H_a : \mu_1 - \mu_2 \neq 21$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 = 15,\ \bar{x_1} = 192,\ s_1 = 12 \\ n_2 = 34,\ \bar{x_2} = 180,\ s_1 = 8 \end{aligned} \tag{9.E.48}$$

Applications

Q9.2.13

A county environmental agency suspects that the fish in a particular polluted lake have elevated mercury level. To confirm that suspicion, five striped bass in that lake were caught and their tissues were tested for mercury. For the purpose of comparison, four striped bass in an unpolluted lake were also caught and tested. The fish tissue mercury levels in mg/kg are given below.

Sample 1 (from polluted lake)	Sample 2(from unpolluted lake)
0.580	0.382
0.711	0.276
0.571	0.570
0.666	0.366
0.598	

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that fish in the polluted lake have elevated levels of mercury in their tissue.

Q9.2.14

A genetic engineering company claims that it has developed a genetically modified tomato plant that yields on average more tomatoes than other varieties. A farmer wants to test the claim on a small scale before committing to a full-scale planting. Ten genetically modified tomato plants are grown from seeds along with ten other tomato plants. At the season's end, the resulting yields in pound are recorded as below.

Sample 1(genetically modified)	Sample 2(regular)
20	21
23	21
27	22
25	18
25	20
25	20
27	18
23	25
24	23
22	20

a. Construct the 99% confidence interval for the difference in the population means based on these data.
b. Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that the mean yield of the genetically modified variety is greater than that for the standard variety.

Q9.2.15

The coaching staff of a professional football team believes that the rushing offense has become increasingly potent in recent years. To investigate this belief, 20 randomly selected games from one year's schedule were compared to 11 randomly selected games from the schedule five years later. The sample information on rushing yards per game (rypg) is summarized below.

	n	$\bar{x}$	s
rypg previously	20	112	24
rypg recently	11	114	21

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data on rushing yards per game provide sufficient evidence to conclude that the rushing offense has become more potent in recent years.

Q9.2.16

The coaching staff of professional football team believes that the rushing offense has become increasingly potent in recent years. To investigate this belief, 20 randomly selected games from one year's schedule were compared to 11 randomly selected games from the schedule five years later. The sample information on passing yards per game (pypg) is summarized below.

	n	$\bar{x}$	s
pypg previously	20	203	38

	n	$\bar{x}$	s
pypg recently	11	232	33

a. Construct the 95% confidence interval for the difference in the population means based on these data.
b. Test, at the 5% level of significance, whether the data on passing yards per game provide sufficient evidence to conclude that the passing offense has become more potent in recent years.

Q9.2.17

A university administrator wishes to know if there is a difference in average starting salary for graduates with master's degrees in engineering and those with master's degrees in business. Fifteen recent graduates with master's degree in engineering and 11 with master's degrees in business are surveyed and the results are summarized below.

	n	$\bar{x}$	s
Engineering	15	68,535	1627
Business	11	63,230	2033

a. Construct the 90% confidence interval for the difference in the population means based on these data.
b. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the average starting salaries are different.

Q9.2.18

A gardener sets up a flower stand in a busy business district and sells bouquets of assorted fresh flowers on weekdays. To find a more profitable pricing, she sells bouquets for 15 dollars each for ten days, then for 10 dollars each for five days. Her average daily profit for the two different prices are given below.

	n	$\bar{x}$	s
$15	10	171	26
$10	5	198	29

a. Construct the 90% confidence interval for the difference in the population means based on these data.
b. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude the gardener's average daily profit will be higher if the bouquets are sold at $\$10$ each.

Answers

1. a. $(16.16, 21.84)$
   b. $(4.28, 11.72)$
2.
3. a. $(0.13, 0.47)$
   b. \((1.14,3.46)\)
4.
5. a. $T = 2.787,\ t_{0.025} = 2.131$, reject H_0
   b. $T = 1.831,\ \pm t_{0.025} = \pm 2.023$, do not reject H_0
6.
7. a. $T = -1.349,\ -t_{0.10} = -1.303$, reject H_0
   b. $T = 1.411,\ \pm t_{0.05} = \pm 1.678$, do not reject H_0
8.
9. a. $T = 2.411,\ df = 28,\ \text{p-value} > 0.01$, do not reject H_0
   b. $T = 1.473,\ df = 49,\ \text{p-value} < 0.10$, reject H_0
10.

11. a. $T = 2.827,\ df = 55,\ \text{p-value} < 0.01$, reject H_0
    b. $T = 1.699,\ df = 63,\ \text{p-value} < 0.10$, reject H_0
12.
13. a. 0.2267 ± 0.2182
    b. $T = 1.699,\ df = 63,\ t_{0.05} = 1.895$, reject H_0 (elevated levels)
14.
15. a. -2 ± 17.7
    b. $T = -0.232,\ df = 29,\ -t_{0.05} = -1.699$, do not reject H_0 (not more potent)
16.
17. a. 5305 ± 1227
    b. $T = 7.395,\ df = 24,\ \pm t_{0.05} = \pm 1.711$, reject H_0 (different)

9.3 Comparison of Two Population Means: Paired Samples

Basic

In all exercises for this section assume that the population of differences is normal.

1. Use the following paired sample data for this exercise.

$$\begin{array}{lcccccccc} \textit{Population } 1 & 35 & 32 & 35 & 35 & 36 & 35 & 35 \\ \textit{Population } 2 & 28 & 26 & 27 & 26 & 29 & 27 & 29 \end{array} \tag{9.E.49}$$

   a. Compute $\bar{d}$ and s_d.
   b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
   c. Construct the 95% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
   d. Test, at the 10% level of significance, the hypothesis that $\mu_1 - \mu_2 > 7$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 7$.

2. Use the following paired sample data for this exercise.

$$\begin{array}{lcccc} \textit{Population } 1 & 103 & 127 & 96 & 110 \\ \textit{Population } 2 & 81 & 106 & 73 & 88 \\ \textit{Population } 1 & 90 & 118 & 130 & 106 \\ \textit{Population } 2 & 70 & 95 & 109 & 83 \end{array} \tag{9.E.50}$$

   a. Compute $\bar{d}$ and s_d.
   b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
   c. Construct the 90% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
   d. Test, at the 1% level of significance, the hypothesis that $\mu_1 - \mu_2 < 247$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 24$.

3. Use the following paired sample data for this exercise.

$$\begin{array}{lcccc} \textit{Population } 1 & 40 & 27 & 55 & 34 \\ \textit{Population } 2 & 53 & 42 & 68 & 50 \end{array} \tag{9.E.51}$$

   a. Compute $\bar{d}$ and s_d.
   b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
   c. Construct the 99% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
   d. Test, at the 10% level of significance, the hypothesis that $\mu_1 - \mu_2 \neq -12$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = -12$.

4. Use the following paired sample data for this exercise.

$$\begin{array}{lccccc} \textit{Population } 1 & 196 & 165 & 181 & 201 & 190 \\ \textit{Population } 2 & 212 & 182 & 199 & 210 & 205 \end{array} \tag{9.E.52}$$

   a. Compute $\bar{d}$ and s_d.
   b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
   c. Construct the 98% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.

d. Test, at the 2% level of significance, the hypothesis that $\mu_1 - \mu_2 \neq -20$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = -20$.

Applications

5. Each of five laboratory mice was released into a maze twice. The five pairs of times to escape were:

Mouse	1	2	3	4	5
First release	129	89	136	163	118
Second release	113	97	139	85	75

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 90% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 10% level of significance, the hypothesis that it takes mice less time to run the maze on the second trial, on average.

6. Eight golfers were asked to submit their latest scores on their favorite golf courses. These golfers were each given a set of newly designed clubs. After playing with the new clubs for a few months, the golfers were again asked to submit their latest scores on the same golf courses. The results are summarized below.

Golfer	1	2	3	4	5	6	7	8
Own clubs	77	80	69	73	73	72	75	77
New clubs	72	81	68	73	75	70	73	75

a. Compute $\bar{d}$ and s_d.
b. Give a point estimate for $\mu_1 - \mu_2 = \mu_d$.
c. Construct the 99% confidence interval for $\mu_1 - \mu_2 = \mu_d$ from these data.
d. Test, at the 1% level of significance, the hypothesis that on average golf scores are lower with the new clubs.

7. A neighborhood home owners association suspects that the recent appraisal values of the houses in the neighborhood conducted by the county government for taxation purposes is too high. It hired a private company to appraise the values of ten houses in the neighborhood. The results, in thousands of dollars, are

House	County Government	Private Company
1	217	219
2	350	338
3	296	291
4	237	237
5	237	235
6	272	269
7	257	239
8	277	275
9	312	320
10	335	335

a. Give a point estimate for the difference between the mean private appraisal of all such homes and the government appraisal of all such homes.
b. Construct the 99% confidence interval based on these data for the difference.

c. Test, at the 1% level of significance, the hypothesis that appraised values by the county government of all such houses is greater than the appraised values by the private appraisal company.

8. In order to cut costs a wine producer is considering using duo or $1+1$ corks in place of full natural wood corks, but is concerned that it could affect buyers's perception of the quality of the wine. The wine producer shipped eight pairs of bottles of its best young wines to eight wine experts. Each pair includes one bottle with a natural wood cork and one with a duo cork. The experts are asked to rate the wines on a one to ten scale, higher numbers corresponding to higher quality. The results are:

Wine Expert	Duo Cork	Wood Cork
1	8.5	8.5
2	8.0	8.5
3	6.5	8.0
4	7.5	8.5
5	8.0	7.5
6	8.0	8.0
7	9.0	9.0
8	7.0	7.5

a. Give a point estimate for the difference between the mean ratings of the wine when bottled are sealed with different kinds of corks.
b. Construct the 90% confidence interval based on these data for the difference.
c. Test, at the 10% level of significance, the hypothesis that on the average duo corks decrease the rating of the wine.

9. Engineers at a tire manufacturing corporation wish to test a new tire material for increased durability. To test the tires under realistic road conditions, new front tires are mounted on each of 11 company cars, one tire made with a production material and the other with the experimental material. After a fixed period the 11 pairs were measured for wear. The amount of wear for each tire (in mm) is shown in the table:

Car	Production	Experimental
1	5.1	5.0
2	6.5	6.5
3	3.6	3.1
4	3.5	3.7
5	5.7	4.5
6	5.0	4.1
7	6.4	5.3
8	4.7	2.6
9	3.2	3.0
10	3.5	3.5
11	6.4	5.1

a. Give a point estimate for the difference in mean wear.
b. Construct the 99% confidence interval for the difference based on these data.
c. Test, at the 1% level of significance, the hypothesis that the mean wear with the experimental material is less than that for the production material.

10. A marriage counselor administered a test designed to measure overall contentment to 30 randomly selected married couples. The scores for each couple are given below. A higher number corresponds to greater contentment or happiness.

Couple	Husband	Wife
1	47	44
2	44	46
3	49	44
4	53	44
5	42	43
6	45	45
7	48	47
8	45	44
9	52	44
10	47	42
11	40	34
12	45	42
13	40	43
14	46	41
15	47	45
16	46	45
17	46	41
18	46	41
19	44	45
20	45	43
21	48	38
22	42	46
23	50	44
24	46	51
25	43	45
26	50	40
27	46	46
28	42	41
29	51	41
30	46	47

a. Test, at the 1% level of significance, the hypothesis that on average men and women are not equally happy in marriage.
b. Test, at the 1% level of significance, the hypothesis that on average men are happier than women in marriage.

https://stats.libretexts.org/@go/page/1113

Large Data Set Exercises

Large Data Sets are absent

11. Large Data Set 5 lists the scores for 25 randomly selected students on practice SAT reading tests before and after taking a two-week SAT preparation course. Denote the population of all students who have taken the course as Population 1 and the population of all students who have not taken the course as Population 2.
    a. Compute the 25 differences in the order **after - before**, their mean $\bar{d}$, and their sample standard deviation s_d.
    b. Give a point estimate for $\mu_d = \mu_1 - \mu_2$, the difference in the mean score of all students who have taken the course and the mean score of all who have not.
    c. Construct a 98% confidence interval for μ_d.
    d. Test, at the 1% level of significance, the hypothesis that the mean SAT score increases by at least ten points by taking the two-week preparation course.
12. Large Data Set 12 lists the scores on one round for 75 randomly selected members at a golf course, first using their own original clubs, then two months later after using new clubs with an experimental design. Denote the population of all golfers using their own original clubs as Population 1 and the population of all golfers using the new style clubs as Population 2.
    a. Compute the 75 differences in the order **original clubs - new clubs**, their mean $\bar{d}$, and their sample standard deviation s_d.
    b. Give a point estimate for $\mu_d = \mu_1 - \mu_2$, the difference in the mean score of all students who have taken the course and the mean score of all who have not.
    c. Construct a 90% confidence interval for μ_d.
    d. Test, at the 1% level of significance, the hypothesis that the mean SAT score increases by at least ten points by taking the two-week preparation course.
13. Consider the previous problem again. Since the data set is so large, it is reasonable to use the standard normal distribution instead of Student's t-distribution with 74 degrees of freedom.
    a. Construct a 90% confidence interval for μ_d using the standard normal distribution, meaning that the formula is $\bar{d} \pm z_{\alpha/2}\frac{s_d}{\sqrt{n}}$. (The computations done in part (a) of the previous problem still apply and need not be redone.) How does the result obtained here compare to the result obtained in part (c) of the previous problem?
    b. Test, at the 1% level of significance, the hypothesis that the mean golf score decreases by at least one stroke by using the new kind of clubs, using the standard normal distribution. (All the work done in part (d) of the previous problem applies, except the critical value is now z_α instead of z_α (or the p-value can be computed exactly instead of only approximated, if you used the p-value approach).) How does the result obtained here compare to the result obtained in part (c) of the previous problem?
    c. Construct the 99% confidence intervals for μ_d using both the t- and z-distributions. How much difference is there in the results now?

Answers

1. a. $\bar{d} = 7.4286,\ s_d = 0.9759$
   b. $\bar{d} = 7.4286$
   c. $(6.53, 8.33)$
   d. $T = 1.162,\ df = 6,\ t_{0.10} = 1.44$, do not reject H_0
2.
3. a. $\bar{d} = -14.25,\ s_d = 1.5$
   b. $\bar{d} = -14.25$
   c. $(-18.63, -9.87)$
   d. $T = -3.000,\ df = 3,\ \pm t_{0.05} = \pm 2.353$, reject H_0
4.
5. a. $\bar{d} = 25.2,\ s_d = 35.6609$
   b. $\bar{d} = 25.2$
   c. 25.2 ± 34.0
   d. $T = 1.580,\ df = 4,\ t_{0.10} = 1.533$, reject H_0 (takes less time)
6.
7. a. 3.2

b. 3.2 ± 7.5

c. $T = 1.392,\ df = 9,\ t_{0.10} = 2.821$, do not reject H_0 (government appraisals not higher)

8.

9. a. 0.65

b. 0.65 ± 0.69

c. $T = 3.014,\ df = 10,\ t_{0.10} = 2.764$, reject H_0 (experimental material wears less)

10.

11. a. $\bar{d} = 16.68,\ s_d = 10.77$

b. $\bar{d} = 16.68$

c. $(11.31, 22.05)$

d. $H_0 : \mu_1 - \mu_2 = 10\ vs\ H_a : \mu_1 - \mu_2 > 10$. Test Statistic: $T = 3.1014,\ df = 11$. Rejection Region: $[2.492, \infty)$. Decision: Reject H_0.

12.

13. a. $(1.6266, 2.6401)$. Endpoints change in the third decimal place.

b. $H_0 : \mu_1 - \mu_2 = 1\ vs\ H_a : \mu_1 - \mu_2 > 1$. Test Statistic: $Z = 3.6791$. Rejection Region: $[2.33, \infty)$. Decision: Reject H_0. The decision is the same as in the previous problem.

c. Using the t-distribution, $(1.3188, 2.9478)$. Using the z-distribution, $(1.3401, 2.9266)$. There is a difference.

9.4: Comparison of Two Population Proportions

Basic

1. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

a. 90% confidence

$$\begin{aligned} n_1 &= 1670,\ \hat{p_1} = 0.42 \\ n_2 &= 900,\ \hat{p_2} = 0.38 \end{aligned} \tag{9.E.53}$$

b. 95% confidence

$$\begin{aligned} n_1 &= 600,\ \hat{p_1} = 0.84 \\ n_2 &= 420,\ \hat{p_2} = 0.67 \end{aligned} \tag{9.E.54}$$

2. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

a. 98% confidence

$$\begin{aligned} n_1 &= 750,\ \hat{p_1} = 0.64 \\ n_2 &= 800,\ \hat{p_2} = 0.51 \end{aligned} \tag{9.E.55}$$

b. 99.5% confidence

$$\begin{aligned} n_1 &= 250,\ \hat{p_1} = 0.78 \\ n_2 &= 250,\ \hat{p_2} = 0.51 \end{aligned} \tag{9.E.56}$$

3. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

a. 80% confidence

$$\begin{aligned} n_1 &= 300,\ \hat{p_1} = 0.255 \\ n_2 &= 400,\ \hat{p_2} = 0.193 \end{aligned} \tag{9.E.57}$$

b. 95% confidence

$$\begin{aligned} n_1 &= 3500,\ \hat{p_1} = 0.147 \\ n_2 &= 3750,\ \hat{p_2} = 0.131 \end{aligned} \tag{9.E.58}$$

4. Construct the confidence interval for $p_1 - p_2$ for the level of confidence and the data given. (The samples are sufficiently large.)

a. 99% confidence

$$\begin{aligned} n_1 &= 2250,\ \hat{p_1} = 0.915 \\ n_2 &= 2525,\ \hat{p_2} = 0.858 \end{aligned} \tag{9.E.59}$$

b. 95% confidence

$$\begin{aligned} n_1 &= 120,\ \hat{p_1} = 0.650 \\ n_2 &= 200,\ \hat{p_2} = 0.505 \end{aligned} \tag{9.E.60}$$

5. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 > 0$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 &= 1200,\ \hat{p_1} = 0.42 \\ n_2 &= 1200,\ \hat{p_2} = 0.40 \end{aligned} \tag{9.E.61}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 &= 550,\ \hat{p_1} = 0.61 \\ n_2 &= 600,\ \hat{p_2} = 0.67 \end{aligned} \tag{9.E.62}$$

6. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.05$ *vs* $H_a : p_1 - p_2 > 0.05$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 &= 1100,\ \hat{p_1} = 0.57 \\ n_2 &= 1100,\ \hat{p_2} = 0.48 \end{aligned} \tag{9.E.63}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 &= 800,\ \hat{p_1} = 0.39 \\ n_2 &= 900,\ \hat{p_2} = 0.43 \end{aligned} \tag{9.E.64}$$

7. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.25$ *vs* $H_a : p_1 - p_2 < 0.25$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 &= 1400,\ \hat{p_1} = 0.57 \\ n_2 &= 1200,\ \hat{p_2} = 0.37 \end{aligned} \tag{9.E.65}$$

b. Test $H_0 : p_1 - p_2 = 0.16$ *vs* $H_a : p_1 - p_2 \neq 0.16$ @ $\alpha = 0.02$

$$\begin{aligned} n_1 &= 750,\ \hat{p_1} = 0.43 \\ n_2 &= 600,\ \hat{p_2} = 0.22 \end{aligned} \tag{9.E.66}$$

8. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0.08$ *vs* $H_a : p_1 - p_2 > 0.08$ @ $\alpha = 0.025$

$$\begin{aligned} n_1 &= 450,\ \hat{p_1} = 0.67 \\ n_2 &= 200,\ \hat{p_2} = 0.52 \end{aligned} \tag{9.E.67}$$

b. Test $H_0 : p_1 - p_2 = 0.02$ *vs* $H_a : p_1 - p_2 \neq 0.02$ @ $\alpha = 0.001$

$$\begin{aligned} n_1 &= 2700,\ \hat{p_1} = 0.837 \\ n_2 &= 2900,\ \hat{p_2} = 0.854 \end{aligned} \tag{9.E.68}$$

9. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)

a. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 < 0$ @ $\alpha = 0.005$

$$\begin{aligned} n_1 &= 1100,\ \hat{p_1} = 0.22 \\ n_2 &= 1300,\ \hat{p_2} = 0.27 \end{aligned} \tag{9.E.69}$$

b. Test $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 \neq 0$ @ $\alpha = 0.01$

$$\begin{aligned} n_1 &= 650,\ \hat{p_1} = 0.35 \\ n_2 &= 650,\ \hat{p_2} = 0.41 \end{aligned} \tag{9.E.70}$$

10. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)
    a. Test $H_0 : p_1 - p_2 = 0.15$ *vs* $H_a : p_1 - p_2 > 0.15$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 &= 950,\ \hat{p_1} = 0.41 \\ n_2 &= 500,\ \hat{p_2} = 0.23 \end{aligned} \tag{9.E.71}$$

    b. Test $H_0 : p_1 - p_2 = 0.10$ *vs* $H_a : p_1 - p_2 \neq 0.10$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 &= 220,\ \hat{p_1} = 0.92 \\ n_2 &= 160,\ \hat{p_2} = 0.78 \end{aligned} \tag{9.E.72}$$

11. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)
    a. Test $H_0 : p_1 - p_2 = 0.22$ *vs* $H_a : p_1 - p_2 > 0.22$ @ $\alpha = 0.05$

$$\begin{aligned} n_1 &= 90,\ \hat{p_1} = 0.72 \\ n_2 &= 75,\ \hat{p_2} = 0.40 \end{aligned} \tag{9.E.73}$$

    b. Test $H_0 : p_1 - p_2 = 0.37$ *vs* $H_a : p_1 - p_2 \neq 0.37$ @ $\alpha = 0.02$

$$n_1 = 425,\ \hat{p_1} = 0.772\ n_2 = 425,\ \hat{p_2} = 0.331 \tag{9.E.74}$$

12. Perform the test of hypotheses indicated, using the data given. Use the critical value approach. Compute the p-value of the test as well. (The samples are sufficiently large.)
    a. Test $H_0 : p_1 - p_2 = 0.50$ *vs* $H_a : p_1 - p_2 < 0.50$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 &= 40,\ \hat{p_1} = 0.65 \\ n_2 &= 55,\ \hat{p_2} = 0.24 \end{aligned} \tag{9.E.75}$$

    b. Test $H_0 : p_1 - p_2 = 0.30$ *vs* $H_a : p_1 - p_2 \neq 0.30$ @ $\alpha = 0.10$

$$\begin{aligned} n_1 &= 7500,\ \hat{p_1} = 0.664 \\ n_2 &= 1000,\ \hat{p_2} = 0.319 \end{aligned} \tag{9.E.76}$$

Applications

In all the remaining exercsises the samples are sufficiently large (so this need not be checked).

13. Voters in a particular city who identify themselves with one or the other of two political parties were randomly selected and asked if they favor a proposal to allow citizens with proper license to carry a concealed handgun in city parks. The results are:

	Party A	Party B
Sample size, *n*	150	200
Number in favor, *x*	90	140

    a. Give a point estimate for the difference in the proportion of all members of Party A and all members of Party B who favor the proposal.
    b. Construct the 95% confidence interval for the difference, based on these data.
    c. Test, at the 5% level of significance, the hypothesis that the proportion of all members of Party A who favor the proposal is less than the proportion of all members of Party B who do.
    d. Compute the p-value of the test.

14. To investigate a possible relation between gender and handedness, a random sample of 320 adults was taken, with the following results:

	Men	Women
Sample size, *n*	168	152
Number of left-handed, *x*	24	9

a. Give a point estimate for the difference in the proportion of all men who are left-handed and the proportion of all women who are left-handed.
b. Construct the 95% confidence interval for the difference, based on these data.
c. Test, at the 5% level of significance, the hypothesis that the proportion of men who are left-handed is greater than the proportion of women who are.
d. Compute the p-value of the test.

15. A local school board member randomly sampled private and public high school teachers in his district to compare the proportions of National Board Certified (NBC) teachers in the faculty. The results were:

	Private Schools	Public Schools
Sample size, *n*	80	520
Proportion of NBC teachers, $\hat{p}$	0.175	0.150

a. Give a point estimate for the difference in the proportion of all teachers in area public schools and the proportion of all teachers in private schools who are National Board Certified.
b. Construct the 90% confidence interval for the difference, based on these data.
c. Test, at the 10% level of significance, the hypothesis that the proportion of all public school teachers who are National Board Certified is less than the proportion of private school teachers who are.
d. Compute the p-value of the test.

16. In professional basketball games, the fans of the home team always try to distract free throw shooters on the visiting team. To investigate whether this tactic is actually effective, the free throw statistics of a professional basketball player with a high free throw percentage were examined. During the entire last season, this player had 656 free throws, 420 in home games and 236 in away games. The results are summarized below.

	Home	Away
Sample size, *n*	420	236
Free throw percent, $\hat{p}$	81.5%	78.8%

a. Give a point estimate for the difference in the proportion of free throws made at home and away.
b. Construct the 90% confidence interval for the difference, based on these data.
c. Test, at the 10% level of significance, the hypothesis that there exists a home advantage in free throws.
d. Compute the p-value of the test.

17. Randomly selected middle-aged people in both China and the United States were asked if they believed that adults have an obligation to financially support their aged parents. The results are summarized below.

	China	USA
Sample size, *n*	1300	150
Number of yes, *x*	1170	110

Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that there exists a cultural difference in attitude regarding this question.

18. A manufacturer of walk-behind push mowers receives refurbished small engines from two new suppliers, A and B. It is not uncommon that some of the refurbished engines need to be lightly serviced before they can be fitted into mowers. The mower manufacturer recently received 100 engines from each supplier. In the shipment from A, 13 needed further service. In the shipment from B, 10 needed further service. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that there exists a difference in the proportions of engines from the two suppliers needing service.

Large Data Set Exercises

Large Data Sets are absent

19. Large Data Sets 6A and 6B record results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Let the population of all voters in region 1 be denoted Population 1 and the population of all voters in region 2 be denoted Population 2. Let p_1 be the proportion of voters in Population 1 who prefer Candidate A, and p_2 the proportion in Population 2 who do.
    a. Find the relevant sample proportions $\hat{p_1}$ and $\hat{p_2}$.
    b. Construct a point estimate for $p_1 - p_2$.
    c. Construct a 95% confidence interval for $p_1 - p_2$.
    d. Test, at the 5% level of significance, the hypothesis that the same proportion of voters in the two regions favor Candidate A, against the alternative that a larger proportion in Population 2 do.
20. Large Data Set 11 records the results of samples of real estate sales in a certain region in the year 2008 (lines 2 through 536) and in the year 2010 (lines 537 through 1106). Foreclosure sales are identified with a 1 in the second column. Let all real estate sales in the region in 2008 be Population 1 and all real estate sales in the region in 2010 be Population 2.
    a. Use the sample data to construct point estimates $\hat{p_1}$ and $\hat{p_2}$ of the proportions p_1 and p_2 of all real estate sales in this region in 2008 and 2010 that were foreclosure sales. Construct a point estimate of $p_1 - p_2$.
    b. Use the sample data to construct a 90% confidence for $p_1 - p_2$.
    c. Test, at the 10% level of significance, the hypothesis that the proportion of real estate sales in the region in 2010 that were foreclosure sales was greater than the proportion of real estate sales in the region in 2008 that were foreclosure sales. (The default is that the proportions were the same.)

Answers

1. a. $(0.0068, 0.0732)$
   b. $(0.1163, 0.2237)$
2.
3. a. $(0.0210, 0.1030)$
   b. $(0.0001, 0.0319)$
4.
5. a. $Z = 0.996,\ z_{0.10} = 1.282,\ \text{p-value} = 0.1587$, do not reject H_0
   b. $Z = -2.120,\ \pm z_{0.025} = \pm 1.960,\ \text{p-value} = 0.0340$, reject H_0
6.
7. a. $Z = -2.602,\ -z_{0.005} = -2.576,\ \text{p-value} = 0.0047$, reject H_0
   b. $Z = 2.020,\ \pm z_{0.01} = \pm 2.326,\ \text{p-value} = 0.0434$, do not reject H_0
8.
9. a. $Z = -2.85,\ \text{p-value} = 0.0022$, reject H_0
   b. $Z = -2.23,\ \text{p-value} = 0.0258$, do not reject H_0
10.
11. a. $Z = 1.36,\ \text{p-value} = 0.0869$, do not reject H_0
    b. $Z = 2.32,\ \text{p-value} = 0.0204$, do not reject H_0
12.
13. a. -0.10
    b. -0.10 ± 0.101
    c. $Z = -1.943,\ -z_{0.05} = -1.645$, reject H_0 (fewer in Party A favor)
    d. $\text{p-value} = 0.0262$
14.
15. a. 0.025
    b. 0.025 ± 0.0745
    c. $Z = 0.552,\ z_{0.10} = 1.282$, do not reject H_0 (as many public school teachers are certified)
    d. $\text{p-value} = 0.2912$
16.
17. $Z = 4.498,\ \pm z_{0.005} = \pm 2.576$, reject H_0 (different)

18.

19. a. $\hat{p_1} = 0.355$ and $\hat{p_2} = 0.41$
    b. $\hat{p_1} - \hat{p_2} = -0.055$
    c. $(-0.1501, 0.0401)$
    d. $H_0 : p_1 - p_2 = 0$ *vs* $H_a : p_1 - p_2 < 0$. Test Statistic: $Z = -1.1335$. Rejection Region: $(-\infty, -1.645]$. Decision: Fail to reject H_0.

9.5 Sample Size Considerations

Basic

1. Estimate the common sample size n of equally sized independent samples needed to estimate $\mu_1 - \mu_2$ as specified when the population standard deviations are as shown.
   a. 90% confidence, to within 3 units, $\sigma_1 = 10$ and $\sigma_2 = 7$
   b. 99% confidence, to within 4 units, $\sigma_1 = 6.8$ and $\sigma_2 = 9.3$
   c. 95% confidence, to within 5 units, $\sigma_1 = 22.6$ and $\sigma_2 = 31.8$
2. Estimate the common sample size n of equally sized independent samples needed to estimate $\mu_1 - \mu_2$ as specified when the population standard deviations are as shown.
   a. 80% confidence, to within 2 units, $\sigma_1 = 14$ and $\sigma_2 = 23$
   b. 90% confidence, to within 0.3 units, $\sigma_1 = 1.3$ and $\sigma_2 = 0.8$
   c. 99% confidence, to within 11 units, $\sigma_1 = 42$ and $\sigma_2 = 37$
3. Estimate the number n of pairs that must be sampled in order to estimate $\mu_d = \mu_1 - \mu_2$ as specified when the standard deviation s_d of the population of differences is as shown.
   a. 80% confidence, to within 6 units, $\sigma_d = 26.5$
   b. 95% confidence, to within 4 units, $\sigma_d = 12$
   c. 90% confidence, to within 5.2 units, $\sigma_d = 11.3$
4. Estimate the number n of pairs that must be sampled in order to estimate $\mu_d = \mu_1 - \mu_2$ as specified when the standard deviation s_d of the population of differences is as shown.
   a. 90% confidence, to within 20 units, $\sigma_d = 75.5$
   b. 95% confidence, to within 11 units, $\sigma_d = 31.4$
   c. 99% confidence, to within 1.8 units, $\sigma_d = 4$
5. Estimate the minimum equal sample sizes $n_1 = n_2$ necessary in order to estimate $p_1 - p_2$ as specified.
   a. 80% confidence, to within 0.05 (five percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.20$ and $p_2 \approx 0.65$
   b. 90% confidence, to within 0.02 (two percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.75$ and $p_2 \approx 0.63$
   c. 95% confidence, to within 0.10 (ten percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.11$ and $p_2 \approx 0.37$
6. Estimate the minimum equal sample sizes $n_1 = n_2$ necessary in order to estimate $p_1 - p_2$ as specified.
   a. 80% confidence, to within 0.02 (two percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.78$ and $p_2 \approx 0.65$
   b. 90% confidence, to within 0.05 (five percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.12$ and $p_2 \approx 0.24$
   c. 95% confidence, to within 0.10 (ten percentage points)
      i. when no prior knowledge of p_1 or p_2 is available
      ii. when prior studies indicate that $p_1 \approx 0.14$ and $p_2 \approx 0.21$

Applications

7. An educational researcher wishes to estimate the difference in average scores of elementary school children on two versions of a 100-point standardized test, at 99% confidence and to within two points. Estimate the minimum equal sample sizes necessary if it is known that the standard deviation of scores on different versions of such tests is 4.9.
8. A university administrator wishes to estimate the difference in mean grade point averages among all men affiliated with fraternities and all unaffiliated men, with 95% confidence and to within 0.15. It is known from prior studies that the standard deviations of grade point averages in the two groups have common value 0.4. Estimate the minimum equal sample sizes necessary to meet these criteria.
9. An automotive tire manufacturer wishes to estimate the difference in mean wear of tires manufactured with an experimental material and ordinary production tire, with 90% confidence and to within 0.5 mm. To eliminate extraneous factors arising from different driving conditions the tires will be tested in pairs on the same vehicles. It is known from prior studies that the standard deviations of the differences of wear of tires constructed with the two kinds of materials is 1.75 mm. Estimate the minimum number of pairs in the sample necessary to meet these criteria.
10. To assess to the relative happiness of men and women in their marriages, a marriage counselor plans to administer a test measuring happiness in marriage to n randomly selected married couples, record the their test scores, find the differences, and then draw inferences on the possible difference. Let μ_1 and μ_2 be the true average levels of happiness in marriage for men and women respectively as measured by this test. Suppose it is desired to find a 90% confidence interval for estimating $\mu_d = \mu_1 - \mu_2$ to within two test points. Suppose further that, from prior studies, it is known that the standard deviation of the differences in test scores is $\sigma_d \approx 10$. What is the minimum number of married couples that must be included in this study?
11. A journalist plans to interview an equal number of members of two political parties to compare the proportions in each party who favor a proposal to allow citizens with a proper license to carry a concealed handgun in public parks. Let p_1 and p_2 be the true proportions of members of the two parties who are in favor of the proposal. Suppose it is desired to find a 95% confidence interval for estimating $p_1 - p_2$ to within 0.05. Estimate the minimum equal number of members of each party that must be sampled to meet these criteria.
12. A member of the state board of education wants to compare the proportions of National Board Certified (NBC) teachers in private high schools and in public high schools in the state. His study plan calls for an equal number of private school teachers and public school teachers to be included in the study. Let p_1 and p_2 be these proportions. Suppose it is desired to find a 99% confidence interval that estimates $p_1 - p_2$ to within 0.05.
    a. Supposing that both proportions are known, from a prior study, to be approximately 0.15, compute the minimum common sample size needed.
    b. Compute the minimum common sample size needed on the supposition that nothing is known about the values of p_1 and p_2.

Answers

1. a. $n_1 = n_2 = 45$
   b. $n_1 = n_2 = 56$
   c. $n_1 = n_2 = 234$
2.
3. a. $n_1 = n_2 = 33$
   b. $n_1 = n_2 = 35$
   c. $n_1 = n_2 = 13$
4.
5. a. i. $n_1 = n_2 = 329$
      ii. $n_1 = n_2 = 255$
   b. i. $n_1 = n_2 = 3383$
      ii. $n_1 = n_2 = 2846$
   c. i. $n_1 = n_2 = 193$
      ii. $n_1 = n_2 = 128$
6.
7. $n_1 = n_2 \approx 80$
8.
9. $n_1 = n_2 \approx 34$

10.
11. $n_1 = n_2 \approx 769$

This page titled 9.E: Two-Sample Problems (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **9.E: Two-Sample Problems (Exercises)** has no license indicated.

CHAPTER OVERVIEW

10: Correlation and Regression

Our interest in this chapter is in situations in which we can associate to each element of a population or sample two measurements x and y, particularly in the case that it is of interest to use the value of x to predict the value of y. For example, the population could be the air in automobile garages, x could be the electrical current produced by an electrochemical reaction taking place in a carbon monoxide meter, and y the concentration of carbon monoxide in the air. In this chapter we will learn statistical methods for analyzing the relationship between variables x and y in this context.

10.1: Linear Relationships Between Variables
10.2: The Linear Correlation Coefficient
10.3: Modelling Linear Relationships with Randomness Present
10.4: The Least Squares Regression Line
10.5: Statistical Inferences About β_1
10.6: The Coefficient of Determination
10.7: Estimation and Prediction
10.8: A Complete Example
10.9: Formula List
10.E: Correlation and Regression (Exercises)

This page titled 10: Correlation and Regression is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Linear Relationships Between Variables

Learning Objectives

To learn what it means for two variables to exhibit a relationship that is close to linear, but which contains an element of randomness

The following table gives examples of the kinds of pairs of variables which could be of interest from a statistical point of view.

x	y
Predictor or independent variable	Response or dependent variable
Temperature in degrees Celsius	Temperature in degrees Fahrenheit
Area of a house (sq.ft.)	Value of the house
Age of a particular make and model car	Resale value of the car
Amount spent by a business on advertising in a year	Revenue received that year
Height of a 25-year-old man	Weight of the man

The first line in the table is different from all the rest because in that case and no other the relationship between the variables is deterministic: once the value of x is known the value of y is completely determined. In fact there is a formula for y in terms of x:

$$y = 95x + 32$$

Choosing several values for x and computing the corresponding value for y for each one using the formula gives the table

x	-40	-15	0	20	50
y	-40	5	32	68	122

We can plot these data by choosing a pair of perpendicular lines in the plane, called the coordinate axes, as shown in Figure 10.1.1. Then to each pair of numbers in the table we associate a unique point in the plane, the point that lies x units to the right of the vertical axis (to the left if $x < 0$) and y units above the horizontal axis (below if $y < 0$). The relationship between x and y is called a linear relationship because the points so plotted all lie on a single straight line. The number 95 in the equation $y = 95x + 32$ is the slope of the line, and measures its steepness. It describes how y changes in response to a change in x: if x increases by 1 unit then y increases (since 95 is positive) by 95 unit. If the slope had been negative then y would have decreased in response to an increase in x. The number 32 in the formula $y = 95x + 32$ is the y-intercept of the line; it identifies where the line crosses the y-axis. You may recall from an earlier course that every non-vertical line in the plane is described by an equation of the form $y = mx + b$, where m is the slope of the line and b is its y-intercept.

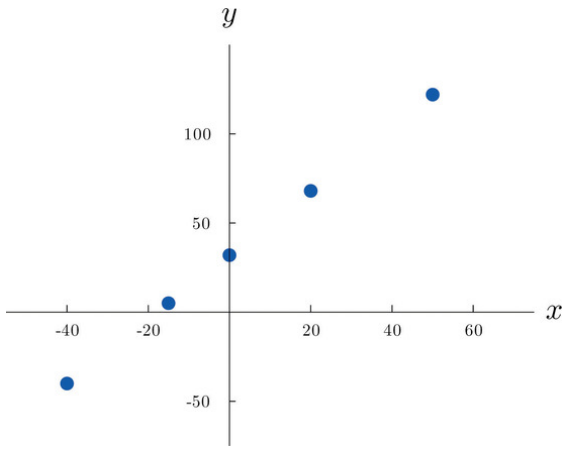


Figure 10.1.1: Plot of Celsius and Fahrenheit Temperature Pairs

The relationship between x and y in the temperature example is deterministic because once the value of x is known, the value of y is completely determined. In contrast, all the other relationships listed in the table above have an element of randomness in them. Consider the relationship described in the last line of the table, the height x of a man aged 25 and his weight y. If we were to randomly select several 25-year-old men and measure the height and weight of each one, we might obtain a collection of (x, y) pairs something like this:

$$(68, 151)\ \ (72, 163)\ \ (69, 146)\ \ (72, 180)\ \ (70, 157)\ \ (73, 170)\ \ (70, 164)\ \ (73, 175)$$
$$(71, 171)\ \ (74, 178)\ \ (72, 160)\ \ (75, 188)$$

A plot of these data is shown in Figure $10.1.2$. Such a plot is called a *scatter diagram* or *scatter plot*. Looking at the plot it is evident that there exists a linear relationship between height x and weight y, but not a perfect one. The points appear to be following a line, but not exactly. There is an element of randomness present.

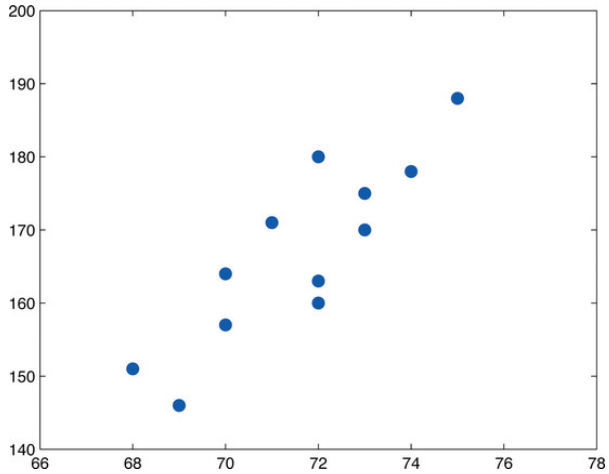


Figure $10.1.2$: Plot of Height and Weight Pairs

In this chapter we will analyze situations in which variables x and y exhibit such a linear relationship with randomness. The level of randomness will vary from situation to situation. In the introductory example connecting an electric current and the level of carbon monoxide in air, the relationship is almost perfect. In other situations, such as the height and weights of individuals, the connection between the two variables involves a high degree of randomness. In the next section we will see how to quantify the strength of the linear relationship between two variables.

Key Takeaway

- Two variables x and y have a deterministic linear relationship if points plotted from (x, y) pairs lie exactly along a single straight line.
- In practice it is common for two variables to exhibit a relationship that is close to linear but which contains an element, possibly large, of randomness.

This page titled 10.1: Linear Relationships Between Variables is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.2: The Linear Correlation Coefficient

Learning Objectives

To learn what the linear correlation coefficient is, how to compute it, and what it tells us about the relationship between two variables x and y

Figure 10.2.1 illustrates linear relationships between two variables x and y of varying strengths. It is visually apparent that in the situation in panel (a), x could serve as a useful predictor of y, it would be less useful in the situation illustrated in panel (b), and in the situation of panel (c) the linear relationship is so weak as to be practically nonexistent. The *linear correlation coefficient* is a number computed directly from the data that measures the strength of the linear relationship between the two variables x and y.

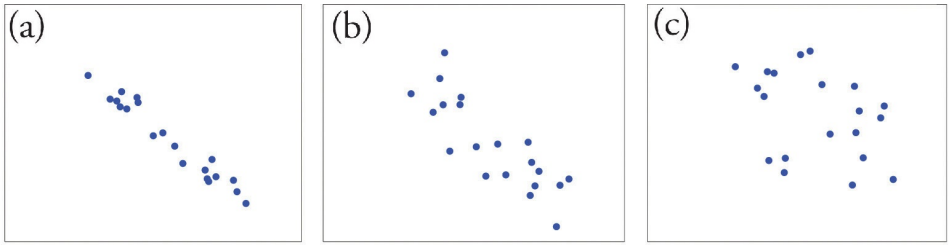


Figure 10.2.1: Linear Relationships of Varying Strengths

Definition: linear correlation coefficient

The *linear correlation coefficient* for a collection of n pairs x of numbers in a sample is the number r given by the formula

The linear correlation coefficient has the following properties, illustrated in Figure 10.2.2

1. The value of r lies between -1 and 1, inclusive.
2. The sign of r indicates the direction of the linear relationship between x and y:
3. The size of $|r|$ indicates the strength of the linear relationship between x and y:
    1. If $|r|$ is near 1 (that is, if r is near either 1 or -1), then the linear relationship between x and y is strong.
    2. If $|r|$ is near 0 (that is, if r is near 0 and of either sign). then the linear relationship between x and y is weak.

so that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{2.44.583}{\sqrt{(46.916)(1690.916)}} = 0.868$$

The number quantifies what is visually apparent from Figure 10.2.2 weights tends to increase linearly with height (r is positive) and although the relationship is not perfect, it is reasonably strong (r is near 1).

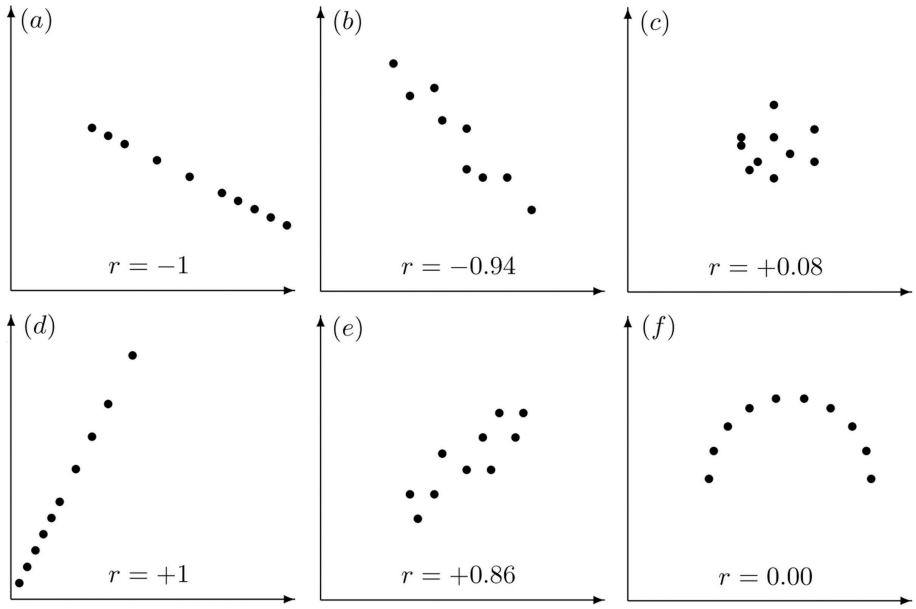


Figure 10.2.2: Linear Correlation Coefficient r. Pay particular attention to panel (f), which shows a perfectly deterministic relationship between x and y, but $f = 0$ because the relationship is not linear. (In this particular case the points lie on the top half of a circle.)

Example 10.2.1

Compute the linear correlation coefficient for the height and weight pairs plotted in Figure 10.2.2.

Solution:

Even for small data sets like this one computations are too long to do completely by hand. In actual practice the data are entered into a calculator or computer and a statistics program is used. In order to clarify the meaning of the formulas we will display the data and related quantities in tabular form. For each

	x	y	x^2	y^2
68	151	4624	10268	22801
69	146	4761	10074	21316
70	157	4900	10990	24649
70	164	4900	11480	26896
71	171	5041	12141	29241
72	160	5184	11520	25600
72	163	5184	11736	26569
72	180	5184	12960	32400
73	170	5329	12410	28900
73	175	5329	12775	30625
74	178	5476	13172	31684
75	188	5625	14100	35344
859	2003	61537	143626	336025

Key Takeaway

- The linear correlation coefficient measures the strength and direction of the linear relationship between two variables x and y.
- The sign of the linear correlation coefficient indicates the direction of the linear relationship between x and y.
- When r is near 1 or -1 the linear relationship is strong; when it is near 0 the linear relationship is weak.

This page titled 10.2: The Linear Correlation Coefficient is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Modelling Linear Relationships with Randomness Present

Learning Objectives

- To learn the framework in which the statistical analysis of the linear relationship between two variables x and y will be done

In this chapter we are dealing with a population for which we can associate to each element two measurements, x and y. We are interested in situations in which the value of x can be used to draw conclusions about the value of y, such as predicting the resale value y of a residential house based on its size x. Since the relationship between x and y is not deterministic, statistical procedures must be applied. For any statistical procedures, given in this book or elsewhere, the associated formulas are valid only under specific assumptions. The set of assumptions in simple linear regression are a mathematical description of the relationship between x and y. Such a set of assumptions is known as a *model*.

For each fixed value of x, a sub-population of the full population is determined, such as the collection of all houses with $2,100$ square feet of living space. For each element of that sub-population there is a measurement y, such as the value of any $2,100$-square-foot house. Let $E(y)$ denote the mean of all the y-values for each particular value of x. $E(y)$ can change from x-value to x-value, such as the mean value of all $2,100$-square-foot houses, the (different) mean value for all $2,500$-square foot-houses, and so on.

Our first assumption is that the relationship between x and the mean of the y-values in the sub-population determined by x is linear. This means that there exist numbers such that

$$y = \beta_1 x + \beta_0$$

This linear relationship is the reason for the word "linear" in "simple linear regression" below. (The word "simple" means that y depends on only one other variable and not two or more.)

Our next assumption is that for each value of x the y-values scatter about the mean $E(y)$ according to a normal distribution centered at $E(y)$ and with a standard deviation σ that is the same for every value of x. This is the same as saying that there exists a normally distributed random variable ε with mean 0 and standard deviation σ so that the relationship between x and y in the whole population is

$$y = \beta_1 x + \beta_0 + \epsilon$$

Our last assumption is that the random deviations associated with different observations are independent.

In summary, the model is:

Simple Linear Regression Model

For each point (x, y) in data set the y-value is an independent observation of

$$y = \beta_1 x + \beta_0 + \varepsilon$$

where β_1 and β_0 are fixed parameters and ε is a normally distributed random variable with mean 0 and an unknown standard deviation σ.

The line with equation

$$y = \beta_1 x + \beta_0$$

is called the **population regression line**.

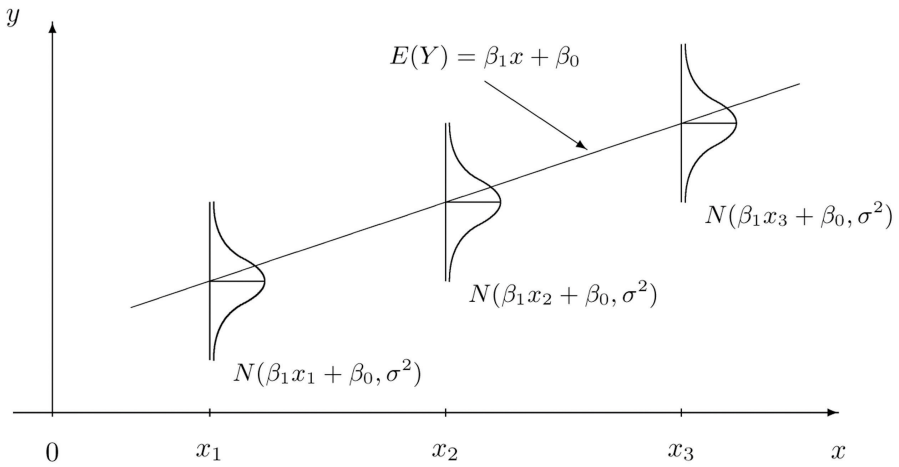


Figure 10.3.1: The Simple Linear Model Concept

It is conceptually important to view the model as a sum of two parts:

$$y = \underbrace{\beta_1 x + \beta_0}_{\text{Deterministic}} + \underbrace{\varepsilon}_{\text{Random}}$$

- **Deterministic Part**. The first part 0 is the equation that describes the trend in y as x increases. The line that we seem to see when we look at the scatter diagram is an approximation of the line

  $$y = \beta_1 x + \beta_0.$$

  There is nothing random in this part, and therefore it is called the *deterministic part* of the model.
- **Random Part**. The second part ε is a random variable, often called the *error term* or the *noise*. This part explains why the actual observed values of y are not exactly on but fluctuate near a line. Information about this term is important since only when one knows how much noise there is in the data can one know how trustworthy the detected trend is.

There are procedures for checking the validity of the three assumptions, but for us it will be sufficient to visually verify the linear trend in the data. If the data set is large then the points in the scatter diagram will form a band about an apparent straight line. The normality of ε with a constant standard deviation corresponds graphically to the band being of roughly constant width, and with most points concentrated near the middle of the band.

Fortunately, the three assumptions do not need to hold exactly in order for the procedures and analysis developed in this chapter to be useful.

Key Takeaway

- Statistical procedures are valid only when certain assumptions are valid. The assumptions underlying the analyses done in this chapter are graphically summarized in Figure 10.3.1.

This page titled 10.3: Modelling Linear Relationships with Randomness Present is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.4: The Least Squares Regression Line

Learning Objectives

- To learn how to measure how well a straight line fits a collection of data.
- To learn how to construct the least squares regression line, the straight line that best fits a collection of data.
- To learn the meaning of the slope of the least squares regression line.
- To learn how to use the least squares regression line to estimate the response variable y in terms of the predictor variable x.

Goodness of Fit of a Straight Line to Data

Once the scatter diagram of the data has been drawn and the model assumptions described in the previous sections at least visually verified (and perhaps the correlation coefficient r computed to quantitatively verify the linear trend), the next step in the analysis is to find the straight line that best fits the data. We will explain how to measure how well a straight line fits a collection of points by examining how well the line $y = \frac{1}{2}x - 1$ fits the data set

x	2	2	6	8	10
y	0	1	2	3	3

(which will be used as a running example for the next three sections). We will write the equation of this line as $\hat{y} = \frac{1}{2}x - 1$ with an accent on the y to indicate that the y-values computed using this equation are not from the data. We will do this with all lines approximating data sets. The line $\hat{y} = \frac{1}{2}x - 1$ was selected as one that seems to fit the data reasonably well.

The idea for measuring the goodness of fit of a straight line to data is illustrated in Figure 10.4.1, in which the graph of the line $\hat{y} = \frac{1}{2}x - 1$ has been superimposed on the scatter plot for the sample data set.

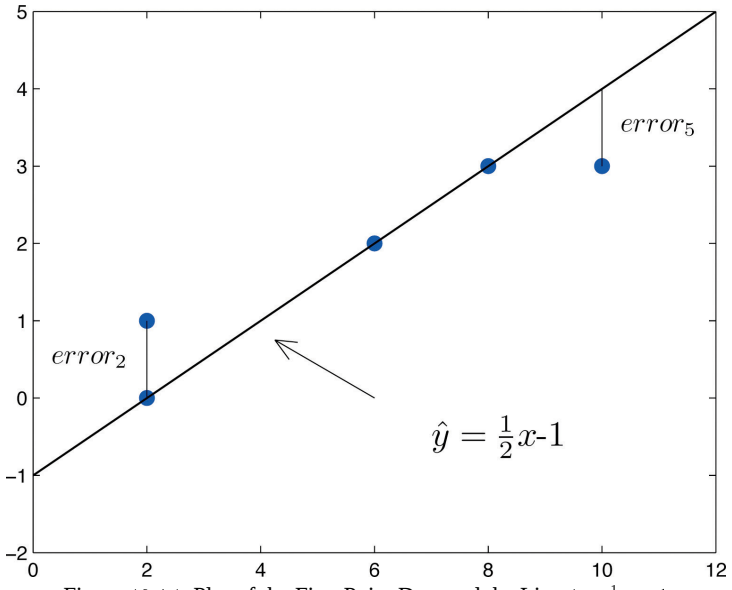


Figure 10.4.1: Plot of the Five-Point Data and the Line $\hat{y} = \frac{1}{2}x - 1$

To each point in the data set there is associated an "error," the positive or negative vertical distance from the point to the line: positive if the point is above the line and negative if it is below the line. The error can be computed as the actual y-value of the point minus the y-value $\hat{y}$ that is "predicted" by inserting the x-value of the data point into the formula for the line:

$$\text{error at data point(x,y)} = (\text{true y}) - (\text{predicted y}) = y - \hat{y}$$

The computation of the error for each of the five points in the data set is shown in Table 10.4.1.

Table 10.4.1: The Errors in Fitting Data with a Straight Line

	x	y	$\hat{y} = \frac{1}{2}x - 1$	$y - \hat{y}$	$(y - \hat{y})^2$
	2	0	0	0	0

	x	y	$\hat{y}=\frac{1}{2}x-1$	$y-\hat{y}$	$(y-\hat{y})^2$
	2	1	0	1	1
	6	2	2	0	0
	8	3	3	0	0
	10	3	4	−1	1
$\sum$	-	-	-	0	2

A first thought for a measure of the goodness of fit of the line to the data would be simply to add the errors at every point, but the example shows that this cannot work well in general. The line does not fit the data perfectly (no line can), yet because of cancellation of positive and negative errors the sum of the errors (the fourth column of numbers) is zero. Instead goodness of fit is measured by the sum of the squares of the errors. Squaring eliminates the minus signs, so no cancellation can occur. For the data and line in Figure 10.4.1 the sum of the squared errors (the last column of numbers) is 2. This number measures the goodness of fit of the line to the data.

Definition: goodness of fit

The goodness of fit of a line $\hat{y}=mx+b$ to a set of n pairs (x,y) of numbers in a sample is the sum of the squared errors

$$\sum(y-\hat{y})^2$$

(n terms in the sum, one for each data pair).

The Least Squares Regression Line

Given any collection of pairs of numbers (except when all the x-values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the least squares regression line. Moreover there are formulas for its slope and y-intercept.

Definition: least squares regression Line

Given a collection of pairs (x,y) of numbers (in which not all the x-values are the same), there is a line $\hat{y}=\hat{\beta}_1x+\hat{\beta}_0$ that best fits the data in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Its slope $\hat{\beta}_1$ and y-intercept $\hat{\beta}_0$ are computed using the formulas

$$\hat{\beta}_1=\frac{SS_{xy}}{SS_{xx}}$$

and

$$\hat{\beta}_0=\overline{y}-\hat{\beta}_1x$$

where

$$SS_{xx}=\sum x^2-\frac{1}{n}\left(\sum x\right)^2$$

and

$$SS_{xy}=\sum xy-\frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$\bar{x}$ is the mean of all the x-values, $\bar{y}$ is the mean of all the y-values, and n is the number of pairs in the data set.

The equation

$$\hat{y}=\hat{\beta}_1x+\hat{\beta}_0$$

specifying the least squares regression line is called the least squares regression equation.

Remember from Section 10.3 that the line with the equation $y = \beta_1 x + \beta_0$ is called the population regression line. The numbers $\hat{\beta}_1$ and $\hat{\beta}_0$ are statistics that estimate the population parameters β_1 and β_0.

We will compute the least squares regression line for the five-point data set, then for a more practical example that will be another running example for the introduction of new concepts in this and the next three sections.

Example 10.4.2

Find the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

and verify that it fits the data better than the line $\hat{y} = \frac{1}{2}x - 1$ considered in Section 10.4.1 above.

Solution

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
$\sum$	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 208 - \frac{1}{5}(28)^2 = 51.2$$

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 68 - \frac{1}{5}(28)(9) = 17.6$$

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{5} = 5.6$$
$$\bar{y} = \frac{\sum y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x - = 1.8 - (0.34375)(5.6) = -0.125$$

The least squares regression line for these data is

$$\hat{y} = 0.34375x - 0.125$$

The computations for measuring how well it fits the sample data are given in Table 10.4.2. The sum of the squared errors is the sum of the numbers in the last column, which is 0.75. It is less than 2, the sum of the squared errors for the fit of the line $\hat{y} = \frac{1}{2}x - 1$ to this data set.

Table 10.4.2 *The Errors in Fitting Data with the Least Squares Regression Line*

x	y	$\hat{y} = 0.34375x - 0.125$	$y - \hat{y}$	$(y - \hat{y})^2$
2	0	0.5625	−0.5625	0.31640625
2	1	0.5625	0.4375	0.19140625
6	2	1.9375	0.0625	0.00390625
8	3	2.6250	0.3750	0.14062500
10	3	3.3125	−0.3125	0.09765625

Example 10.4.3

Table 10.4.3 shows the age in years and the retail value in thousands of dollars of a random sample of ten automobiles of the same make and model.

1. Construct the scatter diagram.
2. Compute the linear correlation coefficient r. Interpret its value in the context of the problem.
3. Compute the least squares regression line. Plot it on the scatter diagram.
4. Interpret the meaning of the slope of the least squares regression line in the context of the problem.
5. Suppose a four-year-old automobile of this make and model is selected at random. Use the regression equation to predict its retail value.
6. Suppose a 20-year-old automobile of this make and model is selected at random. Use the regression equation to predict its retail value. Interpret the result.
7. Comment on the validity of using the regression equation to predict the price of a brand new automobile of this make and model.

Table 10.4.3: *Data on Age and Value of Used Automobiles of a Specific Make and Model*

x	2	3	3	3	4	4	5	5	5	6
y	28.7	24.8	26.0	30.5	23.8	24.6	23.8	20.4	21.6	22.1

Solution

1. The scatter diagram is shown in Figure 10.4.2.

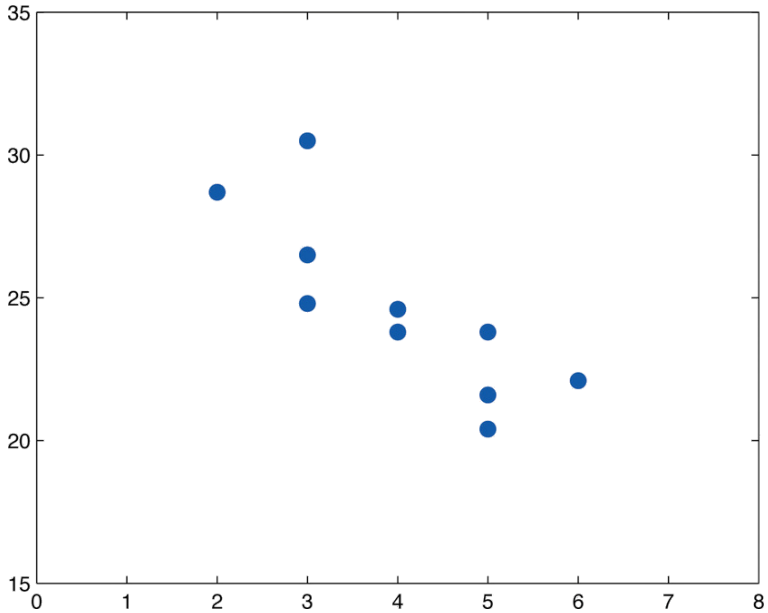


Figure 10.4.2: Scatter Diagram for Age and Value of Used Automobiles

https://stats.libretexts.org/@go/page/545

2. We must first compute $SS_{xx},\ SS_{xy},\ SS_{yy}$, which means computing $\sum x,\ \sum y,\ \sum x^2,\ \sum y^2$ and $\sum xy$. Using a computing device we obtain

$$\sum x = 40 \quad \sum y = 246.3 \quad \sum x^2 = 174 \quad \sum y^2 = 6154.15 \quad \sum xy = 956.5$$

Thus

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 174 - \frac{1}{10}(40)^2 = 14$$
$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 956.5 - \frac{1}{10}(40)(246.3) = -28.7$$
$$SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 6154.15 - \frac{1}{10}(246.3)^2 = 87.781$$

so that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{-28.7}{\sqrt{(14)(87.781)}} = -0.819$$

The age and value of this make and model automobile are moderately strongly negatively correlated. As the age increases, the value of the automobile tends to decrease.

3. Using the values of $\sum x$ and $\sum y$ computed in part (b),

$$\bar{x} = \frac{\sum x}{n} = \frac{40}{10} = 4$$
$$\bar{y} = \frac{\sum y}{n} = \frac{246.3}{10} = 24.63$$

Thus using the values of SS_{xx} and SS_{xy} from part (b),

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-28.7}{14} = -2.05$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x = 24.63 - (-2.05)(4) = 32.83$$

The equation $\bar{y} = \hat{\beta}_1 x + \hat{\beta}_0$ of the least squares regression line for these sample data is

$$\hat{y} = -2.05x + 32.83$$

Figure 10.4.3 shows the scatter diagram with the graph of the least squares regression line superimposed.

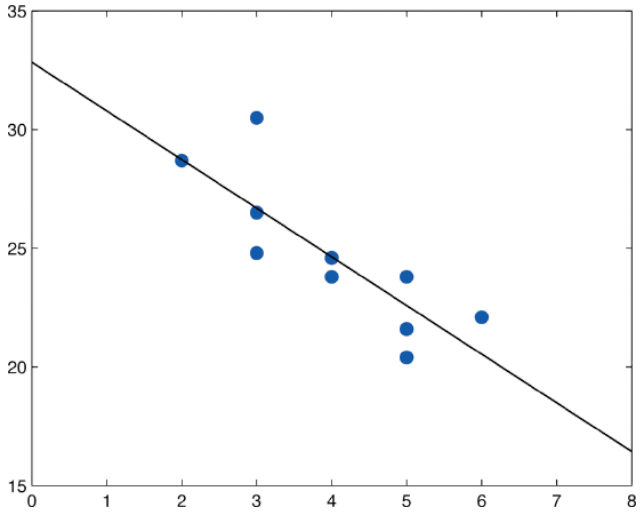


Figure 10.4.3: *Scatter Diagram and Regression Line for Age and Value of Used Automobiles*

4. The slope -2.05 means that for each unit increase in x (additional year of age) the average value of this make and model vehicle decreases by about 2.05 units (about $\$2,050$).

5. Since we know nothing about the automobile other than its age, we assume that it is of about average value and use the average value of all four-year-old vehicles of this make and model as our estimate. The average value is simply the value of $\hat{y}$ obtained when the number 4 is inserted for x in the least squares regression equation:

$$\hat{y} = -2.05(4) + 32.83 = 24.63$$

which corresponds to $\$24,630$.

6. Now we insert $x = 20$ into the least squares regression equation, to obtain

$$\hat{y} = -2.05(20) + 32.83 = -8.17$$

which corresponds to $-\$8,170$. Something is wrong here, since a negative makes no sense. The error arose from applying the regression equation to a value of x not in the range of x-values in the original data, from two to six years. Applying the regression equation $\bar{y} = \hat{\beta}_1 x + \hat{\beta}_0$ to a value of x outside the range of x-values in the data set is called extrapolation. It is an invalid use of the regression equation and should be avoided.

7. The price of a brand new vehicle of this make and model is the value of the automobile at age 0. If the value $x = 0$ is inserted into the regression equation the result is always $\hat{\beta}_0$, the y-intercept, in this case 32.83, which corresponds to $\$32,830$. But this is a case of extrapolation, just as part (f) was, hence this result is invalid, although not obviously so. In the context of the problem, since automobiles tend to lose value much more quickly immediately after they are purchased than they do after they are several years old, the number $\$32,830$ is probably an underestimate of the price of a new automobile of this make and model.

For emphasis we highlight the points raised by parts (f) and (g) of the example.

Definition: extrapolation

The process of using the least squares regression equation to estimate the value of y at a value of x that does not lie in the range of the x-values in the data set that was used to form the regression line is called extrapolation. It is an invalid use of the regression equation that can lead to errors, hence should be avoided.

The Sum of the Squared Errors SSE

In general, in order to measure the goodness of fit of a line to a set of data, we must compute the predicted y-value $\hat{y}$ at every point in the data set, compute each error, square it, and then add up all the squares. In the case of the least squares regression line, however, the line that best fits the data, the sum of the squared errors can be computed directly from the data using the following formula

The sum of the squared errors for the least squares regression line is denoted by SSE. It can be computed using the formula

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

Example 10.4.4

Find the sum of the squared errors SSE for the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

Do so in two ways:

1. using the definition $\sum (y - \hat{y})^2$;
2. using the formula $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$.

Solution

1. The least squares regression line was computed in "Example $10.4.2$" and is $\hat{y} = 0.34375x - 0.125$. SSE was found at the end of that example using the definition $\sum (y - \hat{y})^2$. The computations were tabulated in Table $10.4.2$. SSE is the sum of the numbers in the last column, which is 0.75.
2. The numbers SS_{xy} and $\hat{\beta}_1$ were already computed in "Example $10.4.2$" in the process of finding the least squares regression line. So was the number $\sum y = 9$. We must compute SS_{yy}. To do so it is necessary to first compute

$$\sum y^2 = 0 + 1^2 + 2^2 + 3^2 + 3^2 = 23$$

Then

$$SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 23 - \frac{1}{5}(9)^2 = 6.8$$

so that

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 6.8 - (0.34375)(17.6) = 0.75$$

Example 10.4.5

Find the sum of the squared errors SSE for the least squares regression line for the data set, presented in Table 10.4.3, on age and values of used vehicles in "Example 10.4.3".

Solution

From "Example 10.4.3" we already know that

$$SS_{xy} = -28.7,\ \hat{\beta}_1 = -2.05,\ \text{and}\ \sum y = 246.3$$

To compute SS_{yy} we first compute

$$\sum y^2 = 28.7^2 + 24.8^2 + 26.0^2 + 30.5^2 + 23.8^2 + 24.6^2 + 23.8^2 + 20.4^2 + 21.6^2 + 22.1^2 = 6154.15$$

Then

$$SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 6154.15 - \frac{1}{10}(246.3)^2 = 87.781$$

Therefore

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 87.781 - (-2.05)(-28.7) = 28.946$$

Key Takeaway

- How well a straight line fits a data set is measured by the sum of the squared errors.
- The least squares regression line is the line that best fits the data. Its slope and y-intercept are computed from the data using formulas.
- The slope $\hat{\beta}_1$ of the least squares regression line estimates the size and direction of the mean change in the dependent variable y when the independent variable x is increased by one unit.
- The sum of the squared errors SSE of the least squares regression line can be computed using a formula, without having to compute all the individual errors.

This page titled 10.4: The Least Squares Regression Line is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.5: Statistical Inferences About β1

Learning Objectives

- To learn how to construct a confidence interval for β_1, the slope of the population regression line.
- To learn how to test hypotheses regarding β_1.

The parameter β_1, the slope of the population regression line, is of primary importance in regression analysis because it gives the true rate of change in the mean $E(y)$ in response to a unit increase in the predictor variable x. For every unit increase in x the mean of the response variable y changes by β_1 units, increasing if $\beta_1 > 0$ and decreasing if $\beta_1 < 0$. We wish to construct confidence intervals for β_1 and test hypotheses about it.

Confidence Intervals for β_1

The slope $\hat{\beta}_1$ of the least squares regression line is a point estimate of β_1. A confidence interval for β_1 is given by the following formula.

Definition: $100(1-\alpha)$ Confidence Interval for the Slope β_1 of the Population Regression Line

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{s_\epsilon}{\sqrt{SS_{xx}}}$$

where $S_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ and the number of degrees of freedom is $df = n-2$.

The assumptions listed in Section 10.3 must hold.

Definition: sample standard deviation of errors

The statistic S_ε is called the sample standard deviation of errors. It estimates the standard deviation σ of the errors in the population of y-values for each fixed value of x (see Figure 10.3.1).

Example 10.5.1

Construct the 95% confidence interval for the slope β_1 of the population regression line based on the five-point sample data set

x	2	2	6	8	10
y	0	1	2	3	3

Solution

The point estimate $\hat{\beta}_1$ of β_1 was computed in Example 10.4.2 as $\hat{\beta}_1 = 0.34375$. In the same example SS_{xx} was found to be $SS_{xx} = 51.2$. The sum of the squared errors SSE was computed in Example 10.4.4 as $SSE = 0.75$. Thus

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.75}{3}} = 0.50$$

Confidence level 95% means $\alpha = 1 - 0.95 = 0.05$ so $\alpha/2 = 0.025$. From the row labeled $df = 3$ in Figure 7.1.6 we obtain $t_{0.025} = 3.182$. Therefore

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{S_\varepsilon}{\sqrt{SS_{xx}}} = 0.34375 \pm 3.182\left(\frac{0.50}{\sqrt{51.2}}\right) = 0.34375 \pm 0.2223$$

which gives the interval $(0.1215, 0.5661)$. We are 95% confident that the slope β_1 of the population regression line is between 0.1215 and 0.5661.

Example 10.5.2

Using the sample data in Table 10.4.3 construct a 90% confidence interval for the slope β_1 of the population regression line relating age and value of the automobiles of Example 10.4.3. Interpret the result in the context of the problem.

Solution

The point estimate $\hat{\beta}_1$ of β_1 was computed in Example 10.4.3, as was SS_{xx}. Their values are $\hat{\beta}_1 = -2.05$ and $SS_{xx} = 14$. The sum of the squared errors SSE was computed in Example 10.4.5 as $SSE = 28.946$. Thus

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{28.946}{8}} = 1.902169814$$

Confidence level 90% means $\alpha = 1 - 0.90 = 0.10$ so $\alpha/2 = 0.05$. From the row labeled $df = 8$ in Figure 7.1.6 we obtain $t_{0.05} = 1.860$. Therefore

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{S_\varepsilon}{\sqrt{SS_{xx}}} = -2.05 \pm 1.860\left(\frac{1.902169814}{\sqrt{14}}\right) = -2.05 \pm 0.95$$

which gives the interval $(-3.00, -1.10)$. We are 90% confident that the slope β_1 of the population regression line is between -3.00 and -1.10. In the context of the problem this means that for vehicles of this make and model between two and six years old we are 90% confident that for each additional year of age the average value of such a vehicle decreases by between $\$1,100$ and $\$3,000$.

Testing Hypotheses About β1

Hypotheses regarding β_1 can be tested using the same five-step procedures, either the critical value approach or the p-value approach, that were introduced in Section 8.1 and Section 8.3. The null hypothesis always has the form $H_0 : \beta_1 = B_0$ where B_0 is a number determined from the statement of the problem. The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \beta_1 < B_0$	Left-tailed
$H_a : \beta_1 > B_0$	Right-tailed
$H_a : \beta_1 \neq B_0$	Two-tailed

The value zero for B_0 is of particular importance since in that case the null hypothesis is $H_0 : \beta_1 = 0$, which corresponds to the situation in which x is not useful for predicting y. For if $\beta_1 = 0$ then the population regression line is horizontal, so the mean $E(y)$ is the same for every value of x and we are just as well off in ignoring x completely and approximating y by its average value. Given two variables x and y, the burden of proof is that x is useful for predicting y, not that it is not. Thus the phrase "test whether x is useful for prediction of y," or words to that effect, means to perform the test

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

Standardized Test Statistic for Hypothesis Tests Concerning the Slope β_1 of the Population Regression Line

$$T = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon/\sqrt{SS_{xx}}}$$

The test statistic has Student's t-distribution with $df = n - 2$ degrees of freedom.

The assumptions listed in Section 10.3 must hold.

Example 10.5.3

Test, at the 2% level of significance, whether the variable x is useful for predicting y based on the information in the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

Solution

We will perform the test using the critical value approach.

- **Step 1**. Since x is useful for prediction of y precisely when the slope β_1 of the population regression line is nonzero, the relevant test is

$$H_0 : \beta_1 = 0$$
$$\text{vs.}$$
$$H_a : \beta_1 \neq 0 \ @ \ \alpha = 0.02$$

- **Step 2**. The test statistic is

$$T = \frac{\hat{\beta}_1}{S_\varepsilon / \sqrt{SS_{xx}}}$$

and has Student's t-distribution with $n - 2 = 5 - 2 = 3$ degrees of freedom.

- **Step 3**. From Example 10.4.1, $\beta_1 = 0.34375$ and $SS_{xx} = 51.2$. From "Example 10.5.1", $S_\varepsilon = 0.50$. The value of the test statistic is therefore

$$T = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon / \sqrt{SS_{xx}}} = \frac{0.34375}{0.50/\sqrt{51.2}} = 4.919$$

- **Step 4**. Since the symbol in H_a is "$\neq$" this is a two-tailed test, so there are two critical values $\pm t_{\alpha/2} = \pm t_{0.01}$. Reading from the line in Figure 7.1.6 labeled $df = 3$, $t_{0.01} = 4.541$. The rejection region is

$$(-\infty, -4.541] \cup [4.541, \infty)$$

.

- **Step 5**. As shown in Figure 10.5.1 "Rejection Region and Test Statistic for" the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 2% level of significance, to conclude that the slope of the population regression line is nonzero, so that x is useful as a predictor of y.

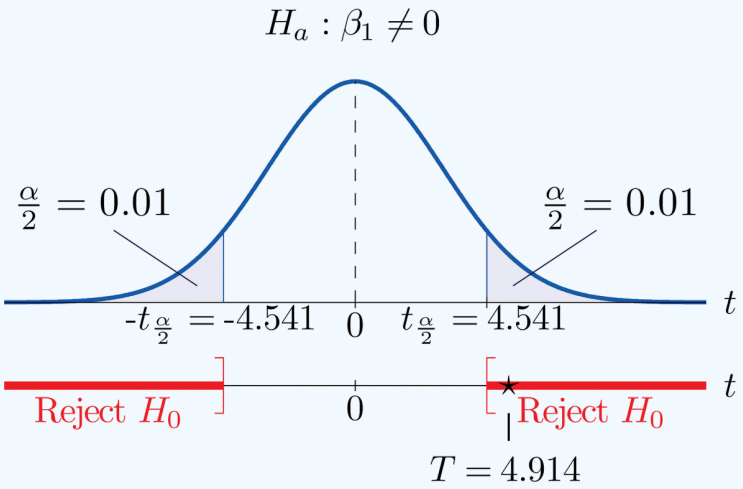


Figure 10.5.1: Rejection Region and Test Statistic for Example 10.5.3

Example 10.5.4

A car salesman claims that automobiles between two and six years old of the make and model discussed in Example 10.4.2 lose more than $\$1,100$ in value each year. Test this claim at the 5% level of significance.

Solution

We will perform the test using the critical value approach.

- **Step 1**. In terms of the variables x and y, the salesman's claim is that if x is increased by 1 unit (one additional year in age), then y decreases by more than 1.1 units (more than $\$1,100$). Thus his assertion is that the slope of the population regression line is negative, and that it is more negative than -1.1. In symbols, $\beta_1 < -1.1$. Since it contains an inequality, this has to be the alternative hypotheses. The null hypothesis has to be an equality and have the same number on the right hand side, so the relevant test is

$$H_0 : \beta_1 = -1.1$$
$$\text{vs.}$$
$$H_a : \beta_1 < -1.1 \ \ @ \ \ \alpha = 0.05$$

- **Step 2**. The test statistic is

$$T = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon / \sqrt{SS_{xx}}}$$

and has Student's t-distribution with 8 degrees of freedom.

- **Step 3**. From Example 10.4.2, $\beta_1 = -2.05$ and $SS_{xx} = 14$. From "Example 10.5.2", $S_\varepsilon = 1.902169814$. The value of the test statistic is therefore

$$T = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon / \sqrt{SS_{xx}}} = \frac{-2.05 - (-1.1)}{1.902169814/\sqrt{14}} = -1.869$$

- **Step 4**. Since the symbol in H_a is "$<$" this is a left-tailed test, so there is a single critical value $-t_{\alpha/2} = -t_{0.05}$. Reading from the line in Figure 7.1.6 labeled $df = 8$, $t_{0.05} = 1.860$. The rejection region is

$$(-\infty, -1.860]$$

.

- **Step 5**. As shown in Figure 10.5.2 "Rejection Region and Test Statistic for " the test statistic falls in the rejection region. The decision is to reject H_0. In the context of the problem our conclusion is:

  The data provide sufficient evidence, at the 5% level of significance, to conclude that vehicles of this make and model and in this age range lose more than $\$1,100$ per year in value, on average.

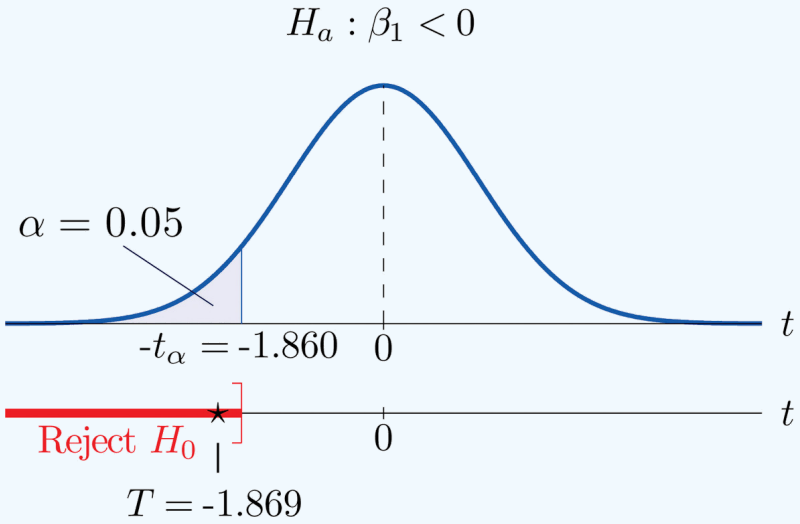


Figure 10.5.2: Rejection Region and Test Statistic for "Example 10.5.4"

Key Takeaway

- The parameter β_1, the slope of the population regression line, is of primary interest because it describes the average change in y with respect to unit increase in x.
- The statistic $\hat{\beta}_1$, the slope of the least squares regression line, is a point estimate of β_1. Confidence intervals for β_1 can be computed using a formula.
- Hypotheses regarding β_1 are tested using the same five-step procedures introduced in Chapter 8.

This page titled 10.5: Statistical Inferences About β₁ is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.6: The Coefficient of Determination

Learning Objectives

- To learn what the coefficient of determination is, how to compute it, and what it tells us about the relationship between two variables x and y.

If the scatter diagram of a set of (x, y) pairs shows neither an upward or downward trend, then the horizontal line $\hat{y} = \overline{y}$ fits it well, as illustrated in Figure 10.6.1. The lack of any upward or downward trend means that when an element of the population is selected at random, knowing the value of the measurement x for that element is not helpful in predicting the value of the measurement y.

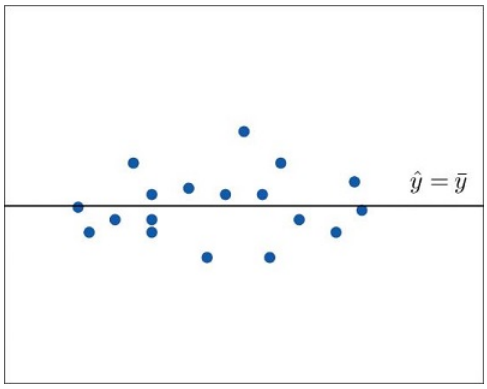


Figure 10.6.1: The line $\hat{y} = \overline{y}$ fits the scatter diagram well.

If the scatter diagram shows a linear trend upward or downward then it is useful to compute the least squares regression line

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$

and use it in predicting y. Figure 10.6.2 illustrates this. In each panel we have plotted the height and weight data of Section 10.1. This is the same scatter plot as Figure 10.6.2, with the average value line $\hat{y} = \overline{y}$ superimposed on it in the left panel and the least squares regression line imposed on it in the right panel. The errors are indicated graphically by the vertical line segments.

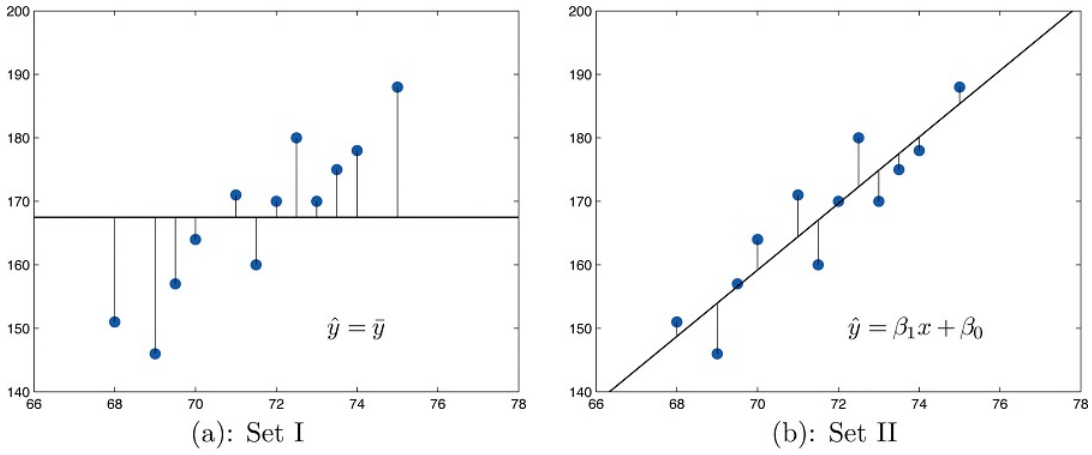


Figure 10.6.2: Same Scatter Diagram with Two Approximating Lines

The sum of the squared errors computed for the regression line, SSE, is smaller than the sum of the squared errors computed for any other line. In particular it is less than the sum of the squared errors computed using the line $\hat{y} = \overline{y}$, which sum is actually the number SS_{yy} that we have seen several times already. A measure of how useful it is to use the regression equation for prediction of y is how much smaller SSE is than SS_{yy}. In particular, the proportion of the sum of the squared errors for the line $\hat{y} = \overline{y}$ that is eliminated by going over to the least squares regression line is

$$\frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

We can think of SSE/SS_{yy} as the proportion of the variability in y that cannot be accounted for by the linear relationship between x and y, since it is still there even when x is taken into account in the best way possible (using the least squares regression line; remember that SSE is the smallest the sum of the squared errors can be for any line). Seen in this light, the coefficient of determination, the complementary proportion of the variability in y, is the proportion of the variability in all the y measurements that is accounted for by the linear relationship between x and y.

In the context of linear regression the coefficient of determination is always the square of the correlation coefficient r discussed in Section 10.2. Thus the coefficient of determination is denoted r^2, and we have two additional formulas for computing it.

Definition: coefficient of determination

The *coefficient of determination* of a collection of (x, y) pairs is the number r^2 computed by any of the following three expressions:

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \hat{\beta}_1 \frac{SS_{xy}}{SS_{yy}}$$

It measures the proportion of the variability in y that is accounted for by the linear relationship between x and y.

If the **correlation coefficient** r is already known then the **coefficient of determination** can be computed simply by squaring r, as the notation indicates, $r^2 = (r)^2$.

Example 10.6.1

The value of used vehicles of the make and model discussed in "Example 10.4.2" in Section 10.4 varies widely. The most expensive automobile in the sample in Table 10.4.3 has value $\$30,500$, which is nearly half again as much as the least expensive one, which is worth $\$20,400$. Find the proportion of the variability in value that is accounted for by the linear relationship between age and value.

Solution

The proportion of the variability in value y that is accounted for by the linear relationship between it and age x is given by the coefficient of determination, r^2. Since the correlation coefficient r was already computed in "Example 10.4.2" in Section 10.4 as

$$r = -0.819$$
$$r^2 = (-0.819)2 = 0.671$$

About 67% of the variability in the value of this vehicle can be explained by its age.

Example 10.6.2

Use each of the three formulas for the coefficient of determination to compute its value for the example of ages and values of vehicles.

Solution

In "Example 10.4.2" in Section 10.4 we computed the exact values

$$SS_{xx} = 14$$
$$SS_{xy} = -28.7$$
$$SS_{yy} = 87.781$$
$$\hat{\beta}_1 = -2.05$$

In "Example 10.4.4" in Section 10.4 we computed the exact value

$$SSE = 28.946$$

Inserting these values into the formulas in the definition, one after the other, gives

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{87.781 - 28.946}{87.781} = 0.6702475479$$

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \frac{(-28.7)^2}{(14)(87.781)} = 0.6702475479$$

$$r^2 = \hat{\beta}_1 \frac{SS_{xy}}{SS_{yy}} = -2.05\frac{-28.7}{87.781} = 0.6702475479$$

which rounds to 0.670. The discrepancy between the value here and in the previous example is because a rounded value of r from "Example 10.4.2" was used there. The actual value of r before rounding is 0.8186864772, which when squared gives the value for r^2 obtained here.

The coefficient of determination r^2 can always be computed by squaring the correlation coefficient r if it is known. Any one of the defining formulas can also be used. Typically one would make the choice based on which quantities have already been computed. What should be avoided is trying to compute r by taking the square root of r^2, if it is already known, since it is easy to make a sign error this way. To see what can go wrong, suppose $r^2 = 0.64$. Taking the square root of a positive number with any calculating device will always return a positive result. The square root of 0.64 is 0.8. However, the actual value of r might be the negative number -0.8.

Key Takeaway

- The coefficient of determination r^2 estimates the proportion of the variability in the variable y that is explained by the linear relationship between y and the variable x.
- There are several formulas for computing r^2. The choice of which one to use can be based on which quantities have already been computed so far.

This page titled 10.6: The Coefficient of Determination is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.7: Estimation and Prediction

Learning Objectives

- To learn the distinction between estimation and prediction.
- To learn the distinction between a confidence interval and a prediction interval.
- To learn how to implement formulas for computing confidence intervals and prediction intervals.

Consider the following pairs of problems, in the context of Example 10.4.2, the automobile age and value example.

Problem 1

1. Estimate the average value of all four-year-old automobiles of this make and model.
2. Construct a 95% confidence interval for the average value of all four-year-old automobiles of this make and model.

Problem 2

1. Shylock intends to buy a four-year-old automobile of this make and model next week. Predict the value of the first such automobile that he encounters.
2. Construct a 95% confidence interval for the value of the first such automobile that he encounters.

The method of solution and answer to the first question in each pair, (1a) and (2a), are the same. When we set x equal to 4 in the least squares regression equation

$$\hat{y} = -2.05x + 32.83$$

that was computed in part (c) of Example 10.4.2, the number returned,

$$\hat{y} = -2.05(4) + 32.83 = 24.63$$

which corresponds to value $\$24,630$, is an estimate of precisely the number sought in question (1a): the mean $E(y)$ of all y values when $x = 4$. Since nothing is known about the first four-year-old automobile of this make and model that Shylock will encounter, our best guess as to its value is the mean value $E(y)$ of all such automobiles, the number 24.63 or $\$24,630$, computed in the same way.

The answers to the second part of each question differ. In question (1b) we are trying to estimate a population parameter: the mean of the all the y-values in the sub-population picked out by the value $x = 4$, that is, the average value of all four-year-old automobiles. In question (2b), however, we are not trying to capture a fixed parameter, but the value of the random variable y in one trial of an experiment: examine the first four-year-old car Shylock encounters. In the first case we seek to construct a confidence interval in the same sense that we have done before. In the second case the situation is different, and the interval constructed has a different name, prediction interval. In the second case we are trying to "predict" where a the value of a random variable will take its value.

$100(1-\alpha)\%$ Confidence Interval for the Mean Value of y at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

where

- x_p is a particular value of x that lies in the range of x-values in the sample data set used to construct the least squares regression line;
- $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
- the number of degrees of freedom for $t_{\alpha/2}$ is $df = n - 2$.

The assumptions listed in Section 10.3 must hold.

The formula for the prediction interval is identical except for the presence of the number 1 underneath the square root sign. This means that the prediction interval is always wider than the confidence interval at the same confidence level and value of x. In practice the presence of the number 1 tends to make it much wider.

$100(1-\alpha)\%$ Prediction Interval for an Individual New Value of of y at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

where

- x_p is a particular value of x that lies in the range of x-values in the data set used to construct the least squares regression line;
- $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
- the number of degrees of freedom for $t_{\alpha/2}$ is $df = n - 2$.

The assumptions listed in Section 10.3 must hold.

Example 10.7.1

Using the sample data of "Example 10.4.2" in Section 10.4 , recorded in Table 10.4.3, construct a 95% confidence interval for the average value of all three-and-one-half-year-old automobiles of this make and model.

Solution

Solving this problem is merely a matter of finding the values of $\hat{y}_p,\ \alpha,\ and\ \ t_{\alpha/2}, S_\varepsilon,\ \bar{x}\ and\ \ SS_{xx}$, and inserting them into the confidence interval formula given just above. Most of these quantities are already known. From Example 10.4.2, $SS_{xx} = 14$ $and\ \ \bar{x} = 4$. From Example 10.5.2, $S\varepsilon = 1.902169814$.

From the statement of the problem $x_p = 3.5$, the value of x of interest. The value of $\hat{y_p}$ is the number given by the regression equation, which by Example 10.4.2 is $\hat{y} = -2.05x + 32.83$, when $x = x_p$, that is, when $x = 3.5$. Thus here $\hat{y} = -2.05(3.5) + 32.83 = 25.655$.

Lastly, confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha/2 = 0.025$. Since the sample size is $n = 10$, there are $n - 2 = 8$ degrees of freedom. By Figure 7.1.6, $t_{0.025} = 2.306$. Thus

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2} S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 25.655 \pm (2.306)(1.902169814)\sqrt{\frac{1}{10} + \frac{(3.5-4)^2}{14}} \\ &= 25.655 \pm 4.386403591\sqrt{0.1178571429} \\ &= 25.655 \pm 1.506\end{aligned}$$

which gives the interval $(24.149, 27.161)$.

We are 95% confident that the average value of all three-and-one-half-year-old vehicles of this make and model is between $\$24,149$ and $\$27,161$.

Example 10.7.2

Using the sample data of Example 10.4.2, recorded in Table 10.4.3, construct a 95% prediction interval for the predicted value of a randomly selected three-and-one-half-year-old automobile of this make and model.

Solution

The computations for this example are identical to those of the previous example, except that now there is the extra number 1 beneath the square root sign. Since we were careful to record the intermediate results of that computation, we have immediately that the 95% prediction interval is

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2} S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 25.655 \pm 4.386403591\sqrt{1.1178571429} \\ &= 25.655 \pm 4\end{aligned}$$

which gives the interval $(21.017, 30.293)$.

We are 95% confident that the value of a randomly selected three-and-one-half-year-old vehicle of this make and model is between $\$21,017$ and $\$30,293$.

Note what an enormous difference the presence of the extra number 1 under the square root sign made. The prediction interval is about two-and-one-half times wider than the confidence interval at the same level of confidence.

KeyTakaways

- A confidence interval is used to estimate the mean value of y in the sub-population determined by the condition that x have some specific value x_p.
- The prediction interval is used to predict the value that the random variable y will take when x has some specific value x_p.

This page titled 10.7: Estimation and Prediction is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.8: A Complete Example

Learning Objectives

- To see a complete linear correlation and regression analysis, in a practical setting, as a cohesive whole

In the preceding sections numerous concepts were introduced and illustrated, but the analysis was broken into disjoint pieces by sections. In this section we will go through a complete example of the use of correlation and regression analysis of data from start to finish, touching on all the topics of this chapter in sequence.

In general educators are convinced that, all other factors being equal, class attendance has a significant bearing on course performance. To investigate the relationship between attendance and performance, an education researcher selects for study a multiple section introductory statistics course at a large university. Instructors in the course agree to keep an accurate record of attendance throughout one semester. At the end of the semester 26 students are selected a random. For each student in the sample two measurements are taken: x, the number of days the student was absent, and y, the student's score on the common final exam in the course. The data are summarized in Table 10.8.1.

Table 10.8.1: Absence and Score Data

Absences	Score	Absences	Score
x	y	x	y
2	76	4	41
7	29	5	63
2	96	4	88
7	63	0	98
2	79	1	99
7	71	0	89
0	88	1	96
0	92	3	90
6	55	1	90
6	70	3	68
2	80	1	84
2	75	3	80
1	63	1	78

A scatter plot of the data is given in Figure 10.8.1. There is a downward trend in the plot which indicates that on average students with more absences tend to do worse on the final examination.

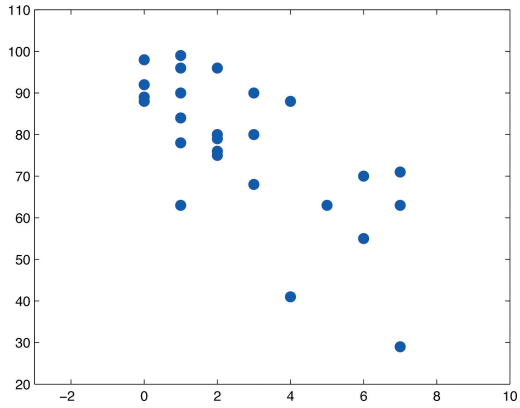


Figure 10.8.1: Plot of the Absence and Exam Score Pairs

The trend observed in Figure 10.8.1 as well as the fairly constant width of the apparent band of points in the plot makes it reasonable to assume a relationship between x and y of the form

$$y = \beta_1 x + \beta_0 + \varepsilon$$

where β_1 and β_0 are unknown parameters and ε is a normal random variable with mean zero and unknown standard deviation $\sum$. Note carefully that this model is being proposed for the population of all students taking this course, not just those taking it this semester, and certainly not just those in the sample. The numbers β_1, β_0, and $\sum$ are parameters relating to this large population.

First we perform preliminary computations that will be needed later. The data are processed in Table 10.8.2.

Table 10.8.2: Processed Absence and Score Data

x	y	x^2	xy	y^2	x	y	x^2	xy	y^2
2	76	4	152	5776	4	41	16	164	1681
7	29	49	203	841	5	63	25	315	3969
2	96	4	192	9216	4	88	16	352	7744
7	63	49	441	3969	0	98	0	0	9604
2	79	4	158	6241	1	99	1	99	9801
7	71	49	497	5041	0	89	0	0	7921
0	88	0	0	7744	1	96	1	96	9216
0	92	0	0	8464	3	90	9	270	8100
6	55	36	330	3025	1	90	1	90	8100
6	70	36	420	4900	3	68	9	204	4624
2	80	4	160	6400	1	84	1	84	7056
2	75	4	150	5625	3	80	9	240	6400
1	63	1	63	3969	1	78	1	78	6084

Adding up the numbers in each column in Table 10.8.2 gives

$$\sum x = 71, \sum y = 2001, \sum x^2 = 329, \sum xy = 4758, \text{ and } \sum y^2 = 161511.$$

Then

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 329 - \frac{1}{26}(71)^2 = 135.1153846$$

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 4758 - \frac{1}{26}(71)(2001) = -706.2692308$$

$$SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 161511 - \frac{1}{26}(2001)^2 = 7510.961538$$

$$SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2 = 161511 - \frac{1}{26}(2001)^2 = 7510.961538$$

and

$$\bar{x} = \frac{\sum x}{n} = \frac{71}{26} = 2.730769231 \;\; and \;\; \bar{y} = \frac{\sum y}{n} = \frac{2001}{26} = 76.96153846$$

We begin the actual modelling by finding the least squares regression line, the line that best fits the data. Its slope and y-intercept are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-706.2692308}{135.1153846} = -5.227156278$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 76.96153846 - (-5.227156278)(2.730769231) = 91.23569553$$

Rounding these numbers to two decimal places, the least squares regression line for these data is

$$\hat{y} = -5.23x + 91.24$$

The goodness of fit of this line to the scatter plot, the sum of its squared errors, is

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 7510.961538 - (-5.227156278)(-706.2692308) = 3819.181894$$

This number is not particularly informative in itself, but we use it to compute the important statistic

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{3819.181894}{24}} = 12.11988495$$

The statistic S_ε estimates the standard deviation $\sum$ of the normal random variable ε in the model. Its meaning is that among all students with the same number of absences, the standard deviation of their scores on the final exam is about 12.1 points. Such a large value on a 100-point exam means that the final exam scores of each sub-population of students, based on the number of absences, are highly variable.

The size and sign of the slope $\beta\hat{}1 = -5.23$ indicate that, for every class missed, students tend to score about 5.23 fewer points lower on the final exam on average. Similarly for every two classes missed students tend to score on average $2 \times 5.23 = 10.46$ fewer points on the final exam, or about a letter grade worse on average.

Since 0 is in the range of x-values in the data set, the y-intercept also has meaning in this problem. It is an estimate of the average grade on the final exam of all students who have perfect attendance. The predicted average of such students is $\hat{\beta}_0 = 91.24$.

Before we use the regression equation further, or perform other analyses, it would be a good idea to examine the utility of the linear regression model. We can do this in two ways: 1) by computing the correlation coefficient r to see how strongly the number of absences x and the score y on the final exam are correlated, and 2) by testing the null hypothesis $H_0 : \hat{\beta}_1 = 0$ (the slope of the population regression line is zero, so x is not a good predictor of y) against the natural alternative $H_a : \hat{\beta}_1 < 0$ (the slope of the population regression line is negative, so final exam scores y go down as absences x go up).

The correlation coefficient r is

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{-706.2692308}{\sqrt{(135.1153846)(7510.961538)}} = -0.7010840977$$

a moderate negative correlation.

Turning to the test of hypotheses, let us test at the commonly used 5% level of significance. The test is

$$H_0 : \beta_1 = 0$$
$$vs.$$
$$H_a : \beta_1 < 0 \;\; @ \;\; \alpha = 0.05$$

From Figure 7.1.6, with $df = 26 - 2 = 24$ degrees of freedom $t_{0.05} = 1.711$, so the rejection region is $(-\infty, -1.711]$. The value of the standardized test statistic is

$$t = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon / \sqrt{SS_{xx}}} = \frac{-5.227156278 - 0}{12.11988495 / \sqrt{135.1153846}} = -5.013$$

which falls in the rejection region. We reject H_0 in favor of H_a. The data provide sufficient evidence, at the 5% level of significance, to conclude that β_1 is negative, meaning that as the number of absences increases average score on the final exam decreases.

As already noted, the value $\beta_1 = -5.23$ gives a point estimate of how much one additional absence is reflected in the average score on the final exam. For each additional absence the average drops by about 5.23 points. We can widen this point estimate to a confidence interval for β_1. At the 95% confidence level, from Figure 7.1.6 with $df = 26 - 2 = 24$ degrees of freedom, $t_{\alpha/2} = t_{0.025} = 2.064$. The 95% confidence interval for β_1 based on our sample data is

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{S_\varepsilon}{\sqrt{SS_{xx}}} = -5.23 \pm 2.064\frac{12.11988495}{\sqrt{135.1153846}} = -5.23 \pm 2.15$$

or $(-7.38, -3.08)$. We are 95% confident that, among all students who ever take this course, for each additional class missed the average score on the final exam goes down by between 3.08 and 7.38 points.

If we restrict attention to the sub-population of all students who have exactly five absences, say, then using the least squares regression equation $\hat{y} = -5.23x + 91.24$ we estimate that the average score on the final exam for those students is

$$\hat{y} = -5.23(5) + 91.24 = 65.09$$

This is also our best guess as to the score on the final exam of any particular student who is absent five times. A 95% confidence interval for the average score on the final exam for all students with five absences is

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2}S_\varepsilon\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 65.09 \pm (2.064)(12.11988495)\sqrt{\frac{1}{26} + \frac{(5 - 2.730769231)^2}{135.1153846}} \\ &= 65.09 \pm 25.01544254\sqrt{0.0765727299} \\ &= 65.09 \pm 6.92\end{aligned}$$

which is the interval $(58.17, 72.01)$. This confidence interval suggests that the true mean score on the final exam for all students who are absent from class exactly five times during the semester is likely to be between 58.17 and 72.01.

If a particular student misses exactly five classes during the semester, his score on the final exam is predicted with 95% confidence to be in the interval

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2}S_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 65.09 \pm 25.01544254\sqrt{1.0765727299} \\ &= 65.09 \pm 25.96\end{aligned}$$

which is the interval $(39.13, 91.05)$. This prediction interval suggests that this individual student's final exam score is likely to be between 39.13 and 91.05. Whereas the 95% confidence interval for the average score of all student with five absences gave real information, this interval is so wide that it says practically nothing about what the individual student's final exam score might be. This is an example of the dramatic effect that the presence of the extra summand 1 under the square sign in the prediction interval can have.

Finally, the proportion of the variability in the scores of students on the final exam that is explained by the linear relationship between that score and the number of absences is estimated by the coefficient of determination, r^2. Since we have already computed r above we easily find that

$$r^2 = (-0.7010840977)^2 = 0.491518912$$

or about 49%. Thus although there is a significant correlation between attendance and performance on the final exam, and we can estimate with fair accuracy the average score of students who miss a certain number of classes, nevertheless less than half the total variation of the exam scores in the sample is explained by the number of absences. This should not come as a surprise, since there are many factors besides attendance that bear on student performance on exams.

Key Takeaway

- It is a good idea to attend class.Edit section

This page titled 10.8: A Complete Example is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.9: Formula List

Learning Objectives

- Listing of all formulas used throughout the chapter.

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 \quad SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) \quad SS_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

Correlation coefficient:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Least squares regression equation (equation of the least squares regression line):

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \text{ where } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sum of the squared errors for the least squares regression line:

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

Sample standard deviation of errors:

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

$100(1-\alpha)\%$ confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{S_\varepsilon}{\sqrt{SS_{xx}}} \quad (df = n-2)$$

Standardized test statistic for hypothesis tests concerning β_1:

$$T = \frac{\hat{\beta}_1 - B_0}{S_\varepsilon/\sqrt{SS_{xx}}} \quad (df = n-2)$$

Coefficient of determination:

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \hat{\beta}_1\frac{SS_{xy}}{SS_{yy}}$$

$100(1-\alpha)\%$ confidence interval for the mean value of y at $x = x_p$:

$$\hat{y}_p \pm t_{\alpha/2}S_\varepsilon\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \quad (df = n-2)$$

$100(1-\alpha)\%$ prediction interval for an individual new value of y at $x = x_p$:

$$\hat{y}_p \pm t_{\alpha/2}S_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} \quad (df = n-2)$$

This page titled 10.9: Formula List is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

10.E: Correlation and Regression (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

10.1 Linear Relationships Between Variables

Basic

1. A line has equation $y = 0.5x + 2$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
2. A line has equation $y = x - 0.5$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
3. A line has equation $y = -2x + 4$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
4. A line has equation $y = -1.5x + 1$.
   a. Pick five distinct x-values, use the equation to compute the corresponding y-values, and plot the five points obtained.
   b. Give the value of the slope of the line; give the value of the y-intercept.
5. Based on the information given about a line, determine how y will change (increase, decrease, or stay the same) when x is increased, and explain. In some cases it might be impossible to tell from the information given.
   a. The slope is positive.
   b. The y-intercept is positive.
   c. The slope is zero.
6. Based on the information given about a line, determine how y will change (increase, decrease, or stay the same) when x is increased, and explain. In some cases it might be impossible to tell from the information given.
   a. The y-intercept is negative.
   b. The y-intercept is zero.
   c. The slope is negative.
7. A data set consists of eight (x, y) pairs of numbers:

$$\begin{array}{cccc} (0,12) & (4,16) & (8,22) & (15,28) \\ (2,15) & (5,14) & (13,24) & (20,30) \end{array} \tag{10.E.1}$$

   a. Plot the data in a scatter diagram.
   b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
   c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.
8. A data set consists of ten (x, y) pairs of numbers:

$$\begin{array}{ccccc} (3,20) & (6,9) & (11,0) & (14,1) & (18,9) \\ (5,13) & (8,4) & (12,0) & (17,6) & (20,16) \end{array} \tag{10.E.2}$$

   a. Plot the data in a scatter diagram.
   b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
   c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.
9. A data set consists of nine (x, y) pairs of numbers:

$$\begin{array}{ccccc} (8,16) & (10,4) & (12,0) & (14,4) & (16,16) \\ (9,9) & (11,1) & (13,1) & (15,9) & \end{array} \tag{10.E.3}$$

   a. Plot the data in a scatter diagram.
   b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
   c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

10. A data set consists of five (x, y) pairs of numbers:

$$(0, 1) \quad (2, 5) \quad (3, 7) \quad (5, 11) \quad (8, 17) \tag{10.E.4}$$

    a. Plot the data in a scatter diagram.
    b. Based on the plot, explain whether the relationship between x and y appears to be deterministic or to involve randomness.
    c. Based on the plot, explain whether the relationship between x and y appears to be linear or not linear.

Applications

11. At $60^{\circ}F$ a particular blend of automotive gasoline weights 6.17 lb/gal. The weight y of gasoline on a tank truck that is loaded with x gallons of gasoline is given by the linear equation

$$y = 6.17x \tag{10.E.5}$$

    a. Explain whether the relationship between the weight y and the amount x of gasoline is deterministic or contains an element of randomness.
    b. Predict the weight of gasoline on a tank truck that has just been loaded with $6,750$ gallons of gasoline.

12. The rate for renting a motor scooter for one day at a beach resort area is $\$25$ plus 30 cents for each mile the scooter is driven. The total cost y in dollars for renting a scooter and driving it x miles is

$$y = 0.30x + 25 \tag{10.E.6}$$

    a. Explain whether the relationship between the cost y of renting the scooter for a day and the distance x that the scooter is driven that day is deterministic or contains an element of randomness.
    b. A person intends to rent a scooter one day for a trip to an attraction 17 miles away. Assuming that the total distance the scooter is driven is 34 miles, predict the cost of the rental.

13. The pricing schedule for labor on a service call by an elevator repair company is $\$150$ plus $\$50$ per hour on site.
    a. Write down the linear equation that relates the labor cost y to the number of hours x that the repairman is on site.
    b. Calculate the labor cost for a service call that lasts 2.5 hours.

14. The cost of a telephone call made through a leased line service is 2.5 cents per minute.
    a. Write down the linear equation that relates the cost y (in cents) of a call to its length x.
    b. Calculate the cost of a call that lasts 23 minutes.

Large Data Set Exercises

Large Data Sets not available

15. Large $\text{Data Set } 1$ lists the SAT scores and GPAs of $1,000$ students. Plot the scatter diagram with SAT score as the independent variable (x) and GPA as the dependent variable (y). Comment on the appearance and strength of any linear trend.
16. Large $\text{Data Set } 12$ lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Plot the scatter diagram with golf score using the original clubs as the independent variable (x) and golf score using the new clubs as the dependent variable (y). Comment on the appearance and strength of any linear trend.
17. Large $\text{Data Set } 13$ records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Plot the scatter diagram with the number of bidders at the auction as the independent variable (x) and the sales price as the dependent variable (y). Comment on the appearance and strength of any linear trend.

Answers

1. a. Answers vary.
   b. Slope $m = 0.5$; y-intercept $b = 2$.
2.
3.
    a. Answers vary.
    b. Slope $m = -2$; y-intercept $b = 4$.
4.
5.

a. y increases.
b. Impossible to tell.
c. y does not change.

6.

7. a. Scatter diagram needed.
b. Involves randomness.
c. Linear.

8.

9. a. Scatter diagram needed.
b. Deterministic.
c. Not linear.

10.

11. a. Deterministic.
b. $41,647.5$ pounds.

12.

13. a. $y = 50x + 150$.
b. $\$275$.

14.

15. There appears to a hint of some positive correlation.

16.

17. There appears to be clear positive correlation.

10.2 The Linear Correlation Coefficient

Basic

With the exception of the exercises at the end of Section 10.3, the first Basic exercise in each of the following sections through Section 10.7 uses the data from the first exercise here, the second Basic exercise uses the data from the second exercise here, and so on, and similarly for the Application exercises. Save your computations done on these exercises so that you do not need to repeat them later.

1. For the sample data

x	0	1	3	5	8
y	2	4	6	5	9

(10.E.7)

a. Draw the scatter plot.
b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

2. For the sample data

x	0	2	3	6	9
y	0	3	3	4	8

(10.E.8)

a. Draw the scatter plot.
b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

3. For the sample data

x	1	3	4	6	8
y	4	1	3	-1	0

(10.E.9)

a. Draw the scatter plot.
b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

4. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 2 & 4 & 7 & 9 \\ \hline y & 5 & 5 & 6 & -3 & 0 \end{array} \tag{10.E.10}$$

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

5. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 1 & 3 & 4 & 5 \\ \hline y & 2 & 1 & 5 & 3 & 4 \end{array} \tag{10.E.11}$$

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

6. For the sample data

$$\begin{array}{c|ccccc} x & 1 & 3 & 5 & 5 & 8 \\ \hline y & 5 & -2 & 2 & -1 & -3 \end{array} \tag{10.E.12}$$

   a. Draw the scatter plot.
   b. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
   c. Compute the linear correlation coefficient and compare its sign to your answer to part (b).

7. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$\begin{gathered} n = 5 \quad \sum x = 25 \quad \sum x^2 = 165 \\ \sum y = 24 \quad \sum y^2 = 134 \quad \sum xy = 144 \\ 1 \le x \le 9 \end{gathered} \tag{10.E.13}$$

8. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$\begin{gathered} n = 5 \quad \sum x = 31 \quad \sum x^2 = 253 \\ \sum y = 18 \quad \sum y^2 = 90 \quad \sum xy = 148 \\ 2 \le x \le 12 \end{gathered} \tag{10.E.14}$$

9. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$\begin{gathered} n = 10 \quad \sum x = 0 \quad \sum x^2 = 60 \\ \sum y = 24 \quad \sum y^2 = 234 \quad \sum xy = -87 \\ -4 \le x \le 4 \end{gathered} \tag{10.E.15}$$

10. Compute the linear correlation coefficient for the sample data summarized by the following information:

$$\begin{gathered} n = 10 \quad \sum x = -3 \quad \sum x^2 = 263 \\ \sum y = 55 \quad \sum y^2 = 917 \quad \sum xy = -355 \\ -10 \le x \le 10 \end{gathered} \tag{10.E.16}$$

Applications

11. The age x in months and vocabulary y were measured for six children, with the results shown in the table.

$$\begin{array}{c|cccccc} x & 13 & 14 & 15 & 16 & 16 & 18 \\ \hline y & 8 & 10 & 15 & 20 & 27 & 30 \end{array} \tag{10.E.17}$$

    Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

12. The curb weight x in hundreds of pounds and braking distance y in feet, at 50 miles per hour on dry pavement, were measured for five vehicles, with the results shown in the table.

$$\begin{array}{c|ccccc} x & 25 & 27.5 & 32.5 & 35 & 45 \end{array}$$

y	105	125	140	140	150

(10.E.18)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

13. The age x and resting heart rate y were measured for ten men, with the results shown in the table.

x	20	23	30	37	35
y	72	71	73	74	74
x	45	51	55	60	63
y	73	72	79	75	77

(10.E.19)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

14. The wind speed x in miles per hour and wave height y in feet were measured under various conditions on an enclosed deep water sea, with the results shown in the table,

x	0	0	2	7	7
y	2.0	0.0	0.3	0.7	3.3
x	9	13	20	22	31
y	4.9	4.9	3.0	6.9	5.9

(10.E.20)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

15. The advertising expenditure x and sales y in thousands of dollars for a small retail business in its first eight years in operation are shown in the table.

x	1.4	1.6	1.6	2.0
y	180	184	190	220
x	2.0	2.2	2.4	2.6
y	186	215	205	240

(10.E.21)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

16. The height x at age 2 and y at age 20, both in inches, for ten women are tabulated in the table.

x	31.3	31.7	32.5	33.5	34.4
y	60.7	61.0	63.1	64.2	65.9
x	35.2	35.8	32.7	33.6	34.8
y	68.2	67.6	62.3	64.9	66.8

(10.E.22)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

17. The course average x just before a final exam and the score y on the final exam were recorded for 15 randomly selected students in a large physics class, with the results shown in the table.

x	69.3	87.7	50.5	51.9	82.7
y	56	89	55	49	61
x	70.5	72.4	91.7	83.3	86.5
y	66	72	83	73	82
x	79.3	78.5	75.7	52.3	62.2
y	92	80	64	18	76

(10.E.23)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

18. The table shows the acres x of corn planted and acres y of corn harvested, in millions of acres, in a particular country in ten successive years.

x	75.7	78.9	78.6	80.9	81.8
y	68.8	69.3	70.9	73.6	75.1
x	78.3	93.5	85.9	86.4	88.2
y	70.6	86.5	78.6	79.5	81.4

(10.E.24)

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

19. Fifty male subjects drank a measured amount x (in ounces) of a medication and the concentration y (in percent) in their blood of the active ingredient was measured 30 minutes later. The sample data are summarized by the following information.

$$n = 50 \quad \sum x = 112.5 \quad \sum y = 4.83 \quad \sum xy = 15.255 \quad 0 \le x \le 4.5 \quad \sum x^2 = 356.25 \quad \sum y^2 = 0.667 \tag{10.E.25}$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

20. In an effort to produce a formula for estimating the age of large free-standing oak trees non-invasively, the girth x (in inches) five feet off the ground of 15 such trees of known age y (in years) was measured. The sample data are summarized by the following information.

$$n = 15 \quad \sum x = 3368 \quad \sum y = 6496 \quad \sum xy = 1,933,219 \quad 74 \le x \le 395 \quad \sum x^2 = 917,780 \quad \sum y^2 = 4,260,666 \tag{10.E.26}$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

21. Construction standards specify the strength of concrete 28 days after it is poured. For 30 samples of various types of concrete the strength x after 3 days and the strength y after 28 days (both in hundreds of pounds per square inch) were measured. The sample data are summarized by the following information.

$$n = 30 \quad \sum x = 501.6 \quad \sum y = 1338.8 \quad \sum xy = 23,246.55 \quad 11 \le x \le 22 \quad \sum x^2 = 8724.74 \quad \sum y^2 = 61,980.14 \tag{10.E.27}$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

22. Power-generating facilities used forecasts of temperature to forecast energy demand. The average temperature x (degrees Fahrenheit) and the day's energy demand y (million watt-hours) were recorded on 40 randomly selected winter days in the region served by a power company. The sample data are summarized by the following information.

$$n = 40 \quad \sum x = 2000 \quad \sum y = 2969 \quad \sum xy = 143,042 \quad 40 \le x \le 60 \quad \sum x^2 = 101,340 \quad \sum y^2 = 243,027 \tag{10.E.28}$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

Additional Exercises

23. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.
    a. the number x of pages in a book and the age y of the author
    b. the number x of pages in a book and the age y of the intended reader
    c. the weight x of an automobile and the fuel economy y in miles per gallon

d. the weight x of an automobile and the reading y on its odometer
e. the amount x of a sedative a person took an hour ago and the time y it takes him to respond to a stimulus

24. In each case state whether you expect the two variables x and y indicated to have positive, negative, or zero correlation.
    a. the length x of time an emergency flare will burn and the length y of time the match used to light it burned
    b. the average length x of time that calls to a retail call center are on hold one day and the number y of calls received that day
    c. the length x of a regularly scheduled commercial flight between two cities and the headwind y encountered by the aircraft
    d. the value x of a house and the its size y in square feet
    e. the average temperature x on a winter day and the energy consumption y of the furnace
25. Changing the units of measurement on two variables x and y should not change the linear correlation coefficient. Moreover, most change of units amount to simply multiplying one unit by the other (for example, 1 foot = 12 inches). Multiply each x value in the table in Exercise 1 by two and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.
26. Refer to the previous exercise. Multiply each x value in the table in Exercise 2 by two, multiply each y value by three, and compute the linear correlation coefficient for the new data set. Compare the new value of r to the one for the original data.
27. Reversing the roles of x and y in the data set of Exercise 1 produces the data set

x	2	4	6	5	9
y	0	1	3	5	8

(10.E.29)

    Compute the linear correlation coefficient of the new set of data and compare it to what you got in Exercise 1.
28. In the context of the previous problem, look at the formula for r and see if you can tell why what you observed there must be true for every data set.

Large Data Set Exercises

Large Data Sets not available

29. Large $\text{Data Set } 1$ lists the SAT scores and GPAs of $1,000$ students. Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the first large data set problem for Section 10.1.
30. Large $\text{Data Set } 12$ lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the second large data set problem for Section 10.1.
31. Large $\text{Data Set } 13$ records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Compute the linear correlation coefficient r. Compare its value to your comments on the appearance and strength of any linear trend in the scatter diagram that you constructed in the third large data set problem for Section 10.1.

Answers

1. $r = 0.921$
2.
3. $r = -0.794$
4.
5. $r = 0.707$
6.
7. 0.875
8.
9. -0.846
10.
11. 0.948
12.
13. 0.709
14.
15. 0.832

16.
17. 0.751
18.
19. 0.965
20.
21. 0.992
22.
23.
    a. zero
    b. positive
    c. negative
    d. zero
    e. positive
24.
25. same value
26.
27. same value
28.
29. $r = 0.4601$
30.
31. $r = 0.9002$

10.3 Modelling Linear Relationships with Randomness Present

Basic

1. State the three assumptions that are the basis for the Simple Linear Regression Model.
2. The Simple Linear Regression Model is summarized by the equation

$$y = \beta_1 x + \beta_0 + \varepsilon \tag{10.E.30}$$

   Identify the deterministic part and the random part.
3. Is the number β_1 in the equation $y = \beta_1 x + \beta_0$ a statistic or a population parameter? Explain.
4. Is the number σ in the Simple Linear Regression Model a statistic or a population parameter? Explain.
5. Describe what to look for in a scatter diagram in order to check that the assumptions of the Simple Linear Regression Model are true.
6. True or false: the assumptions of the Simple Linear Regression Model must hold exactly in order for the procedures and analysis developed in this chapter to be useful.

Answers

1. i. The mean of y is linearly related to x.
   ii. For each given x, y is a normal random variable with mean $\beta_1 x + \beta_0$ and a standard deviation σ.
   iii. All the observations of y in the sample are independent.
2.
3. β_1 is a population parameter.
4.
5. A linear trend.

10.4 The Least Squares Regression Line

Basic

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2.

1. Compute the least squares regression line for the data in Exercise 1 of Section 10.2.
2. Compute the least squares regression line for the data in Exercise 2 of Section 10.2.

3. Compute the least squares regression line for the data in Exercise 3 of Section 10.2.
4. Compute the least squares regression line for the data in Exercise 4 of Section 10.2.
5. For the data in Exercise 5 of Section 10.2
   a. Compute the least squares regression line.
   b. Compute the sum of the squared errors SSE using the definition $\sum (y - \hat{y})^2$.
   c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.
6. For the data in Exercise 6 of Section 10.2
   a. Compute the least squares regression line.
   b. Compute the sum of the squared errors SSE using the definition $\sum (y - \hat{y})^2$.
   c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.
7. Compute the least squares regression line for the data in Exercise 7 of Section 10.2.
8. Compute the least squares regression line for the data in Exercise 8 of Section 10.2.
9. For the data in Exercise 9 of Section 10.2
   a. Compute the least squares regression line.
   b. Can you compute the sum of the squared errors SSE using the definition $\sum (y - \hat{y})^2$? Explain.
   c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.
10. For the data in Exercise 10 of Section 10.2
    a. Compute the least squares regression line.
    b. Can you compute the sum of the squared errors SSE using the definition $\sum (y - \hat{y})^2$? Explain.
    c. Compute the sum of the squared errors SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.

Applications

11. For the data in Exercise 11 of Section 10.2
    a. Compute the least squares regression line.
    b. On average, how many new words does a child from 13 to 18 months old learn each month? Explain.
    c. Estimate the average vocabulary of all 16-month-old children.
12. For the data in Exercise 12 of Section 10.2
    a. Compute the least squares regression line.
    b. On average, how many additional feet are added to the braking distance for each additional 100 pounds of weight? Explain.
    c. Estimate the average braking distance of all cars weighing 3,000 pounds.
13. For the data in Exercise 13 of Section 10.2
    a. Compute the least squares regression line.
    b. Estimate the average resting heart rate of all 40-year-old men.
    c. Estimate the average resting heart rate of all newborn baby boys. Comment on the validity of the estimate.
14. For the data in Exercise 14 of Section 10.2
    a. Compute the least squares regression line.
    b. Estimate the average wave height when the wind is blowing at 10 miles per hour.
    c. Estimate the average wave height when there is no wind blowing. Comment on the validity of the estimate.
15. For the data in Exercise 15 of Section 10.2
    a. Compute the least squares regression line.
    b. On average, for each additional thousand dollars spent on advertising, how does revenue change? Explain.
    c. Estimate the revenue if $\$2,500$ is spent on advertising next year.
16. For the data in Exercise 16 of Section 10.2
    a. Compute the least squares regression line.
    b. On average, for each additional inch of height of two-year-old girl, what is the change in the adult height? Explain.
    c. Predict the adult height of a two-year-old girl who is 33 inches tall.
17. For the data in Exercise 17 of Section 10.2
    a. Compute the least squares regression line.
    b. Compute SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.

c. Estimate the average final exam score of all students whose course average just before the exam is 85.

18. For the data in Exercise 18 of Section 10.2
    a. Compute the least squares regression line.
    b. Compute SSE using the formula $SSE = SS_{yy} - \widehat{\beta_1} SS_{xy}$.
    c. Estimate the number of acres that would be harvested if 90 million acres of corn were planted.
19. For the data in Exercise 19 of Section 10.2
    a. Compute the least squares regression line.
    b. Interpret the value of the slope of the least squares regression line in the context of the problem.
    c. Estimate the average concentration of the active ingredient in the blood in men after consuming 1 ounce of the medication.
20. For the data in Exercise 20 of Section 10.2
    a. Compute the least squares regression line.
    b. Interpret the value of the slope of the least squares regression line in the context of the problem.
    c. Estimate the age of an oak tree whose girth five feet off the ground is 92 inches.
21. For the data in Exercise 21 of Section 10.2
    a. Compute the least squares regression line.
    b. The 28-day strength of concrete used on a certain job must be at least $3,200$ psi. If the 3-day strength is $1,300$ psi, would we anticipate that the concrete will be sufficiently strong on the 28^{th} day? Explain fully.
22. For the data in Exercise 22 of Section 10.2
    a. Compute the least squares regression line.
    b. If the power facility is called upon to provide more than 95 million watt-hours tomorrow then energy will have to be purchased from elsewhere at a premium. The forecast is for an average temperature of 42 degrees. Should the company plan on purchasing power at a premium?

Additional Exercises

23. Verify that no matter what the data are, the least squares regression line always passes through the point with coordinates $(\bar{x}, \bar{y})$. Hint: Find the predicted value of y when $x = \bar{x}$.
24. In Exercise 1 you computed the least squares regression line for the data in Exercise 1 of Section 10.2.
    a. Reverse the roles of *x* and *y* and compute the least squares regression line for the new data set

$$\begin{array}{c|ccccc} x & 2 & 4 & 6 & 5 & 9 \\ \hline y & 0 & 1 & 3 & 5 & 8 \end{array} \tag{10.E.31}$$

    b. Interchanging *x* and *y* corresponds geometrically to reflecting the scatter plot in a 45-degree line. Reflecting the regression line for the original data the same way gives a line with the equation $\bar{y} = 1.346x - 3.600$. Is this the equation that you got in part (a)? Can you figure out why not? Hint: Think about how *x* and *y* are treated differently geometrically in the computation of the goodness of fit.
    c. Compute SSE for each line and see if they fit the same, or if one fits the data better than the other.

Large Data Set Exercises

Large Data Sets not available

25. Large Data Set 1 lists the SAT scores and GPAs of $1,000$ students.
    a. Compute the least squares regression line with SAT score as the independent variable (x) and GPA as the dependent variable (y).
    b. Interpret the meaning of the slope $\widehat{\beta_1}$ of regression line in the context of problem.
    c. Compute SSE the measure of the goodness of fit of the regression line to the sample data.
    d. Estimate the GPA of a student whose SAT score is 1350.
26. Large Data Set 12 lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).
    a. Compute the least squares regression line with scores using the original clubs as the independent variable (x) and scores using the new clubs as the dependent variable (y).
    b. Interpret the meaning of the slope $\widehat{\beta_1}$ of regression line in the context of problem.

c. Compute SSE the measure of the goodness of fit of the regression line to the sample data.
d. Estimate the score with the new clubs of a golfer whose score with the old clubs is 73.

27. Large Data Set 13 records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.
    a. Compute the least squares regression line with the number of bidders present at the auction as the independent variable (x) and sales price as the dependent variable (y).
    b. Interpret the meaning of the slope $\widehat{\beta_1}$ of regression line in the context of problem.
    c. Compute SSE the measure of the goodness of fit of the regression line to the sample data.
    d. Estimate the sales price of a clock at an auction at which the number of bidders is seven.

Answers

1. $\hat{y} = 0.743x + 2.675$
2.
3. $\hat{y} = -0.610x + 4.082$
4.
5. $\hat{y} = 0.625x + 1.25,\ SSE = 5$
6.
7. $\hat{y} = 0.6x + 1.8$
8.
9. $\hat{y} = -1.45x + 2.4,\ SSE = 50.25$ (cannot use the definition to compute)
10.
11. a. $\hat{y} = 4.848x - 56$
    b. 4.8
    c. 21.6
12.
13. a. $\hat{y} = 0.114x + 69.222$
    b. 73.8
    c. 69.2, invalid extrapolation
14.
15. a. $\hat{y} = 42.024x + 119.502$
    b. increases by $\$42,024$
    c. $\$224,562$
16.
17. a. $\hat{y} = 1.045x - 8.527$
    b. 2151.93367
    c. 80.3
18.
19. a. $\hat{y} = 0.043x + 0.001$
    b. For each additional ounce of medication consumed blood concentration of the active ingredient increases by 0.043%
    c. 0.044%
20.
21. a. $\hat{y} = 2.550x + 1.993$
    b. Predicted 28-day strength is $3,514$ psi; sufficiently strong
22.
23.
24.
25. a. $\hat{y} = 0.0016x + 0.022$
    b. On average, every 100 point increase in SAT score adds 0.16 point to the GPA.
    c. $SSE = 432.10$
    d. $\hat{y} = 2.182$
26.

27. a. $\hat{y} = 116.62x + 6955.1$
    b. On average, every 1 additional bidder at an auction raises the price by 116.62 dollars.
    c. $SSE = 1850314.08$
    d. $\hat{y} = 7771.44$

10.5 Statistical Inferences About β1

Basic

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2 and Section 10.4.

1. Construct the 95% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 1 of Section 10.2.
2. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 2 of Section 10.2.
3. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 3 of Section 10.2.
4. Construct the 99% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 4 of Section 10.2.
5. For the data in Exercise 5 of Section 10.2 test, at the 10% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).
6. For the data in Exercise 6 of Section 10.2 test, at the 5% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).
7. Construct the 90% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 7 of Section 10.2.
8. Construct the 95% confidence interval for the slope β_1 of the population regression line based on the sample data set of Exercise 8 of Section 10.2.
9. For the data in Exercise 9 of Section 10.2 test, at the 1% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).
10. For the data in Exercise 10 of Section 10.2 test, at the 1% level of significance, whether x is useful for predicting y (that is, whether $\beta_1 \neq 0$).

Applications

11. For the data in Exercise 11 of Section 10.2 construct a 90% confidence interval for the mean number of new words acquired per month by children between 13 and 18 months of age.
12. For the data in Exercise 12 of Section 10.2 construct a 90% confidence interval for the mean increased braking distance for each additional 100 pounds of vehicle weight.
13. For the data in Exercise 13 of Section 10.2 test, at the 10% level of significance, whether age is useful for predicting resting heart rate.
14. For the data in Exercise 14 of Section 10.2 test, at the 10% level of significance, whether wind speed is useful for predicting wave height.
15. For the situation described in Exercise 15 of Section 10.2
    a. Construct the 95% confidence interval for the mean increase in revenue per additional thousand dollars spent on advertising.
    b. An advertising agency tells the business owner that for every additional thousand dollars spent on advertising, revenue will increase by over $\$25,000$. Test this claim (which is the alternative hypothesis) at the 5% level of significance.
    c. Perform the test of part (b) at the 10% level of significance.
    d. Based on the results in (b) and (c), how believable is the ad agency's claim? (This is a subjective judgement.)
16. For the situation described in Exercise 16 of Section 10.2
    a. Construct the 90% confidence interval for the mean increase in height per additional inch of length at age two.
    b. It is claimed that for girls each additional inch of length at age two means more than an additional inch of height at maturity. Test this claim (which is the alternative hypothesis) at the 10% level of significance.
17. For the data in Exercise 17 of Section 10.2 test, at the 10% level of significance, whether course average before the final exam is useful for predicting the final exam grade.

18. For the situation described in Exercise 18 of Section 10.2, an agronomist claims that each additional million acres planted results in more than $750,000$ additional acres harvested. Test this claim at the 1% level of significance.
19. For the data in Exercise 19 of Section 10.2 test, at the $1/10$th of 1% level of significance, whether, ignoring all other facts such as age and body mass, the amount of the medication consumed is a useful predictor of blood concentration of the active ingredient.
20. For the data in Exercise 20 of Section 10.2 test, at the 1% level of significance, whether for each additional inch of girth the age of the tree increases by at least two and one-half years.
21. For the data in Exercise 21 of Section 10.2
    a. Construct the 95% confidence interval for the mean increase in strength at 28 days for each additional hundred psi increase in strength at 3 days.
    b. Test, at the $1/10$th of 1% level of significance, whether the 3-day strength is useful for predicting 28-day strength.
22. For the situation described in Exercise 22 of Section 10.2
    a. Construct the 99% confidence interval for the mean decrease in energy demand for each one-degree drop in temperature.
    b. An engineer with the power company believes that for each one-degree increase in temperature, daily energy demand will decrease by more than 3.6 million watt-hours. Test this claim at the 1% level of significance.

Large Data Set Exercises

Large Data Sets not available

23. Large $\text{Data Set } 1$ lists the SAT scores and GPAs of $1,000$ students.
    a. Compute the 90% confidence interval for the slope β_1 of the population regression line with SAT score as the independent variable (x) and GPA as the dependent variable (y).
    b. Test, at the 10% level of significance, the hypothesis that the slope of the population regression line is greater than 0.001, against the null hypothesis that it is exactly 0.001.
24. Large $\text{Data Set } 12$ lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).
    a. Compute the 95% confidence interval for the slope β_1 of the population regression line with scores using the original clubs as the independent variable (x) and scores using the new clubs as the dependent variable (y).
    b. Test, at the 10% level of significance, the hypothesis that the slope of the population regression line is different from 1, against the null hypothesis that it is exactly 1.
25. Large $\text{Data Set } 13$ records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.
    a. Compute the 95% confidence interval for the slope β_1 of the population regression line with the number of bidders present at the auction as the independent variable(x) and sales price as the dependent variable (y).
    b. Test, at the 10% level of significance, the hypothesis that the average sales price increases by more than $\$90$ for each additional bidder at an auction, against the default that it increases by exactly $\$90$.

Answers

1. 0.743 ± 0.578
2.
3. -0.610 ± 0.633
4.
5. $T = 1.732,\ \pm t_{0.05} = \pm 2.353$, do not reject H_0
6.
7. 0.6 ± 0.451
8.
9. $T = -4.481,\ \pm t_{0.005} = \pm 3.355$, reject H_0
10.
11. 4.8 ± 1.7 words
12.
13. $T = 2.843,\ \pm t_{0.05} = \pm 1.860$, reject H_0
14.
15. a. 42.024 ± 28.011 thousand dollars

b. $T = 1.487,\ \pm t_{0.05} = \pm 1.943$, do not reject H_0
c. $t_{0.10} = 1.440$, reject H_0

16.
17. $T = 4.096,\ \pm t_{0.05} = \pm 1.771$, reject H_0
18.
19. $T = 25.524,\ \pm t_{0.0005} = \pm 3.505$, reject H_0
20.
21. a. 2.550 ± 0.127 hundred psi
b. $T = 41.072,\ \pm t_{0.005} = \pm 3.674$, reject H_0
22.
23. a. $(0.0014, 0.0018)$
b. $H_0 : \beta_1 = 0.001\ vs\ H_a : \beta_1 > 0.001$. Test Statistic: $Z = 6.1625$. Rejection Region: $[1.28, +\infty)$. Decision: Reject H_0
24.
25. a. $(101.789, 131.4435)$
b. $H_0 : \beta_1 = 90\ vs\ H_a : \beta_1 > 90$. Test Statistic: $T = 3.5938,\ d.f. = 58$. Rejection Region: $[1.296, +\infty)$. Decision: Reject H_0

10.6 The Coefficient of Determination

Basic

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in Section 10.2, Section 10.4, and Section 10.5.

1. For the sample data set of Exercise 1 of Section 10.2 find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
2. For the sample data set of Exercise 2 of Section 10.2" find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
3. For the sample data set of Exercise 3 of Section 10.2 find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
4. For the sample data set of Exercise 4 of Section 10.2 find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
5. For the sample data set of Exercise 5 of Section 10.2 find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
6. For the sample data set of Exercise 6 of Section 10.2 find the coefficient of determination using the formula $r^2 = \widehat{\beta_1} SS_{xy}/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
7. For the sample data set of Exercise 7 of Section 10.2 find the coefficient of determination using the formula $r^2 = (SS_{yy} - SSE)/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
8. For the sample data set of Exercise 7 of Section 10.2 find the coefficient of determination using the formula $r^2 = (SS_{yy} - SSE)/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
9. For the sample data set of Exercise 7 of Section 10.2 find the coefficient of determination using the formula $r^2 = (SS_{yy} - SSE)/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.
10. For the sample data set of Exercise 7 of Section 10.2 find the coefficient of determination using the formula $r^2 = (SS_{yy} - SSE)/SS_{yy}$. Confirm your answer by squaring r as computed in that exercise.

Applications

11. For the data in Exercise 11 of Section 10.2 compute the coefficient of determination and interpret its value in the context of age and vocabulary.
12. For the data in Exercise 12 of Section 10.2" compute the coefficient of determination and interpret its value in the context of vehicle weight and braking distance.
13. For the data in Exercise 13 of Section 10.2 compute the coefficient of determination and interpret its value in the context of age and resting heart rate. In the age range of the data, does age seem to be a very important factor with regard to heart rate?
14. For the data in Exercise 14 of Section 10.2 compute the coefficient of determination and interpret its value in the context of wind speed and wave height. Does wind speed seem to be a very important factor with regard to wave height?

15. For the data in Exercise 15 of Section 10.2 find the proportion of the variability in revenue that is explained by level of advertising.
16. For the data in Exercise 16 of Section 10.2 find the proportion of the variability in adult height that is explained by the variation in length at age two.
17. For the data in Exercise 17 of Section 10.2 compute the coefficient of determination and interpret its value in the context of course average before the final exam and score on the final exam.
18. For the data in Exercise 18 of Section 10.2 compute the coefficient of determination and interpret its value in the context of acres planted and acres harvested.
19. For the data in Exercise 19 of Section 10.2 compute the coefficient of determination and interpret its value in the context of the amount of the medication consumed and blood concentration of the active ingredient.
20. For the data in Exercise 20 of Section 10.2 compute the coefficient of determination and interpret its value in the context of tree size and age.
21. For the data in Exercise 21 of Section 10.2 find the proportion of the variability in 28-day strength of concrete that is accounted for by variation in 3-day strength.
22. For the data in Exercise 22 of Section 10.2 find the proportion of the variability in energy demand that is accounted for by variation in average temperature.

Large Data Set Exercises

Large Data Sets not available

23. Large $\text{Data Set } 1$ lists the SAT scores and GPAs of $1,000$ students. Compute the coefficient of determination and interpret its value in the context of SAT scores and GPAs.
24. Large $\text{Data Set } 12$ lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs). Compute the coefficient of determination and interpret its value in the context of golf scores with the two kinds of golf clubs.
25. Large $\text{Data Set } 13$ records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions. Compute the coefficient of determination and interpret its value in the context of the number of bidders at an auction and the price of this type of antique grandfather clock.

Answers

1. 0.848
2.
3. 0.631
4.
5. 0.5
6.
7. 0.766
8.
9. 0.715
10.
11. 0.898; about 90% of the variability in vocabulary is explained by age
12.
13. 0.503; about 50% of the variability in heart rate is explained by age. Age is a significant but not dominant factor in explaining heart rate.
14.
15. The proportion is $r^2 = 0.692$
16.
17. 0.563; about 56% of the variability in final exam scores is explained by course average before the final exam
18.
19. 0.931; about 93% of the variability in the blood concentration of the active ingredient is explained by the amount of the medication consumed
20.
21. The proportion is $r^2 = 0.984$

22.
23. $r^2 = 21.17\%$
24.
25. $r^2 = 81.04\%$

10.7 Estimation and Prediction

Basic

For the Basic and Application exercises in this section use the computations that were done for the exercises with the same number in previous sections.

1. For the sample data set of Exercise 1 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 4$.
   b. Construct the 90% confidence interval for that mean value.
2. For the sample data set of Exercise 2 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 4$.
   b. Construct the 90% confidence interval for that mean value.
3. For the sample data set of Exercise 3 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 7$.
   b. Construct the 95% confidence interval for that mean value.
4. For the sample data set of Exercise 4 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 2$.
   b. Construct the 80% confidence interval for that mean value.
5. For the sample data set of Exercise 5 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 1$.
   b. Construct the 80% confidence interval for that mean value.
6. For the sample data set of Exercise 6 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 5$.
   b. Construct the 95% confidence interval for that mean value.
7. For the sample data set of Exercise 7 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 6$.
   b. Construct the 99% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = 12$? Explain.
8. For the sample data set of Exercise 8 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 12$.
   b. Construct the 80% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = 0$? Explain.
9. For the sample data set of Exercise 9 of Section 10.2
   a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 0$.
   b. Construct the 90% confidence interval for that mean value.
   c. Is it valid to make the same estimates for $x = -1$? Explain.
10. For the sample data set of Exercise 9 of Section 10.2
    a. Give a point estimate for the mean value of y in the sub-population determined by the condition $x = 8$.
    b. Construct the 95% confidence interval for that mean value.
    c. Is it valid to make the same estimates for $x = 0$? Explain.

Applications

11. For the data in Exercise 11 of Section 10.2
    a. Give a point estimate for the average number of words in the vocabulary of 18-month-old children.
    b. Construct the 95% confidence interval for that mean value.

c. Construct the 95% confidence interval for that mean value.

12. For the data in Exercise 12 of Section 10.2
    a. Give a point estimate for the average braking distance of automobiles that weigh $3,250$ pounds.
    b. Construct the 80% confidence interval for that mean value.
    c. Is it valid to make the same estimates for $5,000$-pound automobiles? Explain.
13. For the data in Exercise 13 of Section 10.2
    a. Give a point estimate for the resting heart rate of a man who is 35 years old.
    b. One of the men in the sample is 35 years old, but his resting heart rate is not what you computed in part (a). Explain why this is not a contradiction.
    c. Construct the 90% confidence interval for the mean resting heart rate of all 35-year-old men.
14. For the data in Exercise 14 of Section 10.2
    a. Give a point estimate for the wave height when the wind speed is 13 miles per hour.
    b. One of the wind speeds in the sample is 13 miles per hour, but the height of waves that day is not what you computed in part (a). Explain why this is not a contradiction.
    c. Construct the 95% confidence interval for the mean wave height on days when the wind speed is 13 miles per hour.
15. For the data in Exercise 15 of Section 10.2
    a. The business owner intends to spend $\$2,500$ on advertising next year. Give an estimate of next year's revenue based on this fact.
    b. Construct the 90% prediction interval for next year's revenue, based on the intent to spend $\$2,500$ on advertising.
16. For the data in Exercise 16 of Section 10.2
    a. A two-year-old girl is 32.3 inches long. Predict her adult height.
    b. Construct the 95% prediction interval for the girl's adult height.
17. For the data in Exercise 17 of Section 10.2
    a. Lodovico has a 78.6 average in his physics class just before the final. Give a point estimate of what his final exam grade will be.
    b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
    c. Based on your answer to (b), construct an interval estimate for Lodovico's final exam grade at the 90% level of confidence.
18. For the data in Exercise 18 of Section 10.2
    a. This year 86.2 million acres of corn were planted. Give a point estimate of the number of acres that will be harvested this year.
    b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
    c. Based on your answer to (b), construct an interval estimate for the number of acres that will be harvested this year, at the 99% level of confidence.
19. For the data in Exercise 19 of Section 10.2
    a. Give a point estimate for the blood concentration of the active ingredient of this medication in a man who has consumed 1.5 ounces of the medication just recently.
    b. Gratiano just consumed 1.5 ounces of this medication 30 minutes ago. Construct a 95% prediction interval for the concentration of the active ingredient in his blood right now.
20. For the data in Exercise 20 of Section 10.2
    a. You measure the girth of a free-standing oak tree five feet off the ground and obtain the value 127 inches. How old do you estimate the tree to be?
    b. Construct a 90% prediction interval for the age of this tree.
21. For the data in Exercise 21 of Section 10.2
    a. A test cylinder of concrete three days old fails at $1,750$ psi. Predict what the 28-day strength of the concrete will be.
    b. Construct a 99% prediction interval for the 28-day strength of this concrete.
    c. Based on your answer to (b), what would be the minimum 28-day strength you could expect this concrete to exhibit?
22. For the data in Exercise 22 of Section 10.2
    a. Tomorrow's average temperature is forecast to be 53 degrees. Estimate the energy demand tomorrow.

b. Construct a 99% prediction interval for the energy demand tomorrow.
c. Based on your answer to (b), what would be the minimum demand you could expect?

Large Data Set Exercises

Large Data Sets not available

23. Large $\text{Data Set } 1$ lists the SAT scores and GPAs of $1,000$ students.
    a. Give a point estimate of the mean GPA of all students who score 1350 on the SAT.
    b. Construct a 90% confidence interval for the mean GPA of all students who score 1350 on the SAT.
24. Large $\text{Data Set } 12$ lists the golf scores on one round of golf for 75 golfers first using their own original clubs, then using clubs of a new, experimental design (after two months of familiarization with the new clubs).
    a. Thurio averages 72 strokes per round with his own clubs. Give a point estimate for his score on one round if he switches to the new clubs.
    b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
    c. Based on your answer to (b), construct an interval estimate for Thurio's score on one round if he switches to the new clubs, at 90% confidence.
25. Large $\text{Data Set } 13$ records the number of bidders and sales price of a particular type of antique grandfather clock at 60 auctions.
    a. There are seven likely bidders at the Verona auction today. Give a point estimate for the price of such a clock at today's auction.
    b. Explain whether an interval estimate for this problem is a confidence interval or a prediction interval.
    c. Based on your answer to (b), construct an interval estimate for the likely sale price of such a clock at today's sale, at 95% confidence.

Answers

1. a. 5.647
   b. 5.647 ± 1.253
2.
3. a. -0.188
   b. -0.188 ± 3.041
4.
5. a. 1.875
   b. 1.875 ± 1.423
6.
7. a. 5.4
   b. 5.4 ± 3.355
   c. invalid (extrapolation)
8.
9. a. 2.4
   b. 2.4 ± 1.474
   c. valid (-1 is in the range of the x-values in the data set)
10.
11. a. 31.3 words
    b. 31.3 ± 7.1 words
    c. not valid, since two years is 24 months, hence this is extrapolation
12.
13. a. 73.2 beats/min
    b. The man's heart rate is not the predicted average for all men his age.
    c. 73.2 ± 1.2 beats/min
14.
15. a. $\$224,562$

b. $\$224,562 \pm \$28,699$

16.

17. a. 74
    b. Prediction (one person, not an average for all who have average 78.6 before the final exam)
    c. 74 ± 24

18.

19. a. 0.066%
    b. $0.066 \pm 0.034\%$

20.

21. a. $4,656$ psi
    b. $4,656 \pm 321$ psi
    c. $4,656 - 321 = 4,335$ psi

22.

23. a. 2.19
    b. $(2.1421, 2.2316)$

24.

25. a. 7771.39
    b. A prediction interval.
    c. $(7410.41, 8132.38)$

10.8 A Complete Example

Basic

The exercises in this section are unrelated to those in previous sections.

1. The data give the amount x of silicofluoride in the water (mg/L) and the amount y of lead in the bloodstream (μg/dL) of ten children in various communities with and without municipal water. Perform a complete analysis of the data, in analogy with the discussion in this section (that is, make a scatter plot, do preliminary computations, find the least squares regression line, find SSE, s_ε and r, and so on). In the hypothesis test use as the alternative hypothesis $\beta_1 > 0$, and test at the 5% level of significance. Use confidence level 95% for the confidence interval for β_1. Construct 95% confidence and predictions intervals at $x_p = 2$ at the end.

x	0.0	0.0	1.1	1.4	1.6
y	0.3	0.1	4.7	3.2	5.1
x	1.7	2.0	2.0	2.2	2.2
y	7.0	5.0	6.1	8.6	9.5

(10.E.32)

2. The table gives the weight x (thousands of pounds) and available heat energy y (million BTU) of a standard cord of various species of wood typically used for heating. Perform a complete analysis of the data, in analogy with the discussion in this section (that is, make a scatter plot, do preliminary computations, find the least squares regression line, find SSE, s_ε and r, and so on). In the hypothesis test use as the alternative hypothesis β_1, and test at the 5% level of significance. Use confidence level 95% for the confidence interval for β_1. Construct 95% confidence and predictions intervals at $x_p = 5$ at the end.

x	3.37	3.50	4.29	4.00	4.64
y	23.6	17.5	20.1	21.6	28.1
x	4.99	4.94	5.48	3.26	4.16
y	25.3	27.0	30.7	18.9	20.7

(10.E.33)

Large Data Set Exercises

Large Data Sets not available

3. Large Data Sets 3 and 3A list the shoe sizes and heights of 174 customers entering a shoe store. The gender of the customer is not indicated in Large Data Set 3. However, men's and women's shoes are not measured on the same scale; for example, a size 8 shoe for men is not the same size as a size 8 shoe for women. Thus it would not be meaningful to apply regression analysis to

Large Data Set 3. Nevertheless, compute the scatter diagrams, with shoe size as the independent variable (x) and height as the dependent variable (y), for (i) just the data on men, (ii) just the data on women, and (iii) the full mixed data set with both men and women. Does the third, invalid scatter diagram look markedly different from the other two?

4. Separate out from Large Data Set 3A just the data on men and do a complete analysis, with shoe size as the independent variable (x) and height as the dependent variable (y). Use $\alpha = 0.05$ and $x_p = 10$ whenever appropriate.
5. Separate out from Large Data Set 3A just the data on women and do a complete analysis, with shoe size as the independent variable (x) and height as the dependent variable (y). Use $\alpha = 0.05$ and $x_p = 10$ whenever appropriate.

Answers

1. $$\sum x = 14.2,\ \sum y = 49.6,\ \sum xy = 91.73,\ \sum x^2 = 26.3,\ \sum y^2 = 333.86 \quad (10.E.34)$$
$$SS_{xx} = 6.136,\ SS_{xy} = 21.298,\ SS_{yy} = 87.844$$
$$\bar{x} = 1.42,\ \bar{y} = 4.96$$
$$\widehat{\beta_1} = 3.47,\ \widehat{\beta_0} = 0.03$$
$$SSE = 13.92$$
$$s_\varepsilon = 1.32$$
$$r = 0.9174, r^2 = 0.8416$$
$$df = 8, T = 6.518$$

   The 95% confidence interval for β_1 is: $(2.24, 4.70)$
   At $x_p = 2$ the 95% confidence interval for $E(y)$ is $(5.77, 8.17)$
   At $x_p = 2$ the 95% confidence interval for y is $(3.73, 10.21)$
2.
3. The positively correlated trend seems less profound than that in each of the previous plots.
4.
5. The regression line: $\hat{y} = 3.3426x + 138.7692$. Coefficient of Correlation: $r = 0.9431$. Coefficient of Determination: $r^2 = 0.8894$. $SSE = 283.2473$. $s_e = 1.9305$. A 95% confidence interval for β_1: $(3.0733, 3.6120)$. Test Statistic for $H_0 : \beta_1 = 0 : T = 24.7209$. At $x_p = 10$, $\hat{y} = 172.1956$; a 95% confidence interval for the mean value of y is: $(171.5577, 172.8335)$; and a 95% prediction interval for an individual value of y is: $(168.2974, 176.0938)$.

This page titled 10.E: Correlation and Regression (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **10.E: Correlation and Regression (Exercises)** has no license indicated.

CHAPTER OVERVIEW

11: Chi-Square Tests and F-Tests

In previous chapters you saw how to test hypotheses concerning population means and population proportions. The idea of testing hypotheses can be extended to many other situations that involve different parameters and use different test statistics. Whereas the standardized test statistics that appeared in earlier chapters followed either a normal or Student t-distribution, in this chapter the tests will involve two other very common and useful distributions, the chi-square and the F-distributions. The chi-square distribution arises in tests of hypotheses concerning the independence of two random variables and concerning whether a discrete random variable follows a specified distribution. The F-distribution arises in tests of hypotheses concerning whether or not two population variances are equal and concerning whether or not three or more population means are equal.

11.1: Chi-Square Tests for Independence
11.2: Chi-Square One-Sample Goodness-of-Fit Tests
11.3: F-tests for Equality of Two Variances
11.4: F-Tests in One-Way ANOVA
11.E: Chi-Square Tests and F-Tests (Exercises)

This page titled 11: Chi-Square Tests and F-Tests is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

11.1: Chi-Square Tests for Independence

Learning Objectives

- To understand what chi-square distributions are.
- To understand how to use a chi-square test to judge whether two factors are independent.

Chi-Square Distributions

As you know, there is a whole family of t-distributions, each one specified by a parameter called the degrees of freedom, denoted df. Similarly, all the chi-square distributions form a family, and each of its members is also specified by a parameter df, the number of degrees of freedom. Chi is a Greek letter denoted by the symbol χ and chi-square is often denoted by χ^2.

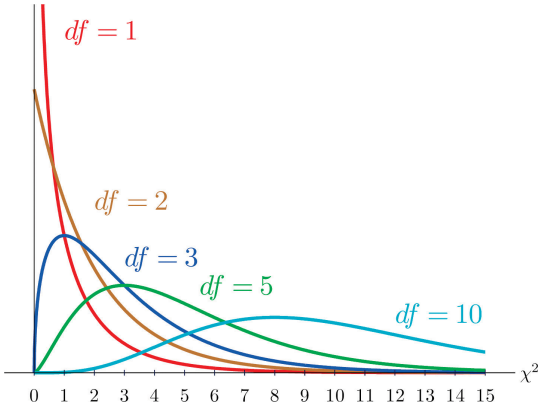


Figure 11.1.1: Many χ Distributions

Figure 11.1.1 shows several χ-square distributions for different degrees of freedom. A chi-square random variable is a random variable that assumes only positive values and follows a χ-square distribution.

Definition: critical value

The value of the chi-square random variable χ^2 with $df = k$ that cuts off a right tail of area c is denoted χ_c^2 and is called a critical value (Figure 11.1.2).

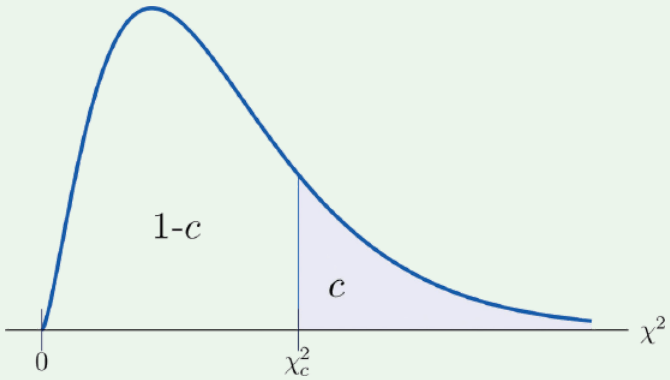
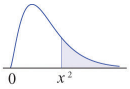


Figure 11.1.2: χ_c^2 Illustrated

Figure 11.1.3 below gives values of χ_c^2 for various values of c and under several chi-square distributions with various degrees of freedom.



Critical Values of Chi-Square Distributions

	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 11.1.3: Critical Values of Chi-Square Distributions

Tests for Independence

Hypotheses tests encountered earlier in the book had to do with how the numerical values of two population parameters compared. In this subsection we will investigate hypotheses that have to do with whether or not two random variables take their values independently, or whether the value of one has a relation to the value of the other. Thus the hypotheses will be expressed in words, not mathematical symbols. We build the discussion around the following example.

There is a theory that the gender of a baby in the womb is related to the baby's heart rate: baby girls tend to have higher heart rates. Suppose we wish to test this theory. We examine the heart rate records of 40 babies taken during their mothers' last prenatal checkups before delivery, and to each of these 40 randomly selected records we compute the values of two random measures: 1) gender and 2) heart rate. In this context these two random measures are often called factors. Since the burden of proof is that heart rate and gender are related, not that they are unrelated, the problem of testing the theory on baby gender and heart rate can be formulated as a test of the following hypotheses:

$$H_0 : \text{Baby gender and baby heart rate are independent}$$
$$vs.$$

H_a : Baby gender and baby heart rate are not independent

The factor gender has two natural categories or levels: boy and girl. We divide the second factor, heart rate, into two levels, low and high, by choosing some heart rate, say 145 beats per minute, as the cutoff between them. A heart rate below 145 beats per minute will be considered low and 145 and above considered high. The 40 records give rise to a 2×2 *contingency table*. By adjoining row totals, column totals, and a grand total we obtain the table shown as Table 11.1.1. The four entries in boldface type are counts of observations from the sample of $n = 40$. There were 11 girls with low heart rate, 17 boys with low heart rate, and so on. They form the *core* of the expanded table.

Table 11.1.1: Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	11	7	18
	Boy	17	5	22
Column Total		28	12	Total = 40

In analogy with the fact that the probability of independent events is the product of the probabilities of each event, if heart rate and gender were independent then we would expect the number in each core cell to be close to the product of the row total R and column total C of the row and column containing it, divided by the sample size n. Denoting such an expected number of observations E, these four expected values are:

- 1st row and 1st column: $E = (R \times C)/n = 18 \times 28/40 = 12.6$
- 1st row and 2nd column: $E = (R \times C)/n = 18 \times 12/40 = 5.4$
- 2nd row and 1st column: $E = (R \times C)/n = 22 \times 28/40 = 15.4$
- 2nd row and 2nd column: $E = (R \times C)/n = 22 \times 12/40 = 6.6$

We update Table 11.1.1 by placing each expected value in its corresponding core cell, right under the observed value in the cell. This gives the updated table Table 11.1.2.

Table 11.1.2: Updated Baby Gender and Heart Rate

		Heart Rate		
		Low	High	Row Total
Gender	Girl	$O = 11$ $E = 12.6$	$O = 7$ $E = 5.4$	$R = 18$
	Boy	$O = 17$ $E = 15.4$	$O = 5$ $E = 6.6$	$R = 22$
Column Total		$C = 28$	$C = 12$	$n = 40$

A measure of how much the data deviate from what we would expect to see if the factors really were independent is the sum of the squares of the difference of the numbers in each core cell, or, standardizing by dividing each square by the expected number in the cell, the sum $\sum (O - E)^2/E$. We would reject the null hypothesis that the factors are independent only if this number is large, so the test is right-tailed. In this example the random variable $\sum (O - E)^2/E$ has the chi-square distribution with one degree of freedom. If we had decided at the outset to test at the 10% level of significance, the critical value defining the rejection region would be, reading from Figure 11.1.3, $\chi^2_\alpha = \chi^2_{0.10} = 2.706$, so that the rejection region would be the interval $[2.706, \infty)$. When we compute the value of the standardized test statistic we obtain

$$\sum \frac{(O-E)^2}{E} = \frac{(11-12.6)^2}{12.6} + \frac{(7-5.4)^2}{5.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-6.6)^2}{6.6} = 1.231$$

Since $1.231 < 2.706$, the decision is not to reject H_0. See Figure 11.1.4. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that heart rate and gender are related.

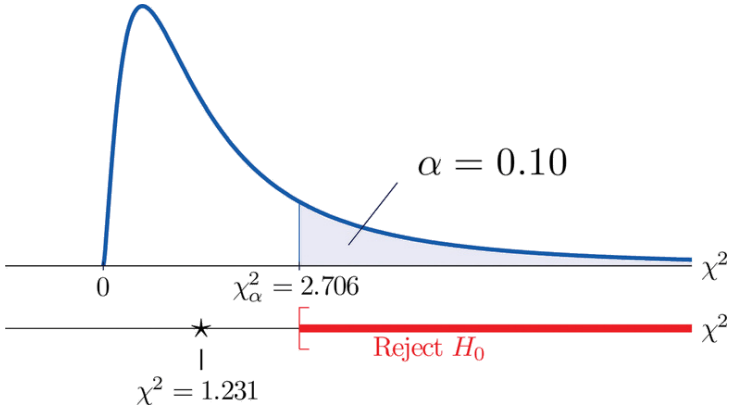


Figure 11.1.4*: Baby Gender Prediction*

With this specific example in mind, now turn to the general situation. In the general setting of testing the independence of two factors, call them Factor 1 and Factor 2, the hypotheses to be tested are

$$H_0 : \text{The two factors are independent}$$
$$vs.$$
$$H_a : \text{The two factors are not independent}$$

As in the example each factor is divided into a number of categories or levels. These could arise naturally, as in the boy-girl division of gender, or somewhat arbitrarily, as in the high-low division of heart rate. Suppose Factor 1 has I levels and Factor 2 has J levels. Then the information from a random sample gives rise to a general $I \times J$ contingency table, which with row totals, column totals, and a grand total would appear as shown in Table 11.1.3. Each cell may be labeled by a pair of indices (i, j). O_{ij} stands for the observed count of observations in the cell in row i and column j, R_i for the i^{th} row total and C_j for the j^{th} column total. To simplify the notation we will drop the indices so Table 11.1.3 becomes Table 11.1.4. Nevertheless it is important to keep in mind that the Os, the Rs and the Cs, though denoted by the same symbols, are in fact different numbers.

Table 11.1.3: General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O_{11}	$\cdots$	O_{1j}	$\cdots$	O_{1J}	R_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O_{i1}	$\cdots$	O_{ij}	$\cdots$	O_{iJ}	R_i
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O_{I1}	$\cdots$	O_{Ij}	$\cdots$	O_{IJ}	R_I
Column Total		C_1	$\cdots$	C_j	$\cdots$	C_J	n

Table 11.1.4: Simplified General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O	$\cdots$	O	$\cdots$	O	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O	$\cdots$	O	$\cdots$	O	R

	Factor 2 Levels					
	1	$\cdots$	j	$\cdots$	J	Row Total
Column Total	C	$\cdots$	C	$\cdots$	C	n

As in the example, for each core cell in the table we compute what would be the expected number E of observations if the two factors were independent. E is computed for each core cell (each cell with an O in it) of Table 11.1.4 by the rule applied in the example:

$$E = R \times Cn$$

where R is the row total and C is the column total corresponding to the cell, and n is the sample size

After the expected number is computed for every cell, Table 11.1.4 is updated to form Table 11.1.5 by inserting the computed value of E into each core cell.

Table 11.1.5: Updated General Contingency Table

		Factor 2 Levels					
		1	$\cdots$	j	$\cdots$	J	Row Total
Factor 1 Levels	1	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	O E	$\cdots$	O E	$\cdots$	O E	R
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	O E	$\cdots$	O E	$\cdots$	O E	R
Column Total		C	$\cdots$	C	$\cdots$	C	n

Here is the test statistic for the general hypothesis based on Table 11.1.5, together with the conditions that it follow a chi-square distribution.

Test Statistic for Testing the Independence of Two Factors

$$\chi^2 = \sum (O - E)^2 E$$

where the sum is over all core cells of the table.

If

1. the two study factors are independent, and
2. the observed count O of each cell in Table 11.1.5 is at least 5,

then χ^2 approximately follows a chi-square distribution with $df = (I - 1) \times (J - 1)$ degrees of freedom.

The same five-step procedures, either the critical value approach or the p-value approach, that were introduced in Section 8.1 and Section 8.3 are used to perform the test, which is always right-tailed.

Example 11.1.1

A researcher wishes to investigate whether students' scores on a college entrance examination (CEE) have any indicative power for future college performance as measured by GPA. In other words, he wishes to investigate whether the factors CEE

and GPA are independent or not. He randomly selects $n = 100$ students in a college and notes each student's score on the entrance examination and his grade point average at the end of the sophomore year. He divides entrance exam scores into two levels and grade point averages into three levels. Sorting the data according to these divisions, he forms the contingency table shown as Table 11.1.6, in which the row and column totals have already been computed.

Table 11.1.6: *CEE versus GPA Contingency Table*

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	35	12	5	52
	≥ 1800	6	24	18	48
Column Total		41	36	23	Total = 100

Test, at the 1% level of significance, whether these data provide sufficient evidence to conclude that CEE scores indicate future performance levels of incoming college freshmen as measured by GPA.

Solution

We perform the test using the critical value approach, following the usual five-step method outlined at the end of Section 8.1.

- **Step 1**. The hypotheses are

$$H_0 : \text{CEE and GPA are independent factors}$$
$$vs.$$
$$H_a : \text{CEE and GPA are not independent factors}$$

- **Step 2**. The distribution is chi-square.
- **Step 3**. To compute the value of the test statistic we must first computed the expected number for each of the six core cells (the ones whose entries are boldface):
  - 1st row and 1st column: $E = (R \times C)/n = 41 \times 52/100 = 21.32$
  - 1st row and 2nd column: $E = (R \times C)/n = 36 \times 52/100 = 18.72$
  - 1st row and 3rd column: $E = (R \times C)/n = 23 \times 52/100 = 11.96$
  - 2nd row and 1st column: $E = (R \times C)/n = 41 \times 48/100 = 19.68$
  - 2nd row and 2nd column: $E = (R \times C)/n = 36 \times 48/100 = 17.28$
  - 2nd row and 3rd column: $E = (R \times C)/n = 23 \times 48/100 = 11.04$

Table 11.1.6 is updated to Table 11.1.6.

Table 11.1.7: Updated CEE versus GPA Contingency Table

		GPA			
		< 2.7	2.7 to 3.2	> 3.2	Row Total
CEE	< 1800	$O = 35$ $E = 21.32$	$O = 12$ $E = 18.72$	$O = 5$ $E = 11.96$	$R = 52$
	≥ 1800	$O = 6$ $E = 19.68$	$O = 24$ $E = 17.28$	$O = 18$ $E = 11.04$	$R = 48$
Column Total		$C = 41$	$C = 36$	$C = 23$	$n = 100$

The test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(35-21.32)^2}{21.32} + \frac{(12-18.72)^2}{18.72} + \frac{(5-11.96)^2}{11.96} + \frac{(6-19.68)^2}{19.68} \\ &\quad + \frac{(24-17.28)^2}{17.28} + \frac{(18-11.04)^2}{11.04} \\ &= 31.75\end{aligned}$$

- **Step 4**. Since the CEE factor has two levels and the GPA factor has three, $I = 2$ and $J = 3$. Thus the test statistic follows the chi-square distribution with $df = (2-1) \times (3-1) = 2$ degrees of freedom.

  Since the test is right-tailed, the critical value is $\chi^2_{0.01}$. Reading from Figure 7.1.6 "Critical Values of Chi-Square Distributions", $\chi^2_{0.01} = 9.210$, so the rejection region is $[9.210, \infty)$.

- **Step 5**. Since $31.75 > 9.21$ the decision is to reject the null hypothesis. See Figure 11.1.5. The data provide sufficient evidence, at the 1% level of significance, to conclude that CEE score and GPA are not independent: the entrance exam score has predictive power.

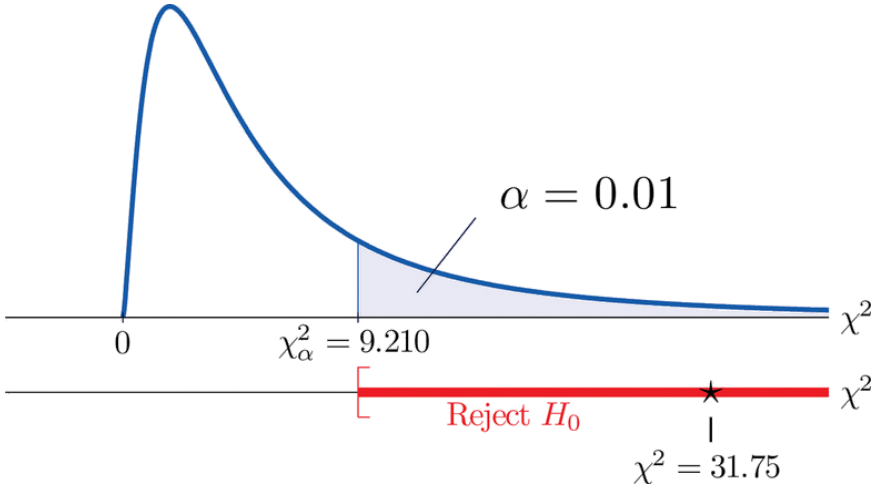


Figure 11.1.5: "Example 11.1.1"

Key Takeaway

- Critical values of a chi-square distribution with degrees of freedom df are found in Figure 7.1.6.
- A chi-square test can be used to evaluate the hypothesis that two random variables or factors are independent.

This page titled 11.1: Chi-Square Tests for Independence is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

11.2: Chi-Square One-Sample Goodness-of-Fit Tests

Learning Objectives

- To understand how to use a chi-square test to judge whether a sample fits a particular population well.

Suppose we wish to determine if an ordinary-looking six-sided die is fair, or balanced, meaning that every face has probability $1/6$ of landing on top when the die is tossed. We could toss the die dozens, maybe hundreds, of times and compare the actual number of times each face landed on top to the expected number, which would be $1/6$ of the total number of tosses. We wouldn't expect each number to be exactly $1/6$ of the total, but it should be close. To be specific, suppose the die is tossed $n = 60$ times with the results summarized in Table 11.2.1. For ease of reference we add a column of expected frequencies, which in this simple example is simply a column of $10s$. The result is shown as Table 11.2.2. In analogy with the previous section we call this an "updated" table. A measure of how much the data deviate from what we would expect to see if the die really were fair is the sum of the squares of the differences between the observed frequency O and the expected frequency E in each row, or, standardizing by dividing each square by the expected number, the sum

$$\frac{\Sigma(O-E)^2}{E}$$

If we formulate the investigation as a test of hypotheses, the test is

$$H_0 : \text{The die is fair}$$
$$vs.$$
$$H_a : \text{The die is not fair}$$

Table 11.2.1: Die Contingency Table

Die Value	Assumed Distribution	Observed Frequency
1	1/6	9
2	1/6	15
3	1/6	9
4	1/6	8
5	1/6	6
6	1/6	13

Table 11.2.2: Updated Die Contingency Table

Die Value	Assumed Distribution	Observed Freq.	Expected Freq.
1	1/6	9	10
2	1/6	15	10
3	1/6	9	10
4	1/6	8	10
5	1/6	6	10
6	1/6	13	10

We would reject the null hypothesis that the die is fair only if the number $\frac{\Sigma(O-E)^2}{E}$ is large, so the test is right-tailed. In this example the random variable $\frac{\Sigma(O-E)^2}{E}$ has the chi-square distribution with five degrees of freedom. If we had decided at the outset to test at the 10% level of significance, the critical value defining the rejection region would be, reading from Figure 7.1.6, $\chi^2_\alpha = \chi^2_{0.10} = 9.236$, so that the rejection region would be the interval

$[9.236, \infty)$

. When we compute the value of the standardized test statistic using the numbers in the last two columns of Table 11.2.2, we obtain

$$\begin{aligned}\sum \frac{(O-E)^2}{E} &= \frac{(-1)^2}{10} + \frac{(5)^2}{10} + \frac{(-1)^2}{10} + \frac{(-2)^2}{10} + \frac{(-4)^2}{10} + \frac{(3)^2}{10} \\ &= 0.1 + 2.5 + 0.1 + 0.4 + 1.6 + 0.9 \\ &= 5.6\end{aligned}$$

Since $5.6 < 9.236$ the decision is not to reject H_0. See Figure 11.2.1. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the die is loaded.

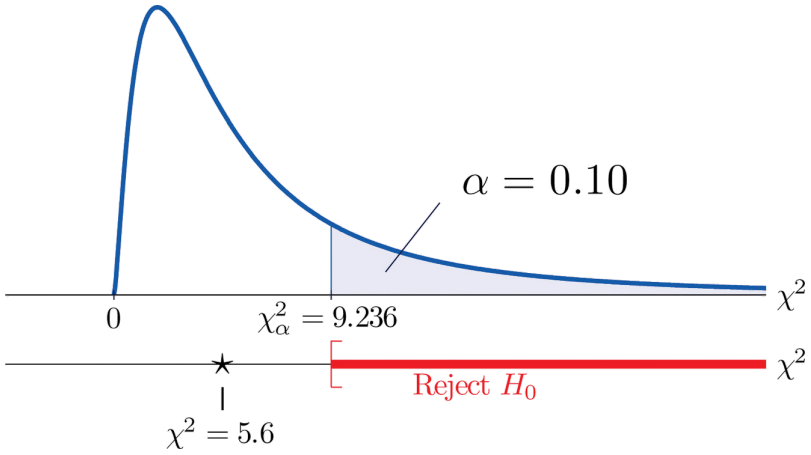


Figure 11.2.1: Balanced Die

In the general situation we consider a discrete random variable that can take I different values, $x_1, x_2, \cdots, x_I$, for which the default assumption is that the probability distribution is

x	x_1	x_2	$\cdots$	x_I
$P(x)$	p_1	p_2	$\cdots$	p_I

We wish to test the hypotheses:

$$H_0 : \text{The assumed probability distribution for X is valid}$$
$$vs.$$
$$H_a : \text{The assumed probability distribution for X is not valid}$$

We take a sample of size n and obtain a list of observed frequencies. This is shown in Table 11.2.3. Based on the assumed probability distribution we also have a list of assumed frequencies, each of which is defined and computed by the formula

$$Ei = n \times pi$$

Table 11.2.3: General Contingency Table

Factor Levels	Assumed Distribution	Observed Frequency
1	p_1	O_1
2	p_2	O_2
$\vdots$	$\vdots$	$\vdots$
I	p_I	O_I

Table 11.2.3 is updated to Table 11.2.4 by adding the expected frequency for each value of X. To simplify the notation we drop indices for the observed and expected frequencies and represent Table 11.2.4 by Table 11.2.5.

Table 11.2.4: Updated General Contingency Table

Factor Levels	Assumed Distribution	Observed Freq.	Expected Freq.
1	p_1	O_1	E_1

Factor Levels	Assumed Distribution	Observed Freq.	Expected Freq.
2	p_2	O_2	E_2
⋮	⋮	⋮	⋮
I	p_I	O_I	E_I

Table 11.2.5: Simplified Updated General Contingency Table

Factor Levels	Assumed Distribution	Observed Freq.	Expected Freq.
1	p_1	O	E
2	p_2	O	E
⋮	⋮	⋮	⋮
I	p_I	O	E

Here is the test statistic for the general hypothesis based on Table 11.2.5, together with the conditions that it follow a chi-square distribution.

Test Statistic for Testing Goodness of Fit to a Discrete Probability Distribution

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where the sum is over all the rows of the table (one for each value of X).

If

1. the true probability distribution of X is as assumed, and
2. the observed count O of each cell in Table 11.2.5 is at least 5,

then χ^2 approximately follows a chi-square distribution with $df = I - 1$ degrees of freedom.

The test is known as a goodness-of-fit χ^2 test since it tests the null hypothesis that the sample fits the assumed probability distribution well. It is always right-tailed, since deviation from the assumed probability distribution corresponds to large values of χ^2.

Testing is done using either of the usual five-step procedures.

Example 11.2.1

Table 11.2.6 shows the distribution of various ethnic groups in the population of a particular state based on a decennial U.S. census. Five years later a random sample of 2,500 residents of the state was taken, with the results given in Table 11.2.7 (along with the probability distribution from the census year). Test, at the 1% level of significance, whether there is sufficient evidence in the sample to conclude that the distribution of ethnic groups in this state five years after the census had changed from that in the census year.

Table 11.2.6: Ethnic Groups in the Census Year

Ethnicity	White	Black	Amer.-Indian	Hispanic	Asian	Others
Proportion	0.743	0.216	0.012	0.012	0.008	0.009

Table 11.2.7: Sample Data Five Years After the Census Year

Ethnicity	Assumed Distribution	Observed Frequency
White	0.743	1732

Ethnicity	Assumed Distribution	Observed Frequency
Black	0.216	538
American-Indian	0.012	32
Hispanic	0.012	42
Asian	0.008	133
Others	0.009	23

Solution

We test using the critical value approach.

- **Step 1**. The hypotheses of interest in this case can be expressed as

$$H_0 : \text{The distribution of ethnic groups has not changed}$$
$$vs.$$
$$H_a : \text{The distribution of ethnic groups has changed}$$

- **Step 2**. The distribution is chi-square.
- **Step 3**. To compute the value of the test statistic we must first compute the expected number for each row of Table 11.2.7. Since $n = 2500$, using the formula $E_i = n \times p_i$ and the values of p_i from either Table 11.2.6 or Table 11.2.7,

$$E_1 = 2500 \times 0.743 = 1857.5$$
$$E_2 = 2500 \times 0.216 = 540$$
$$E_3 = 2500 \times 0.012 = 30$$
$$E_4 = 2500 \times 0.012 = 30$$
$$E_5 = 2500 \times 0.008 = 20$$
$$E_6 = 2500 \times 0.009 = 22.5$$

Table 11.2.7 is updated to Table 11.2.8.

Table 11.2.8: Observed and Expected Frequencies Five Years After the Census Year

Ethnicity	Assumed Dist.	Observed Freq.	Expected Freq.
White	0.743	1732	1857.5
Black	0.216	538	540
American-Indian	0.012	32	30
Hispanic	0.012	42	30
Asian	0.008	133	20
Others	0.009	23	22.5

The value of the test statistic is

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(1732-1857.5)^2}{1857.5} + \frac{(538-540)^2}{540} + \frac{(32-30)^2}{30} \\ &\quad + \frac{(42-30)^2}{30} + \frac{(133-20)^2}{20} + \frac{(23-22.5)^2}{22.5} \\ &= 651.881\end{aligned}$$

Since the random variable takes six values, $I = 6$. Thus the test statistic follows the chi-square distribution with $df = 6 - 1 = 5$ degrees of freedom.

Since the test is right-tailed, the critical value is $\chi^2_{0.01}$. Reading from Figure 7.1.6, $\chi^2_{0.01} = 15.086$, so the rejection region is $[15.086, \infty)$.

Since $651.881 > 15.086$ the decision is to reject the null hypothesis. See Figure 11.2.2. The data provide sufficient evidence, at the 1% level of significance, to conclude that the ethnic distribution in this state has changed in the five years since the U.S. census.

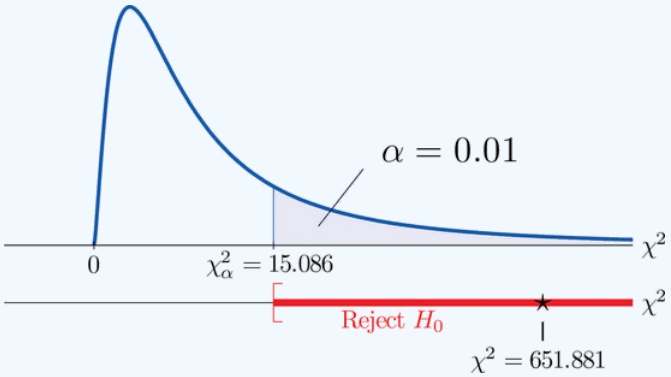


Figure 11.2.2: "Example 11.2.1"

Key Takeaway

- The chi-square goodness-of-fit test can be used to evaluate the hypothesis that a sample is taken from a population with an assumed specific probability distribution.

This page titled 11.2: Chi-Square One-Sample Goodness-of-Fit Tests is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

11.3: F-tests for Equality of Two Variances

Learning Objectives

- To understand what F-distributions are.
- To understand how to use an F-test to judge whether two population variances are equal.

F-Distributions

Another important and useful family of distributions in statistics is the family of F-distributions. Each member of the F-distribution family is specified by a pair of parameters called degrees of freedom and denoted df_1 and df_2. Figure 11.3.1 shows several F-distributions for different pairs of degrees of freedom. An F random variable is a random variable that assumes only positive values and follows an F-distribution.

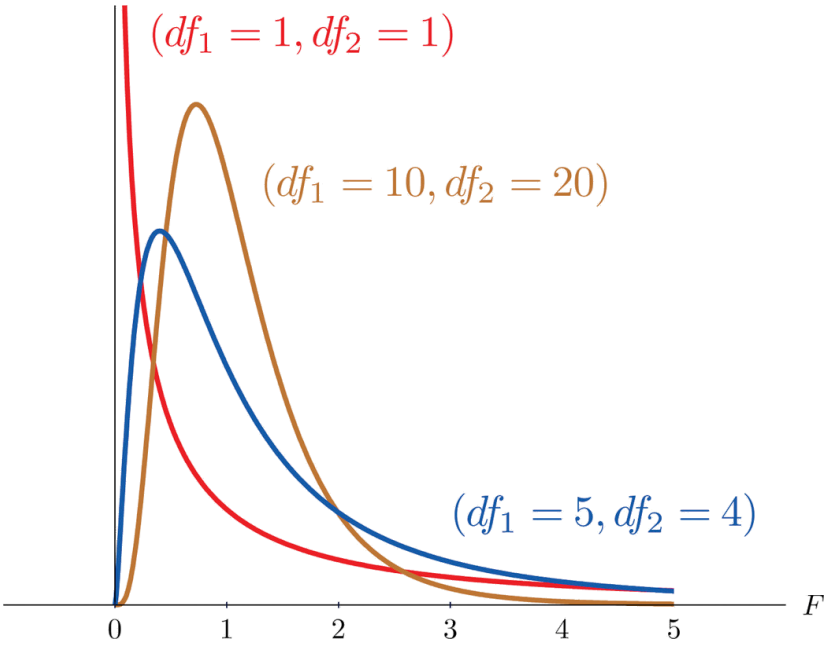


Figure 11.3.1: Many F-Distributions

The parameter df_1 is often referred to as the numerator degrees of freedom and the parameter df_2 as the denominator degrees of freedom. It is important to keep in mind that they are not interchangeable. For example, the F-distribution with degrees of freedom $df_1 = 3$ and $df_2 = 8$ is a different distribution from the F-distribution with degrees of freedom $df_1 = 8$ and $df_2 = 3$.

Definition: critical value

The value of the F random variable F with degrees of freedom df_1 and df_2 that cuts off a right tail of area c is denoted F_c and is called a critical value (Figure 11.3.2).

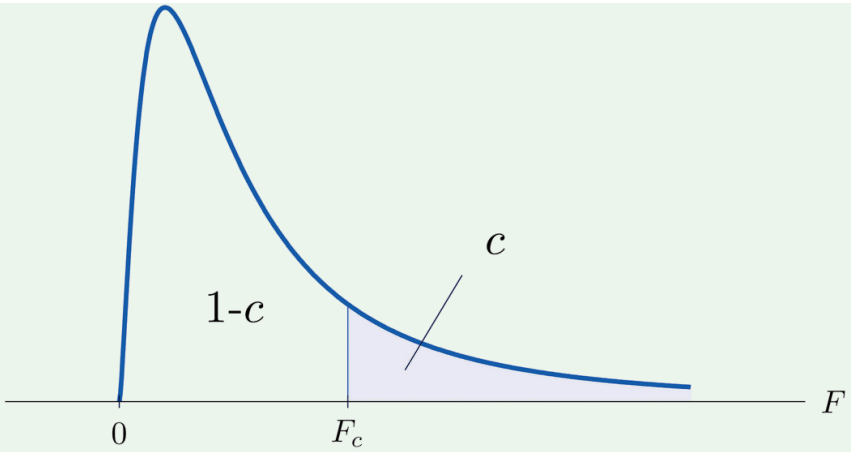


Figure 11.3.2: F_c Illustrated

Tables containing the values of F_c are given in Chapter 11. Each of the tables is for a fixed collection of values of c, either 0.900, 0.950, 0.975, 0.990, and 0.995 (yielding what are called "lower" critical values), or 0.005, 0.010, 0.025, 0.050, and 0.100 (yielding what are called "upper" critical values). In each table critical values are given for various pairs (df_1, df_2). We illustrate the use of the tables with several examples.

Example 11.3.1: an F random variable

Suppose F is an F random variable with degrees of freedom $df_1 = 5$ and $df_2 = 4$. Use the tables to find

1. $F_{0.10}$
2. $F_{0.95}$

Solution

1. The column headings of all the tables contain $df_1 = 5$. Look for the table for which 0.10 is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 4$ in the left margin of the table. A portion of the relevant table is provided. The entry in the intersection of the column with heading $df_1 = 5$ and the row with the headings 0.10 and $df_2 = 4$, which is shaded in the table provided, is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯	5	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.005	4	⋯	⋯	⋯	22.5	⋯
0.01	4	⋯	⋯	⋯	15.5	⋯
0.025	4	⋯	⋯	⋯	9.36	⋯
0.05	4	⋯	⋯	⋯	6.26	⋯
0.10	4	⋯	⋯	⋯	4.05	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

2. Look for the table for which 0.95 is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2 = 4$ in the left margin of the table. A portion of the relevant table is provided. The entry in the intersection of the column with heading $df_1 = 5$ and the row with the headings 0.95 and $df_2 = 4$, which is shaded in the table provided, is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯	5	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.90	4	⋯	⋯	⋯	0.28	⋯

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯	5	⋯
0.95	4	⋯	⋯	⋯	0.19	⋯
0.975	4	⋯	⋯	⋯	0.14	⋯
0.99	4	⋯	⋯	⋯	0.09	⋯
0.995	4	⋯	⋯	⋯	0.06	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Example 11.3.2

Suppose F is an F random variable with degrees of freedom $df_1 = 2$ and $df_2 = 20$. Let $\alpha = 0.05$. Use the tables to find

1. F_{α}
2. $F_{\alpha/2}$
3. $F_{1-\alpha}$
4. $F_{1-\alpha/2}$

Solution

1. The column headings of all the tables contain $df_1 = 2$. Look for the table for which $\alpha = 0.05$ is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1 = 2$ and the row with the headings 0.05 and $df_2 = 20$ is the answer,

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.005	20	⋯	6.99	⋯
0.01	20	⋯	5.85	⋯
0.025	20	⋯	4.46	⋯
0.05	20	⋯	3.49	⋯
0.10	20	⋯	2.59	⋯
⋮	⋮	⋮	⋮	⋮

2. Look for the table for which $\alpha/2 = 0.025$ is one of the entries on the extreme left (a table of upper critical values) and that has a row heading $df_2 = 20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1 = 2$ and the row with the headings 0.025 and $df_2 = 20$ is the answer, $F_{0.025} = 4.46$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.005	20	⋯	6.99	⋯
0.01	20	⋯	5.85	⋯
0.025	20	⋯	4.46	⋯
0.05	20	⋯	3.49	⋯
0.10	20	⋯	2.59	⋯
⋮	⋮	⋮	⋮	⋮

3. Look for the table for which $1-\alpha=0.95$ is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2=20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1=2$ and the row with the headings 0.95 and $df_2=20$ is the answer, $F_{0.95}=0.05$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.90	20	⋯	0.11	⋯
0.95	20	⋯	0.05	⋯
0.975	20	⋯	0.03	⋯
0.99	20	⋯	0.01	⋯
0.995	20	⋯	0.01	⋯
⋮	⋮	⋮	⋮	⋮

4. Look for the table for which $1-\alpha/2=0.975$ is one of the entries on the extreme left (a table of lower critical values) and that has a row heading $df_2=20$ in the left margin of the table. A portion of the relevant table is provided. The shaded entry, in the intersection of the column with heading $df_1=2$ and the row with the headings 0.975 and $df_2=20$ is the answer, $F_{0.975}=0.03$.

F Tail Area	$\frac{df_1}{df_2}$	1	2	⋯
⋮	⋮	⋮	⋮	⋮
0.90	20	⋯	0.11	⋯
0.95	20	⋯	0.05	⋯
0.975	20	⋯	0.03	⋯
0.99	20	⋯	0.01	⋯
0.995	20	⋯	0.01	⋯
⋮	⋮	⋮	⋮	⋮

A fact that sometimes allows us to find a critical value from a table that we could not read otherwise is:

If $F_u(r,s)$ denotes the value of the F-distribution with degrees of freedom $df_1=r$ and $df_2=s$ that cuts off a right tail of area u, then

$$F_c(k,l)=\frac{1}{F_{1-c}(l,k)}$$

Example 11.3.3

Use the tables to find

1. $F_{0.01}$ for an F random variable with $df_1=13$ and $df_2=8$.
2. $F_{0.975}$ for an F random variable with $df_1=40$ and $df_2=10$.

Solution

1. There is no table with $df_1=13$, but there is one with $df_1=8$. Thus we use the fact that

$$F_{0.01}(13,8)=\frac{1}{F_{0.99}(8,13)}$$

Using the relevant table we find that $F_{0.99}(8,13)=0.18$, hence $F_{0.01}(13,8)=0.18^{-1}=5.556$.

2. There is no table with $df_1=40$, but there is one with $df_1=10$. Thus we use the fact that

$$F_{0.975}(40,10) = \frac{1}{F_{0.025}(10,40)}$$

Using the relevant table we find that $F_{0.025}(10,40) = 3.31$, hence $F_{0.975}(40,10) = 3.31^{-1} = 0.302$.

F-Tests for Equality of Two Variances

In Chapter 9 we saw how to test hypotheses about the difference between two population means μ_1 and μ_2. In some practical situations the difference between the population standard deviations σ_1 and σ_2 is also of interest. Standard deviation measures the variability of a random variable. For example, if the random variable measures the size of a machined part in a manufacturing process, the size of standard deviation is one indicator of product quality. A smaller standard deviation among items produced in the manufacturing process is desirable since it indicates consistency in product quality.

For theoretical reasons it is easier to compare the squares of the population standard deviations, the population variances σ_1^2 and σ_2^2. This is not a problem, since $\sigma_1 = \sigma_2$ precisely when $\sigma_1^2 = \sigma_2^2$, $\sigma_1 < \sigma_2$ precisely when $\sigma_1^2 < \sigma_2^2$, and $\sigma_1 > \sigma_2$ precisely when $\sigma_1^2 > \sigma_2^2$.

The null hypothesis always has the form $H_0 : \sigma_1^2 = \sigma_2^2$. The three forms of the alternative hypothesis, with the terminology for each case, are:

Form of H_a	Terminology
$H_a : \sigma_1^2 > \sigma_2^2$	Right-tailed
$H_a : \sigma_1^2 < \sigma_2^2$	Left-tailed
$H_a : \sigma_1^2 \neq \sigma_2^2$	Two-tailed

Just as when we test hypotheses concerning two population means, we take a random sample from each population, of sizes n_1 and n_2, and compute the sample standard deviations s_1 and s_2. In this context the samples are always independent. The populations themselves must be normally distributed.

Test Statistic for Hypothesis Tests Concerning the Difference Between Two Population Variances

$$F = \frac{s_1^2}{s_2^2}$$

If the two populations are normally distributed and if $H_0 : \sigma_1^2 = \sigma_2^2$ is true then under independent sampling F approximately follows an F-distribution with degrees of freedom $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$.

A test based on the test statistic F is called an *F-test*.

A most important point is that while the rejection region for a right-tailed test is exactly as in every other situation that we have encountered, because of the asymmetry in the F-distribution the critical value for a left-tailed test and the lower critical value for a two-tailed test have the special forms shown in the following table:

Terminology	Alternative Hypothesis	Rejection Region
Right-tailed	$H_a : \sigma_1^2 > \sigma_2^2$	$F \geq F_\alpha$
Left-tailed	$H_a : \sigma_1^2 < \sigma_2^2$	$F \leq F_{1-\alpha}$
Two-tailed	$H_a : \sigma_1^2 \neq \sigma_2^2$	$F \leq F_{1-\alpha/2}$ or $F \geq F_{\alpha/2}$

Figure 11.3.3 illustrates these rejection regions.

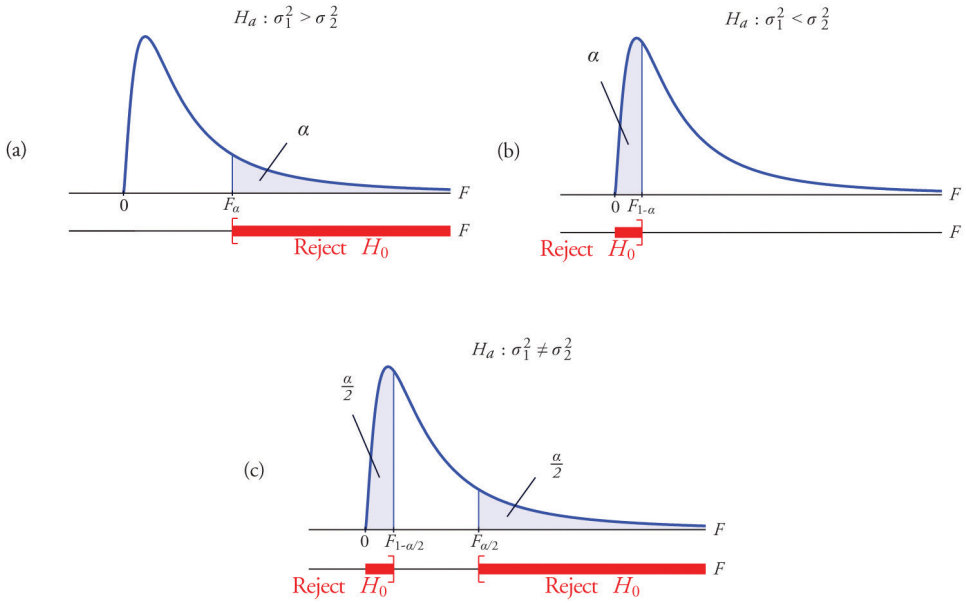


Figure 11.3.3: Rejection Regions: (a) Right-Tailed; (b) Left-Tailed; (c) Two-Tailed

The test is performed using the usual five-step procedure described at the end of Section 8.1.

Example 11.3.4

One of the quality measures of blood glucose meter strips is the consistency of the test results on the same sample of blood. The consistency is measured by the variance of the readings in repeated testing. Suppose two types of strips, A and B, are compared for their respective consistencies. We arbitrarily label the population of Type A strips Population 1 and the population of Type B strips Population 2. Suppose 15 Type A strips were tested with blood drops from a well-shaken vial and 20 Type B strips were tested with the blood from the same vial. The results are summarized in Table 11.3.3. Assume the glucose readings using Type A strips follow a normal distribution with variance $\sigma_1^2$and those using Type B strips follow a normal distribution with variance with σ_2^2. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the consistencies of the two types of strips are different.

Table 11.3.3: Two Types of Test Strips

Strip Type	Sample Size	Sample Variance
A	$n_1 = 16$	$s_1^2 = 2.09$
B	$n_2 = 21$	$s_2^2 = 1.10$

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$vs.$$
$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ @ } \alpha = 0.10$$

- **Step 2**. The distribution is the F-distribution with degrees of freedom $df_1 = 16 - 1 = 15$ and $df_2 = 21 - 1 = 20$.
- **Step 3**. The test is two-tailed. The left or lower critical value is $F_{1-\alpha} = F_{0.95} = 0.43$. The right or upper critical value is $F_{\alpha/2} = F_{0.05} = 2.20$. Thus the rejection region is $[0, -0.43] \cup [2.20, \infty)$, as illustrated in Figure 11.3.4.

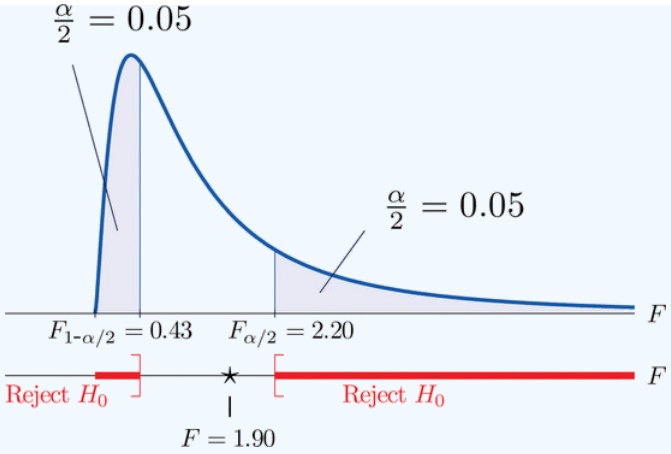


Figure 11.3.4: Rejection Region and Test Statistic for "Example 11.3.4"

- **Step 4**. The value of the test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{2.09}{1.10} = 1.90$$

- **Step 5**. As shown in Figure 11.3.4, the test statistic 1.90 does not lie in the rejection region, so the decision is not to reject H_0. The data do not provide sufficient evidence, at the 10% level of significance, to conclude that there is a difference in the consistency, as measured by the variance, of the two types of test strips.

Example 11.3.5

In the context of "Example 11.3.4", suppose Type A test strips are the current market leader and Type B test strips are a newly improved version of Type A. Test, at the 10% level of significance, whether the data given in Table 11.3.3 provide sufficient evidence to conclude that Type B test strips have better consistency (lower variance) than Type A test strips.

Solution

- **Step 1**. The test of hypotheses is now

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$vs.$$
$$H_a : \sigma_1^2 > \sigma_2^2 \text{ @ } \alpha = 0.10$$

- **Step 2**. The distribution is the F-distribution with degrees of freedom $df_1 = 16 - 1 = 15$ and $df_2 = 21 - 1 = 20$.
- **Step 3**. The value of the test statistic is

$$F = \frac{s_1^2}{s_2^2} = \frac{2.09}{1.10} = 1.90$$

- **Step 4**. The test is right-tailed. The single critical value is $F_\alpha = F_{0.10} = 1.84$. Thus the rejection region is $[1.84, \infty)$, as illustrated in Figure 11.3.5.

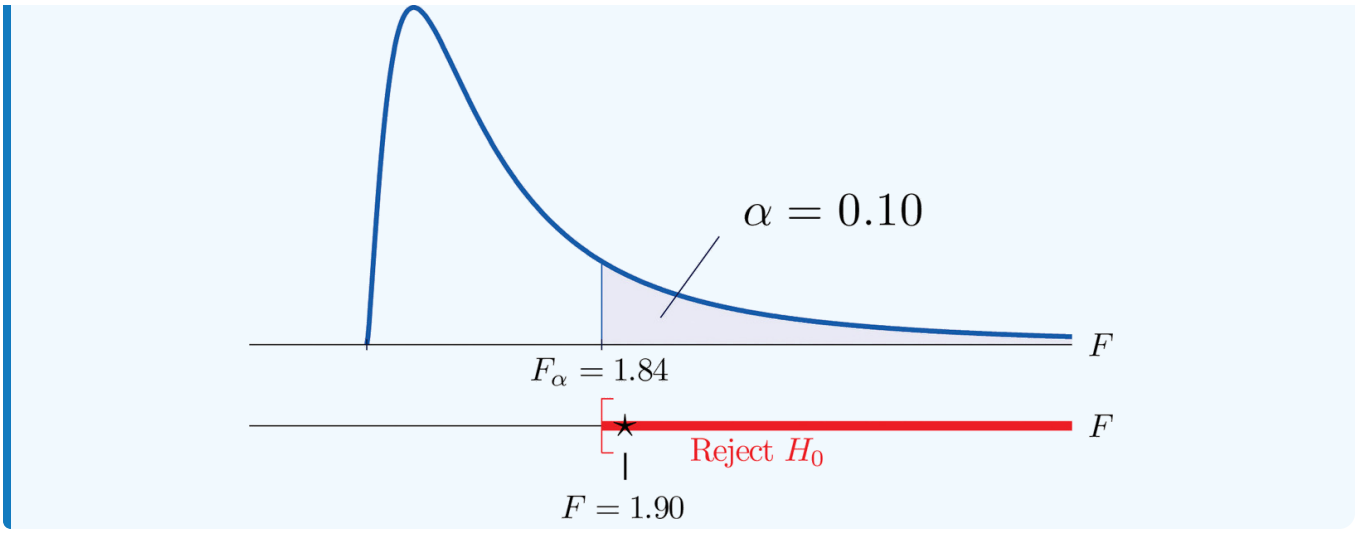
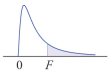


Figure 11.3.5*: Rejection Region and Test Statistic for "Example* 11.3.5*"*

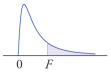
- **Step 5**. As shown in Figure 11.3.5, the test statistic 1.90 lies in the rejection region, so the decision is to reject H_0 The data provide sufficient evidence, at the 10% level of significance, to conclude that Type B test strips have better consistency (lower variance) than Type A test strips do.

Upper Critical Values of F-Distributions



Upper Critical Values of *F*-Distributions

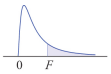
F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24630	24836	25044	25253
0.01	1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6157	6209	6261	6313
0.025	1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1010
0.05	1	161	200	216	225	230	234	237	239	241	242	246	248	250	252
0.10	1	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	61.2	61.7	62.3	62.8
0.005	2	199	199	199	199	199	199	199	199	199	199	199	199	199	199
0.01	2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
0.025	2	38.5	39.0	39.2	39.3	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
0.05	2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
0.10	2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.46	9.47
0.005	3	55.6	49.8	47.5	46.2	45.4	44.9	44.4	44.1	43.9	43.7	43.1	42.8	42.5	42.2
0.01	3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4	27.2	26.9	26.7	26.5	26.3
0.025	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.2	14.1	14.0
0.05	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.57
0.10	3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.15
0.005	4	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.4	20.2	19.9	19.6
0.01	4	21.2	18.0	16.8	16.0	15.5	15.2	15.0	14.8	14.7	14.6	14.2	14.0	13.9	13.7
0.025	4	12.2	10.7	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.36
0.05	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.69
0.10	4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.82	3.79
0.005	5	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.2	12.9	12.7	12.4
0.01	5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38	9.20
0.025	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.12
0.05	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.43
0.10	5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.21	3.17	3.14
0.005	6	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.3	9.81	9.59	9.36	9.12
0.01	6	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.06
0.025	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	4.96
0.05	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.74
0.10	6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.80	2.76
0.005	7	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	7.97	7.75	7.53	7.31
0.01	7	12.3	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.82
0.025	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.25
0.05	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.30
0.10	7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.56	2.51



Upper Critical Values of *F*-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	8	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	6.81	6.61	6.40	6.18
0.01	8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.03
0.025	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.78
0.05	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.01
0.10	8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.38	2.34
0.005	9	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.03	5.83	5.62	5.41
0.01	9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.48
0.025	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.45
0.05	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.79
0.10	9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.25	2.21
0.005	10	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.47	5.27	5.07	4.86
0.01	10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.08
0.025	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.20
0.05	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.62
0.10	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.16	2.11
0.005	11	12.2	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.05	4.86	4.65	4.45
0.01	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.78
0.025	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.12	3.00
0.05	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.49
0.10	11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.08	2.03
0.005	12	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.72	4.53	4.33	4.12
0.01	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.54
0.025	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	2.96	2.85
0.05	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.38
0.10	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.01	1.96
0.005	13	11.4	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.46	4.27	4.07	3.87
0.01	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.34
0.025	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.84	2.72
0.05	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.30
0.10	13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.96	1.90
0.005	14	11.1	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.25	4.06	3.86	3.66
0.01	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.18
0.025	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.73	2.61
0.05	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.22
0.10	14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.91	1.86

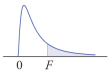
https://stats.libretexts.org/@go/page/515



Upper Critical Values of *F*-Distributions

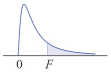
F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.005	15	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.07	3.88	3.69	3.48
0.01	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.05
0.025	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.52
0.05	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.16
0.10	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.87	1.82
0.005	20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.50	3.32	3.12	2.92
0.01	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.61
0.025	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.22
0.05	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.95
0.10	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.74	1.68
0.005	30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.01	2.82	2.63	2.42
0.01	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.21
0.025	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	1.94
0.05	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.74
0.10	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.61	1.54
0.005	40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.78	2.60	2.40	2.18
0.01	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.02
0.025	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.80
0.05	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.64
0.10	40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.54	1.47
0.005	50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.65	2.47	2.27	2.05
0.01	50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.27	2.10	1.91
0.025	50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.99	1.87	1.72
0.05	50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.58
0.10	50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63	1.57	1.50	1.42
0.005	60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.57	2.39	2.19	1.96
0.01	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.84
0.025	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.67
0.05	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.53
0.10	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.48	1.40
0.005	100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.41	2.23	2.02	1.79
0.01	100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.69
0.025	100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.71	1.56
0.05	100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.45
0.10	100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56	1.49	1.42	1.34

Lower Critical Values of F-Distributions



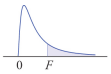
Lower Critical Values of F-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	1	0.03	0.12	0.18	0.22	0.25	0.26	0.28	0.29	0.30	0.30	0.33	0.34	0.35	0.36
0.95	1	0.01	0.05	0.10	0.13	0.15	0.17	0.18	0.19	0.20	0.20	0.22	0.23	0.24	0.25
0.975	1	0.00	0.03	0.06	0.08	0.10	0.11	0.12	0.13	0.14	0.14	0.16	0.17	0.18	0.19
0.99	1	0.00	0.01	0.03	0.05	0.06	0.07	0.08	0.09	0.09	0.10	0.12	0.12	0.13	0.14
0.995	1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12
0.90	2	0.02	0.11	0.18	0.23	0.26	0.29	0.31	0.32	0.33	0.34	0.37	0.39	0.40	0.42
0.95	2	0.01	0.05	0.10	0.14	0.17	0.19	0.21	0.22	0.23	0.24	0.27	0.29	0.30	0.32
0.975	2	0.00	0.03	0.06	0.09	0.12	0.14	0.15	0.17	0.17	0.18	0.21	0.22	0.24	0.25
0.99	2	0.00	0.01	0.03	0.06	0.08	0.09	0.10	0.12	0.12	0.13	0.16	0.17	0.19	0.20
0.995	2	0.00	0.01	0.02	0.04	0.05	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.17
0.90	3	0.02	0.11	0.19	0.24	0.28	0.30	0.33	0.34	0.36	0.37	0.40	0.42	0.44	0.46
0.95	3	0.00	0.05	0.11	0.15	0.18	0.21	0.23	0.25	0.26	0.27	0.30	0.32	0.34	0.36
0.975	3	0.00	0.03	0.06	0.10	0.13	0.15	0.17	0.18	0.20	0.21	0.24	0.26	0.28	0.30
0.99	3	0.00	0.01	0.03	0.06	0.08	0.10	0.12	0.13	0.14	0.15	0.18	0.20	0.22	0.24
0.995	3	0.00	0.01	0.02	0.04	0.06	0.08	0.09	0.10	0.11	0.12	0.15	0.17	0.19	0.21
0.90	4	0.02	0.11	0.19	0.24	0.28	0.31	0.34	0.36	0.37	0.38	0.42	0.44	0.47	0.49
0.95	4	0.00	0.05	0.11	0.16	0.19	0.22	0.24	0.26	0.28	0.29	0.33	0.35	0.37	0.40
0.975	4	0.00	0.03	0.07	0.10	0.14	0.16	0.18	0.20	0.21	0.22	0.26	0.28	0.31	0.33
0.99	4	0.00	0.01	0.03	0.06	0.09	0.11	0.13	0.14	0.16	0.17	0.20	0.23	0.25	0.27
0.995	4	0.00	0.01	0.02	0.04	0.06	0.08	0.10	0.11	0.13	0.14	0.17	0.19	0.22	0.24
0.90	5	0.02	0.11	0.19	0.25	0.29	0.32	0.35	0.37	0.38	0.40	0.44	0.46	0.49	0.51
0.95	5	0.00	0.05	0.11	0.16	0.20	0.23	0.25	0.27	0.29	0.30	0.34	0.37	0.39	0.42
0.975	5	0.00	0.03	0.07	0.11	0.14	0.17	0.19	0.21	0.22	0.24	0.28	0.30	0.33	0.36
0.99	5	0.00	0.01	0.04	0.06	0.09	0.11	0.13	0.15	0.17	0.18	0.22	0.24	0.27	0.30
0.995	5	0.00	0.01	0.02	0.04	0.07	0.09	0.11	0.12	0.13	0.15	0.19	0.21	0.24	0.27
0.90	6	0.02	0.11	0.19	0.25	0.29	0.33	0.35	0.37	0.39	0.41	0.45	0.48	0.50	0.53
0.95	6	0.00	0.05	0.11	0.16	0.20	0.23	0.26	0.28	0.30	0.31	0.36	0.38	0.41	0.44
0.975	6	0.00	0.03	0.07	0.11	0.14	0.17	0.20	0.21	0.23	0.25	0.29	0.32	0.35	0.38
0.99	6	0.00	0.01	0.04	0.07	0.09	0.12	0.14	0.16	0.17	0.19	0.23	0.26	0.29	0.32
0.995	6	0.00	0.01	0.02	0.05	0.07	0.09	0.11	0.13	0.14	0.15	0.2	00.22	0.25	0.29
0.90	7	0.02	0.11	0.19	0.25	0.30	0.33	0.36	0.38	0.40	0.41	0.46	0.49	0.52	0.55
0.95	7	0.00	0.05	0.11	0.16	0.21	0.24	0.26	0.29	0.30	0.32	0.37	0.40	0.43	0.46
0.975	7	0.00	0.03	0.07	0.11	0.15	0.18	0.20	0.22	0.24	0.25	0.30	0.33	0.36	0.40
0.99	7	0.00	0.01	0.04	0.07	0.10	0.12	0.14	0.16	0.18	0.19	0.24	0.27	0.30	0.34
0.995	7	0.00	0.01	0.02	0.05	0.07	0.09	0.11	0.13	0.15	0.16	0.21	0.23	0.27	0.3



Lower Critical Values of F-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	8	0.02	0.11	0.19	0.25	0.30	0.34	0.36	0.39	0.40	0.42	0.47	0.50	0.53	0.56
0.95	8	0.00	0.05	0.11	0.17	0.21	0.24	0.27	0.29	0.31	0.33	0.38	0.41	0.44	0.48
0.975	8	0.00	0.03	0.07	0.11	0.15	0.18	0.20	0.23	0.24	0.26	0.31	0.34	0.38	0.41
0.99	8	0.00	0.01	0.04	0.07	0.10	0.12	0.15	0.17	0.18	0.20	0.25	0.28	0.32	0.35
0.995	8	0.00	0.01	0.02	0.05	0.07	0.09	0.12	0.13	0.15	0.16	0.21	0.24	0.28	0.32
0.90	9	0.02	0.11	0.19	0.25	0.30	0.34	0.37	0.39	0.41	0.43	0.48	0.51	0.54	0.58
0.95	9	0.00	0.05	0.11	0.17	0.21	0.24	0.27	0.30	0.31	0.33	0.39	0.42	0.45	0.49
0.975	9	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.23	0.25	0.26	0.32	0.35	0.39	0.43
0.99	9	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.20	0.26	0.29	0.33	0.37
0.995	9	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.15	0.17	0.22	0.25	0.29	0.33
0.90	10	0.02	0.11	0.19	0.26	0.30	0.34	0.37	0.39	0.41	0.43	0.49	0.52	0.55	0.59
0.95	10	0.00	0.05	0.11	0.17	0.21	0.25	0.27	0.30	0.32	0.34	0.39	0.43	0.46	0.50
0.975	10	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.23	0.25	0.27	0.33	0.36	0.40	0.44
0.99	10	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.21	0.26	0.30	0.34	0.38
0.995	10	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.17	0.23	0.26	0.30	0.34
0.90	11	0.02	0.11	0.19	0.26	0.30	0.34	0.37	0.40	0.42	0.43	0.49	0.52	0.56	0.60
0.95	11	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.30	0.32	0.34	0.40	0.43	0.47	0.51
0.975	11	0.00	0.03	0.07	0.11	0.15	0.18	0.21	0.24	0.26	0.27	0.33	0.37	0.41	0.45
0.99	11	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.17	0.19	0.21	0.27	0.30	0.34	0.39
0.995	11	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.17	0.23	0.27	0.31	0.36
0.90	12	0.02	0.11	0.19	0.26	0.31	0.34	0.37	0.40	0.42	0.44	0.50	0.53	0.56	0.60
0.95	12	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.30	0.33	0.34	0.40	0.44	0.48	0.52
0.975	12	0.00	0.03	0.07	0.11	0.15	0.19	0.21	0.24	0.26	0.28	0.34	0.37	0.41	0.46
0.99	12	0.00	0.01	0.04	0.07	0.10	0.13	0.15	0.18	0.20	0.21	0.27	0.31	0.35	0.40
0.995	12	0.00	0.01	0.02	0.05	0.07	0.10	0.12	0.14	0.16	0.18	0.24	0.27	0.31	0.36
0.90	13	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.40	0.42	0.44	0.50	0.53	0.57	0.61
0.95	13	0.00	0.05	0.11	0.17	0.21	0.25	0.28	0.31	0.33	0.35	0.41	0.44	0.48	0.53
0.975	13	0.00	0.03	0.07	0.11	0.15	0.19	0.22	0.24	0.26	0.28	0.34	0.38	0.42	0.47
0.99	13	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.31	0.36	0.41
0.995	13	0.00	0.01	0.02	0.05	0.08	0.10	0.12	0.14	0.16	0.18	0.24	0.28	0.32	0.37
0.90	14	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.40	0.43	0.44	0.50	0.54	0.58	0.62
0.95	14	0.00	0.05	0.11	0.17	0.22	0.25	0.28	0.31	0.33	0.35	0.41	0.45	0.49	0.54
0.975	14	0.00	0.03	0.07	0.12	0.15	0.19	0.22	0.24	0.26	0.28	0.35	0.38	0.43	0.48
0.99	14	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.32	0.36	0.42
0.995	14	0.00	0.01	0.02	0.05	0.08	0.10	0.12	0.15	0.16	0.18	0.24	0.28	0.33	0.38



Lower Critical Values of *F*-Distributions

F tail area	df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	60
0.90	15	0.02	0.11	0.19	0.26	0.31	0.35	0.38	0.41	0.43	0.45	0.51	0.54	0.58	0.62
0.95	15	0.00	0.05	0.11	0.17	0.22	0.25	0.28	0.31	0.33	0.35	0.42	0.45	0.50	0.54
0.975	15	0.00	0.03	0.07	0.12	0.16	0.19	0.22	0.24	0.27	0.28	0.35	0.39	0.43	0.49
0.99	15	0.00	0.01	0.04	0.07	0.10	0.13	0.16	0.18	0.20	0.22	0.28	0.32	0.37	0.43
0.995	15	0.00	0.01	0.02	0.05	0.08	0.10	0.13	0.15	0.17	0.18	0.25	0.29	0.33	0.39
0.90	20	0.02	0.11	0.19	0.26	0.31	0.35	0.39	0.41	0.44	0.45	0.52	0.56	0.60	0.65
0.95	20	0.00	0.05	0.12	0.17	0.22	0.26	0.29	0.32	0.34	0.36	0.43	0.47	0.52	0.57
0.975	20	0.00	0.03	0.07	0.12	0.16	0.19	0.22	0.25	0.27	0.29	0.36	0.41	0.46	0.51
0.99	20	0.00	0.01	0.04	0.07	0.10	0.14	0.16	0.19	0.21	0.23	0.30	0.34	0.39	0.45
0.995	20	0.00	0.01	0.02	0.05	0.08	0.1	00.13	0.15	0.17	0.19	0.26	0.30	0.35	0.42
0.90	30	0.02	0.11	0.19	0.26	0.32	0.36	0.39	0.42	0.44	0.46	0.53	0.58	0.62	0.68
0.95	30	0.00	0.05	0.12	0.17	0.22	0.26	0.30	0.32	0.35	0.37	0.45	0.49	0.54	0.61
0.975	30	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.28	0.30	0.38	0.43	0.48	0.55
0.99	30	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.19	0.22	0.24	0.31	0.36	0.42	0.49
0.995	30	0.00	0.01	0.02	0.05	0.08	0.11	0.13	0.16	0.18	0.20	0.27	0.32	0.38	0.46
0.90	40	0.02	0.11	0.19	0.26	0.32	0.36	0.39	0.42	0.45	0.47	0.54	0.59	0.64	0.70
0.95	40	0.00	0.05	0.12	0.17	0.22	0.26	0.30	0.33	0.35	0.38	0.45	0.50	0.56	0.63
0.975	40	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.29	0.31	0.39	0.44	0.50	0.57
0.99	40	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.32	0.37	0.43	0.52
0.995	40	0.00	0.01	0.02	0.05	0.08	0.11	0.13	0.16	0.18	0.20	0.28	0.33	0.40	0.48
0.90	50	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.45	0.47	0.55	0.59	0.64	0.71
0.95	50	0.00	0.05	0.12	0.18	0.23	0.27	0.30	0.33	0.36	0.38	0.46	0.51	0.57	0.64
0.975	50	0.00	0.03	0.07	0.12	0.16	0.20	0.23	0.26	0.29	0.31	0.39	0.44	0.51	0.59
0.99	50	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.32	0.38	0.45	0.53
0.995	50	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.18	0.20	0.28	0.34	0.41	0.50
0.90	60	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.45	0.47	0.55	0.60	0.65	0.72
0.95	60	0.00	0.05	0.12	0.18	0.23	0.27	0.30	0.33	0.36	0.38	0.46	0.51	0.57	0.65
0.975	60	0.00	0.03	0.07	0.12	0.16	0.20	0.24	0.26	0.29	0.31	0.40	0.45	0.52	0.60
0.99	60	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.22	0.24	0.33	0.38	0.45	0.54
0.995	60	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.18	0.21	0.29	0.34	0.41	0.51
0.90	100	0.02	0.11	0.19	0.26	0.32	0.36	0.40	0.43	0.46	0.48	0.56	0.61	0.66	0.74
0.95	100	0.00	0.05	0.12	0.18	0.23	0.27	0.31	0.34	0.36	0.39	0.47	0.52	0.59	0.68
0.975	100	0.00	0.03	0.07	0.12	0.16	0.20	0.24	0.27	0.29	0.32	0.40	0.46	0.53	0.63
0.99	100	0.00	0.01	0.04	0.07	0.11	0.14	0.17	0.20	0.23	0.25	0.34	0.39	0.47	0.57
0.995	100	0.00	0.01	0.02	0.05	0.08	0.11	0.14	0.16	0.19	0.21	0.29	0.35	0.43	0.54

Key Takeaway

- Critical values of an F-distribution with degrees of freedom df_1 and df_2 are found in tables above.
- An F-test can be used to evaluate the hypothesis of two identical normal population variances.

This page titled 11.3: F-tests for Equality of Two Variances is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

11.4: F-Tests in One-Way ANOVA

Learning Objectives

- To understand how to use an F-test to judge whether several population means are all equal

In Chapter 9, we saw how to compare two population means μ_1 and μ_2.

In this section we will learn to compare three or more population means at the same time, which is often of interest in practical applications. For example, an administrator at a university may be interested in knowing whether student grade point averages are the same for different majors. In another example, an oncologist may be interested in knowing whether patients with the same type of cancer have the same average survival times under several different competing cancer treatments.

In general, suppose there are K normal populations with possibly different means, $\mu_1, \mu_2, \ldots, \mu_K$, but all with the same variance σ^2. The study question is whether all the K population means are the same. We formulate this question as the test of hypotheses

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$vs.$$
$$H_a : \text{not all K population means are equal}$$

To perform the test K independent random samples are taken from the K normal populations. The K sample means, the K sample variances, and the K sample sizes are summarized in the table:

Population	Sample Size	Sample Mean	Sample Variance
1	n_1	$\bar{x_1}$	s_1^2
2	n_2	$\bar{x_2}$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	n_K	$\bar{x_K}$	s_K^2

Define the following quantities:

Definitions

The *combined sample size*:

$$n = n_1 + n_2 + \ldots + n_K$$

The *mean of the combined sample* of all n observations:

$$x = \frac{\Sigma x}{n} = \frac{n_1\overline{x} + n_2\overline{x}_2 + \ \ldots \ + n_K\overline{x}_K}{n}$$

The *mean square for treatment:*

$$MST = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots \ + n_K(\overline{x}K - \overline{x})^2}{K - 1}$$

The *mean square for error:*

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ \ldots \ + (n_{K-1})s_K^2}{n - K}$$

MST can be thought of as the variance between the K individual independent random samples and MSE as the variance within the samples. This is the reason for the name “analysis of variance,” universally abbreviated ANOVA. The adjective “one-way” has to do with the fact that the sampling scheme is the simplest possible, that of taking one random sample from each population under consideration. If the means of the K populations are all the same then the two quantities MST and MSE should be close to the same, so the null hypothesis will be rejected if the ratio of these two quantities is significantly greater than 1. This yields the following test statistic and methods and conditions for its use.

Test Statistic for Testing the Null Hypothesis that K Population Means Are Equal

$$F = \frac{MST}{MSE}$$

If the K populations are normally distributed with a common variance and if $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is true then under independent random sampling F approximately follows an F-distribution with degrees of freedom $df_1 = K - 1$ and $df_2 = n - K$.

The test is right-tailed: H_0 is rejected at level of significance α if $F \geq F_\alpha$.

As always the test is performed using the usual five-step procedure.

Example 11.4.1

The average of grade point averages (GPAs) of college courses in a specific major is a measure of difficulty of the major. An educator wishes to conduct a study to find out whether the difficulty levels of different majors are the same. For such a study, a random sample of major grade point averages (GPA) of 11 graduating seniors at a large university is selected for each of the four majors mathematics, English, education, and biology. The data are given in Table 11.4.1. Test, at the 5% level of significance, whether the data contain sufficient evidence to conclude that there are differences among the average major GPAs of these four majors.

Table 11.4.1: Difficulty Levels of College Majors

Mathematics	English	Education	Biology
2.59	3.64	4.00	2.78
3.13	3.19	3.59	3.51
2.97	3.15	2.80	2.65
2.50	3.78	2.39	3.16
2.53	3.03	3.47	2.94
3.29	2.61	3.59	2.32
2.53	3.20	3.74	2.58
3.17	3.30	3.77	3.21
2.70	3.54	3.13	3.23
3.88	3.25	3.00	3.57
2.64	4.00	3.47	3.22

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$vs.$$
$$H_a : \text{not all four population means are equal @ } \alpha = 0.05$$

- **Step 2**. The test statistic is $F = MST/MSE$ with (since $n = 44$ and $K = 4$) degrees of freedom $df_1 = K - 1 = 4 - 1 = 3$ and $df_2 = n - K = 44 - 4 = 40$.
- **Step 3**. If we index the population of mathematics majors by 1, English majors by 2, education majors by 3, and biology majors by 4, then the sample sizes, sample means, and sample variances of the four samples in Table 11.4.1 are summarized (after rounding for simplicity) by:

Major	Sample Size	Sample Mean	Sample Variance

Major	Sample Size	Sample Mean	Sample Variance
Mathematics	$n_1 = 11$	$\bar{x}_1 = 2.90$	$s_1^2 = 0.188$
English	$n_2 = 11$	$\bar{x}_2 = 3.34$	$s_2^2 = 0.148$
Education	$n_3 = 11$	$\bar{x}_3 = 3.36$	$s_3^2 = 0.229$
Biology	$n_4 = 11$	$\bar{x}_4 = 3.02$	$s_4^2 = 0.157$

The average of all 44 observations is (after rounding for simplicity) $\overline{x} = 3.15$. We compute (rounding for simplicity)

$$\begin{aligned} MST &= \frac{11(2.90-3.15)^2 + 11(3.34-3.15)^2 + 11(3.36-3.15)^2 + 11(3.02-3.15)^2}{4-1} \\ &= \frac{1.7556}{3} \\ &= 0.585 \end{aligned}$$

and

$$\begin{aligned} MSE &= \frac{(11-1)(0.188) + (11-1)(0.148) + (11-1)(0.229) + (11-1)(0.157)}{44-4} \\ &= \frac{7.22}{40} \\ &= 0.181 \end{aligned}$$

so that

$$F = \frac{MST}{MSE} = \frac{0.585}{0.181} = 3.232$$

- **Step 4**. The test is right-tailed. The single critical value is (since $df_1 = 3$ and $df_2 = 40$) $F_\alpha = F_{0.05} = 2.84$. Thus the rejection region is $[2.84, \infty)$, as illustrated in Figure 11.4.1.

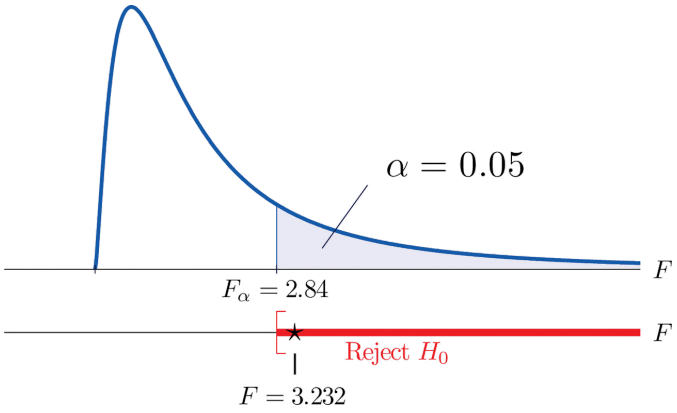


Figure 11.4.1: Rejection Region

- **Step 5**. Since $F = 3.232 > 2.84$, we reject H_0. The data provide sufficient evidence, at the 5% level of significance, to conclude that the averages of major GPAs for the four majors considered are not all equal.

Example 11.4.2: Mice Survival Times

A research laboratory developed two treatments which are believed to have the potential of prolonging the survival times of patients with an acute form of thymic leukemia. To evaluate the potential treatment effects 33 laboratory mice with thymic leukemia were randomly divided into three groups. One group received Treatment 1, one received Treatment 2, and the third was observed as a control group. The survival times of these mice are given in Table 11.4.2. Test, at the 1% level of significance, whether these data provide sufficient evidence to confirm the belief that at least one of the two treatments affects the average survival time of mice with thymic leukemia.

Table 11.4.2 Mice Survival Times in Days

Treatment 1		Treatment 2	Control
71	75	77	81
72	73	67	79
75	72	79	73
80	65	78	71
60	63	81	75
65	69	72	84
63	64	71	77
78	71	84	67
		91	

Solution

- **Step 1**. The test of hypotheses is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$vs.$$
$$H_a : \text{not all three population means are equal @ } \alpha = 0.01$$

- **Step 2**. The test statistic is $F = \dfrac{MST}{MSE}$ with (since $n = 33$ and $K = 3$) degrees of freedom $df_1 = K - 1 = 3 - 1 = 2$ and $df_2 = n - K = 33 - 3 = 30$.
- **Step 3**. If we index the population of mice receiving Treatment 1 by 1, Treatment 2 by 2, and no treatment by 3, then the sample sizes, sample means, and sample variances of the three samples in Table 11.4.2 are summarized (after rounding for simplicity) by:

Table 11.4.2: Mice Survival Times in Days

Group	Sample Size	Sample Mean	Sample Variance
Treatment 1	$n_1 = 16$	$\bar{x}_1 = 69.75$	$s_1^2 = 34.47$
Treatment 2	$n_2 = 9$	$\bar{x}_2 = 77.78$	$s_2^2 = 52.69$
Control	$n_3 = 8$	$\bar{x}_3 = 75.88$	$s_3^2 = 30.69$

The average of all 33 observations is (after rounding for simplicity) $\overline{x} = 73.42$. We compute (rounding for simplicity)

$$\begin{aligned} MST &= \frac{16(69.75 - 73.42)^2 + 9(77.78 - 73.42)^2 + 8(75.88 - 73.42)^2}{31} \\ &= \frac{434.63}{2} \\ &= 217.50 \end{aligned}$$

and

$$\begin{aligned} MSE &= \frac{(16 - 1)(34.47) + (9 - 1)(52.69) + (8 - 1)(30.69)}{33 - 3} \\ &= \frac{1153.4}{30} \\ &= 38.45 \end{aligned}$$

so that

$$F = \frac{MST}{MSE} = \frac{217.50}{38.45} = 5.65$$

- **Step 4**. The test is right-tailed. The single critical value is $F_\alpha = F_{0.01} = 5.39$. Thus the rejection region is $[5.39, \infty)$, as illustrated in Figure 11.4.2.

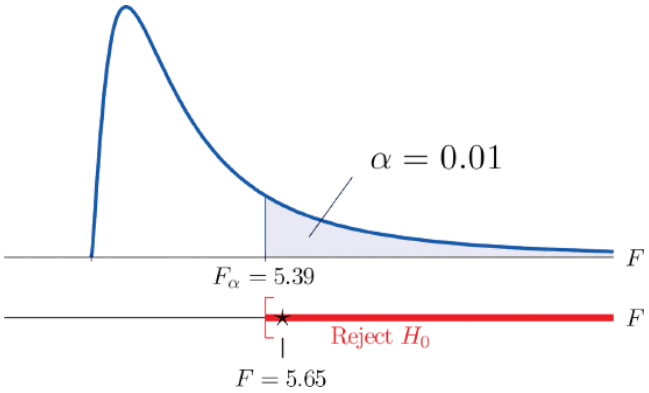


Figure 11.4.2: Rejection Region

- **Step 5**. Since $F = 5.65 > 5.39$, we reject H_0. The data provide sufficient evidence, at the 1% level of significance, to conclude that a treatment effect exists at least for one of the two treatments in increasing the mean survival time of mice with thymic leukemia.

It is important to to note that, if the null hypothesis of equal population means is rejected, the statistical implication is that not all population means are equal. It does not however tell which population mean is different from which. The inference about where the suggested difference lies is most frequently made by a follow-up study.

Key Takeaway

- An F-test can be used to evaluate the hypothesis that the means of several normal populations, all with the same standard deviation, are identical.

This page titled 11.4: F-Tests in One-Way ANOVA is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

11.E: Chi-Square Tests and F-Tests (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by Shafer and Zhang.

11.1: Chi-Square Tests for Independence

Basic

Q11.1.1

Find $\chi^2_{0.01}$ for each of the following number of degrees of freedom.

a. $df = 5$
b. $df = 11$
c. $df = 25$

Q11.1.2

Find $\chi^2_{0.05}$ for each of the following number of degrees of freedom.

a. $df = 6$
b. $df = 12$
c. $df = 30$

Q11.1.3

Find $\chi^2_{0.10}$ for each of the following number of degrees of freedom.

a. $df = 6$
b. $df = 12$
c. $df = 30$

Q11.1.4

Find $\chi^2_{0.01}$ for each of the following number of degrees of freedom.

a. $df = 7$
b. $df = 10$
c. $df = 20$

Q11.1.5

For $df = 7$ and $\alpha = 0.05$

a. χ^2_{α}
b. $\chi^2_{\frac{\alpha}{2}}$

Q11.1.6

For $df = 17$ and $\alpha = 0.01$

a. χ^2_{α}
b. $\chi^2_{\frac{\alpha}{2}}$

Q11.1.7

A data sample is sorted into a 2×2 contingency table based on two factors, each of which has two levels.

		Factor 1		
		Level 1	Level 2	Row Total
Factor 2	Level 1	20	10	R
	Level 2	15	5	R
Column Total		C	C	n

a. Find the column totals, the row totals, and the grand total, n, of the table.
b. Find the expected number E of observations for each cell based on the assumption that the two factors are independent (that is, just use the formula $E = (R \times C)/n$).
c. Find the value of the chi-square test statistic χ^2.
d. Find the number of degrees of freedom of the chi-square test statistic.

Q11.1.8

A data sample is sorted into a 3×2 contingency table based on two factors, one of which has three levels and the other of which has two levels.

		Factor 1		
		Level 1	Level 2	Row Total
Factor 2	Level 1	20	10	R
	Level 2	15	5	R
	Level 3	10	20	R
Column Total		C	C	n

a. Find the column totals, the row totals, and the grand total, n, of the table.
b. Find the expected number E of observations for each cell based on the assumption that the two factors are independent (that is, just use the formula $E = (R \times C)/n$).
c. Find the value of the chi-square test statistic χ^2.
d. Find the number of degrees of freedom of the chi-square test statistic.

Applications

Q11.1.9

A child psychologist believes that children perform better on tests when they are given perceived freedom of choice. To test this belief, the psychologist carried out an experiment in which 200 third graders were randomly assigned to two groups, A and B. Each child was given the same simple logic test. However in group B, each child was given the freedom to choose a text booklet from many with various drawings on the covers. The performance of each child was rated as Very Good, Good, and Fair. The results are summarized in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to support the psychologist's belief.

		Group	
		A	B
Performance	Very Good	32	29
	Good	55	61
	Fair	10	13

Q11.1.10

In regard to wine tasting competitions, many experts claim that the first glass of wine served sets a reference taste and that a different reference wine may alter the relative ranking of the other wines in competition. To test this claim, three wines, A, B and C, were served at a wine tasting event. Each person was served a single glass of each wine, but in different orders for different guests. At the close, each person was asked to name the best of the three. One hundred seventy-two people were at the event and their top picks are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to support the claim that wine experts' preference is dependent on the first served wine.

	Top Pick

		Top Pick		
		A	*B*	*C*
First Glass	*A*	12	31	27
	B	15	40	21
	C	10	9	7

9. Is being left-handed hereditary? To answer this question, 250 adults are randomly selected and their handedness and their parents' handedness are noted. The results are summarized in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that there is a hereditary element in handedness.

		Number of Parents Left-Handed		
		0	1	2
Handedness	Left	8	10	12
	Right	178	21	21

10. Some geneticists claim that the genes that determine left-handedness also govern development of the language centers of the brain. If this claim is true, then it would be reasonable to expect that left-handed people tend to have stronger language abilities. A study designed to text this claim randomly selected 807 students who took the Graduate Record Examination (GRE). Their scores on the language portion of the examination were classified into three categories: *low*, *average*, and *high*, and their handedness was also noted. The results are given in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that left-handed people tend to have stronger language abilities.

		GRE English Scores		
		Low	Average	High
Handedness	Left	18	40	22
	Right	201	360	166

11. It is generally believed that children brought up in stable families tend to do well in school. To verify such a belief, a social scientist examined 290 randomly selected students' records in a public high school and noted each student's family structure and academic status four years after entering high school. The data were then sorted into a 2×3 contingency table with two factors. $\text{Factor } 1$ has two levels: *graduated* and *did not graduate*. $\text{Factor } 2$ has three levels: *no parent*, *one parent*, and *two parents*. The results are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that family structure matters in school performance of the students.

		Academic Status	
		Graduated	Did Not Graduate
Family	No parent	18	31
	One parent	101	44
	Two parents	70	26

12. A large middle school administrator wishes to use celebrity influence to encourage students to make healthier choices in the school cafeteria. The cafeteria is situated at the center of an open space. Everyday at lunch time students get their lunch and a drink in three separate lines leading to three separate serving stations. As an experiment, the school administrator displayed a poster of a popular teen pop star drinking milk at each of the three areas where drinks are provided, except the milk in the poster is different at each location: one shows white milk, one shows strawberry-flavored pink milk, and one shows chocolate milk.

After the first day of the experiment the administrator noted the students' milk choices separately for the three lines. The data are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that the posters had some impact on the students' drink choices.

	Student Choice		
	Regular	Strawberry	Chocolate
Poster Choice			
Regular	38	28	40
Strawberry	18	51	24
Chocolate	32	32	53

Large Data Set Exercise

Large Data Sets not available

15. Large $\text{Data Set } 8$ records the result of a survey of 300 randomly selected adults who go to movie theaters regularly. For each person the gender and preferred type of movie were recorded. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that the factors "gender" and "preferred type of movie" are dependent.

Answers

1. a. 15.09
   b. 24.72
   c. 44.31
2.
3. a. 10.64
   b. 18.55
   c. 40.26
4.
5. a. 14.07
   b. 16.01
6.
7. a. $C_1 = 35,\ C_2 = 15,\ R_1 = 30,\ R_2 = 20,\ n = 50$
   b. $E_{11} = 21,\ E_{12} = 9,\ E_{21} = 14,\ E_{22} = 6$
   c. $\chi^2 = 0.3968$
   d. $df = 1$
8.
9. $\chi^2 = 0.6698,\ \chi^2_{0.05} = 5.99$, do not reject H_0
10.
11. $\chi^2 = 72.35,\ \chi^2_{0.01} = 9.21$, reject H_0
12.
13. $\chi^2 = 21.2784,\ \chi^2_{0.01} = 9.21$, reject H_0
14.
15. $\chi^2 = 28.4539$, $df = 3$, Rejection Region: $[7.815, \infty)$, Decision: reject H_0 of independence

11.2: Chi-Square One-Sample Goodness-of-Fit Tests

Basic

1. A data sample is sorted into five categories with an assumed probability distribution.

Factor Levels	Assumed Distribution	Observed Frequency
1	$p_1 = 0.1$	10

Factor Levels	Assumed Distribution	Observed Frequency
2	$p_2 = 0.4$	35
3	$p_3 = 0.4$	45
4	$p_4 = 0.1$	10

a. Find the size n of the sample.
b. Find the expected number E of observations for each level, if the sampled population has a probability distribution as assumed (that is, just use the formula $E_i = n \times p_i$).
c. Find the chi-square test statistic χ^2.
d. Find the number of degrees of freedom of the chi-square test statistic.

2. A data sample is sorted into five categories with an assumed probability distribution.

Factor Levels	Assumed Distribution	Observed Frequency
1	msub$p_1 = 0.3$	23
2	msub$p_2 = 0.3$	30
3	$p_3 = 0.2$	19
4	$p_4 = 0.1$	8
5	$p_5 = 0.1$	10

a. Find the size n of the sample.
b. Find the expected number E of observations for each level, if the sampled population has a probability distribution as assumed (that is, just use the formula $E_i = n \times p_i$).
c. Find the chi-square test statistic χ^2.
d. Find the number of degrees of freedom of the chi-square test statistic.

Applications

3. Retailers of collectible postage stamps often buy their stamps in large quantities by weight at auctions. The prices the retailers are willing to pay depend on how old the postage stamps are. Many collectible postage stamps at auctions are described by the proportions of stamps issued at various periods in the past. Generally the older the stamps the higher the value. At one particular auction, a lot of collectible stamps is advertised to have the age distribution given in the table provided. A retail buyer took a sample of 73 stamps from the lot and sorted them by age. The results are given in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that the age distribution of the lot is different from what was claimed by the seller.

Year	Claimed Distribution	Observed Frequency
Before 1940	0.10	6
1940 to 1959	0.25	15
1960 to 1979	0.45	30
After 1979	0.20	22

4. The litter size of Bengal tigers is typically two or three cubs, but it can vary between one and four. Based on long-term observations, the litter size of Bengal tigers in the wild has the distribution given in the table provided. A zoologist believes that Bengal tigers in captivity tend to have different (possibly smaller) litter sizes from those in the wild. To verify this belief, the zoologist searched all data sources and found 316 litter size records of Bengal tigers in captivity. The results are given in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that the distribution of litter sizes in captivity differs from that in the wild.

Litter Size	Wild Litter Distribution	Observed Frequency

Litter Size	Wild Litter Distribution	Observed Frequency
1	0.11	41
2	0.69	243
3	0.18	27
4	0.02	5

5. An online shoe retailer sells men's shoes in sizes 8 to 13. In the past orders for the different shoe sizes have followed the distribution given in the table provided. The management believes that recent marketing efforts may have expanded their customer base and, as a result, there may be a shift in the size distribution for future orders. To have a better understanding of its future sales, the shoe seller examined $1,040$ sales records of recent orders and noted the sizes of the shoes ordered. The results are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that the shoe size distribution of future sales will differ from the historic one.

Shoe Size	Past Size Distribution	Recent Size Frequency
8.0	0.03	25
8.5	0.06	43
9.0	0.09	88
9.5	0.19	221
10.0	0.23	272
10.5	0.14	150
11.0	0.10	107
11.5	0.06	51
12.0	0.05	37
12.5	0.03	35
13.0	0.02	11

6. An online shoe retailer sells women's shoes in sizes 5 to 10. In the past orders for the different shoe sizes have followed the distribution given in the table provided. The management believes that recent marketing efforts may have expanded their customer base and, as a result, there may be a shift in the size distribution for future orders. To have a better understanding of its future sales, the shoe seller examined $1,174$ sales records of recent orders and noted the sizes of the shoes ordered. The results are given in the table provided. Test, at the 1% level of significance, whether there is sufficient evidence in the data to conclude that the shoe size distribution of future sales will differ from the historic one.

Shoe Size	Past Size Distribution	Recent Size Frequency
5.0	0.02	20
5.5	0.03	23
6.0	0.07	88
6.5	0.08	90
7.0	0.20	222
7.5	0.20	258
8.0	0.15	177
8.5	0.11	121

Shoe Size	Past Size Distribution	Recent Size Frequency
9.0	0.08	91
9.5	0.04	53
10.0	0.02	31

7. A chess opening is a sequence of moves at the beginning of a chess game. There are many well-studied named openings in chess literature. French Defense is one of the most popular openings for black, although it is considered a relatively weak opening since it gives black probability 0.344 of winning, probability 0.405 of losing, and probability 0.251 of drawing. A chess master believes that he has discovered a new variation of French Defense that may alter the probability distribution of the outcome of the game. In his many Internet chess games in the last two years, he was able to apply the new variation in 77 games. The wins, losses, and draws in the 77 games are given in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that the newly discovered variation of French Defense alters the probability distribution of the result of the game.

Result for Black	Probability Distribution	New Variation Wins
Win	0.344	31
Loss	0.405	25
Draw	0.251	21

8. The Department of Parks and Wildlife stocks a large lake with fish every six years. It is determined that a healthy diversity of fish in the lake should consist of 10% largemouth bass, 15% smallmouth bass, 10% striped bass, 10% trout, and 20% catfish. Therefore each time the lake is stocked, the fish population in the lake is restored to maintain that particular distribution. Every three years, the department conducts a study to see whether the distribution of the fish in the lake has shifted away from the target proportions. In one particular year, a research group from the department observed a sample of 292 fish from the lake with the results given in the table provided. Test, at the 5% level of significance, whether there is sufficient evidence in the data to conclude that the fish population distribution has shifted since the last stocking.

Fish	Target Distribution	Fish in Sample
Largemouth Bass	0.10	14
Smallmouth Bass	0.15	49
Striped Bass	0.10	21
Trout	0.10	22
Catfish	0.20	75
Other	0.35	111

Large Data Set Exercise

Large Data Sets not available

9. Large $\text{Data Set } 4$ records the result of 500 tosses of six-sided die. Test, at the 10% level of significance, whether there is sufficient evidence in the data to conclude that the die is not “fair” (or “balanced”), that is, that the probability distribution differs from probability $1/6$ for each of the six faces on the die.

Answers

S11.2.1

a. $n = 100$
b. $E = 10, E = 40, E = 40, E = 10$
c. $\chi^2 = 1.25$

d. $df = 3$

S11.2.3

$\chi^2 = 4.8082,\ \chi^2_{0.05} = 7.81,$ do not reject H_0

S11.2.5

$\chi^2 = 26.5765,\ \chi^2_{0.01} = 23.21,$ reject H_0

S11.2.7

$\chi^2 = 2.1401,\ \chi^2_{0.05} = 5.99,$ do not reject H_0

S11.2.9

$\chi^2 = 2.944,\ df = 5,$ Rejection Region: $[9.236, \infty),$ Decision: Fail to reject H_0 of balance

11.3 F-tests for Equality of Two Variances

Basic

1. Find $F_{0.01}$ for each of the following degrees of freedom.
   a. $df_1 = 5$ and $df_2 = 5$
   b. $df_1 = 5$ and $df_2 = 12$
   c. $df_1 = 12$ and $df_2 = 20$
2. Find $F_{0.05}$ for each of the following degrees of freedom.
   a. $df_1 = 6$ and $df_2 = 6$
   b. $df_1 = 6$ and $df_2 = 12$
   c. $df_1 = 12$ and $df_2 = 30$
3. Find $F_{0.95}$ for each of the following degrees of freedom.
   a. $df_1 = 6$ and $df_2 = 6$
   b. $df_1 = 6$ and $df_2 = 12$
   c. $df_1 = 12$ and $df_2 = 30$
4. Find $F_{0.90}$ for each of the following degrees of freedom.
   a. $df_1 = 5$ and $df_2 = 5$
   b. $df_1 = 5$ and $df_2 = 12$
   c. $df_1 = 12$ and $df_2 = 20$
5. For $df_1 = 7$, $df_2 = 10$ and $\alpha = 0.05$, find
   a. F_{α}
   b. $F_{1-\alpha}$
   c. $F_{\alpha/2}$
   d. $F_{1-\alpha/2}$
6. For $df_1 = 15$, $df_2 = 8$ and $\alpha = 0.01$, find
   a. F_{α}
   b. $F_{1-\alpha}$
   c. $F_{\alpha/2}$
   d. $F_{1-\alpha/2}$
7. For each of the two samples

$$\begin{aligned} &\text{Sample 1} : \{8, 2, 11, 0, -2\} \\ &\text{Sample 2} : \{-2, 0, 0, 0, 2, 4, -1\} \end{aligned} \tag{11.E.1}$$

   find
   a. the sample size
   b. the sample mean
   c. the sample variance

8. For each of the two samples

$$\begin{aligned} &\text{Sample 1}: \{0.8, 1.2, 1.1, 0.8, -2.0\} \\ &\text{Sample 2}: \{-2.0, 0.0, 0.7, 0.8, 2.2, 4.1, -1.9\} \end{aligned} \tag{11.E.2}$$

find

   a. the sample size
   b. the sample mean
   c. the sample variance

9. Two random samples taken from two normal populations yielded the following information:

Sample	Sample Size	Sample Variance
1	$n_1 = 16$	$s_1^2 = 53$
2	$n_2 = 21$	$s_2^2 = 32$

   a. Find the statistic $F = s_1^2/s_2^2$
   b. Find the degrees of freedom df_1 and df_2.
   c. Find $F_{0.05}$ using df_1 and df_2 computed above.
   d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 > \alpha_2^2$ at the 5% level of significance.

10. Two random samples taken from two normal populations yielded the following information:

Sample	Sample Size	Sample Variance
1	$n_1 = 11$	$s_1^2 = 61$
2	$n_2 = 8$	$s_2^2 = 44$

   a. Find the statistic $F = s_1^2/s_2^2$.
   b. Find the degrees of freedom df_1 and df_2.
   c. Find $F_{0.05}$ using df_1 and df_2computed above.
   d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 > \alpha_2^2$ at the 5% level of significance.

11. Two random samples taken from two normal populations yielded the following information:
   a. Find the statistic $F = s_1^2/s_2^2$.
   b. Find the degrees of freedom df_1 and df_2.
   c. For $\alpha = 0.05$ find $F_{1-\alpha}$ using df_1 and df_2 computed above.
   d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 < \alpha_2^2$ at the 5% level of significance.

12.

Sample	Sample Size	Sample Variance
1	$n_1 = 10$	$s_1^2 = 12$
2	$n_2 = 13$	$s_2^2 = 23$

13. Two random samples taken from two normal populations yielded the following information:

Sample	Sample Size	Sample Variance
1	$n_1 = 8$	$s_1^2 = 102$
2	$n_2 = 8$	$s_2^2 = 603$

   a. Find the statistic $F = s_1^2/s_2^2$
   b. Find the degrees of freedom df_1 and df_2.
   c. For $\alpha = 0.05$ find $F_{1-\alpha}$using df_1 and df_2computed above.
   d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 < \alpha_2^2$at the 5% level of significance.

14. Two random samples taken from two normal populations yielded the following information:

Sample	Sample Size	Sample Variance
1	$n_1 = 9$	$s_1^2 = 123$
2	$n_2 = 31$	$s_2^2 = 543$

a. Find the statistic $F = s_1^2/s_2^2$
b. Find the degrees of freedom df_1 and df_2.
c. For $\alpha = 0.05$ find $F_{1-\alpha/2}$ and $F_{\alpha/2}$ using df_1 and df_2computed above.
d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 \neq \alpha_2^2$at the 5% level of significance.

15. Two random samples taken from two normal populations yielded the following information:

Sample	Sample Size	Sample Variance
1	$n_1 = 21$	$s_1^2 = 199$
2	$n_2 = 21$	$s_2^2 = 66$

a. Find the statistic $F = s_1^2/s_2^2$
b. Find the degrees of freedom df_1 and df_2.
c. For $\alpha = 0.05$ find df_1 and df_2computed above.
d. Perform the test the hypotheses $H_0 : \alpha_1^2 = \alpha_2^2$ vs $H_a : \alpha_1^2 \neq \alpha_2^2$at the 5% level of significance.

Applications

15. Japanese sturgeon is a subspecies of the sturgeon family indigenous to Japan and the Northwest Pacific. In a particular fish hatchery newly hatched baby Japanese sturgeon are kept in tanks for several weeks before being transferred to larger ponds. Dissolved oxygen in tank water is very tightly monitored by an electronic system and rigorously maintained at a target level of 6.5 milligrams per liter (mg/l). The fish hatchery looks to upgrade their water monitoring systems for tighter control of dissolved oxygen. A new system is evaluated against the old one currently being used in terms of the variance in measured dissolved oxygen. Thirty-one water samples from a tank operated with the new system were collected and 16 water samples from a tank operated with the old system were collected, all during the course of a day. The samples yield the following information:

$$\begin{aligned} &\text{New Sample 1: } n_1 = 31 \; s_1^2 = 0.0121 \\ &\text{Old Sample 2: } n_1 = 16 \; s_2^2 = 0.0319 \end{aligned} \tag{11.E.3}$$

Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the new system will provide a tighter control of dissolved oxygen in the tanks.

16. The risk of investing in a stock is measured by the volatility, or the variance, in changes in the price of that stock. Mutual funds are baskets of stocks and offer generally lower risk to investors. Different mutual funds have different focuses and offer different levels of risk. Hippolyta is deciding between two mutual funds, A and B, with similar expected returns. To make a final decision, she examined the annual returns of the two funds during the last ten years and obtained the following information:

$$\begin{aligned} &\text{Mutual Fund A Sample 1: } n_1 = 10 \; s_1^2 = 0.012 \\ &\text{Mutual Fund B Sample 2: } n_1 = 10 \; s_2^2 = 0.005 \end{aligned} \tag{11.E.4}$$

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that the two mutual funds offer different levels of risk.

17. It is commonly acknowledged that grading of the writing part of a college entrance examination is subject to inconsistency. Every year a large number of potential graders are put through a rigorous training program before being given grading assignments. In order to gauge whether such a training program really enhances consistency in grading, a statistician conducted an experiment in which a reference essay was given to 61 trained graders and 31 untrained graders. Information on the scores given by these graders is summarized below:

$$\begin{aligned} &\text{Trained Sample 1: } n_1 = 61 \; s_1^2 = 2.15 \\ &\text{Untrained Sample 2: } n_1 = 31 \; s_2^2 = 3.91 \end{aligned} \tag{11.E.5}$$

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that the training program enhances the consistency in essay grading.

18. A common problem encountered by many classical music radio stations is that their listeners belong to an increasingly narrow band of ages in the population. The new general manager of a classical music radio station believed that a new playlist offered by a professional programming agency would attract listeners from a wider range of ages. The new list was used for a year. Two random samples were taken before and after the new playlist was adopted. Information on the ages of the listeners in the sample are summarized below:

$$\begin{aligned} &\text{Before Sample 1: } n_1 = 21 \; s_1^2 = 56.25 \\ &\text{After Sample 2: } n_1 = 16 \; s_2^2 = 76.56 \end{aligned} \tag{11.E.6}$$

Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the new playlist has expanded the range of listener ages.

19. A laptop computer maker uses battery packs supplied by two companies, A and B. While both brands have the same average battery life between charges (LBC), the computer maker seems to receive more complaints about shorter LBC than expected for battery packs supplied by company B. The computer maker suspects that this could be caused by higher variance in LBC for Brand B. To check that, ten new battery packs from each brand are selected, installed on the same models of laptops, and the laptops are allowed to run until the battery packs are completely discharged. The following are the observed LBCs in hours.

Brand A	Brand B
3.2	3.0
3.4	3.5
2.8	2.9
3.0	3.1
3.0	2.3
3.0	2.0
2.8	3.0
2.9	2.9
3.0	3.0
3.0	4.1

Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that the LBCs of Brand B have a larger variance that those of Brand A.

20. A manufacturer of a blood-pressure measuring device for home use claims that its device is more consistent than that produced by a leading competitor. During a visit to a medical store a potential buyer tried both devices on himself repeatedly during a short period of time. The following are readings of systolic pressure.

Manufacturer	Competitor
132	129
134	132
129	129
129	138
130	
132	

a. Test, at the 5% level of significance, whether the data provide sufficient evidence to conclude that the manufacturer's claim is true.

b. Repeat the test at the 10% level of significance. Quote as many computations from part (a) as possible.

Large Data Set Exercises

Large Data Sets not available

21. Large Data Sets 1A and 1B record SAT scores for 419 male and 581 female students. Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that the variances of scores of male and female students differ.
22. Large Data Sets 7, 7A, and 7B record the survival times of 140 laboratory mice with thymic leukemia. Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the variances of survival times of male mice and female mice differ.

Answers

1. a. 11.0
   b. 5.06
   c. 3.23
2.
3. a. 0.23
   b. 0.25
   c. 0.40
4.
5. a. 3.14
   b. 0.27
   c. 3.95
   d. 0.21
6.
7. I. Sample 1
      a. $n_1 = 5$
      b. $\bar{x_1} = 3.8$
      c. $s_1^2 = 30.2$
   II. Sample 1
      a. $n_2 = 7$
      b. $\bar{x_1} = 0.4286$
      c. $s_2^2 = 3.95$
8.
9. a. 1.6563
   b. $df_1 = 15,\ df_2 = 20$
   c. $F_{0.05} = 2.2$
   d. do not reject H_0
10.
11. a. 0.5217
    b. $df_1 = 9,\ df_2 = 12$
    c. $F_{0.95} = 0.3254$
    d. do not reject H_0
12.
13. a. 0.1692
    b. $df_1 = 8,\ df_2 = 30$
    c. $F_{0.975} = 0.26,\ F_{0.025} = 2.65$
    d. reject H_0
14.
15. $F = 0.3793,\ F_{0.90} = 0.58$, reject H_0

16.
17. $F = 0.5499,\ F_{0.95} = 0.61$, reject H_0
18.
19. $F = 0.0971,\ F_{0.95} = 0.31$, reject H_0
20.
21. $F = 0.893131, df_1 = 418,\ df_2 = 580$. Rejection Region: $(0, 0.7897] \cup [1.2614, \infty)$. Decision: Fail to reject H_0 of equal variances.

11.4 F-Tests in One-Way ANOVA

Basic

1. The following three random samples are taken from three normal populations with respective means μ_1, μ_2 and μ_3, and the same variance σ^2.

Sample 1	Sample 2	Sample 3
2	3	0
2	5	1
3	7	2
5		1
3		

   a. Find the combined sample size n.
   b. Find the combined sample mean $\bar{x}$.
   c. Find the sample mean for each of the three samples.
   d. Find the sample variance for each of the three samples.
   e. Find MST.
   f. Find MSE.
   g. Find $F = MST/MSE$.

2. The following three random samples are taken from three normal populations with respective means μ_1, μ_2 and μ_3, and the same variance σ^2.

Sample 1	Sample 2	Sample 3
0.0	1.3	0.2
0.1	1.5	0.2
0.2	1.7	0.3
0.1		0.5
		0.0

   a. Find the combined sample size n.
   b. Find the combined sample mean $\bar{x}$.
   c. Find the sample mean for each of the three samples.
   d. Find the sample variance for each of the three samples.
   e. Find MST.
   f. Find MSE.
   g. Find $F = MST/MSE$.

3. Refer to Exercise 1.
   a. Find the number of populations under consideration K.
   b. Find the degrees of freedom $df_1 = K - 1$ and $df_2 = n - K$
   c. For $\alpha = 0.05$, find F_α with the degrees of freedom computed above.
   d. At $\alpha = 0.05$, test hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad (11.E.7)$$
$$vs\ H_a : \text{at least one pair of the population means are not equal}$$

4. Refer to Exercise 2.
   a. Find the number of populations under consideration K.
   b. Find the degrees of freedom $df_1 = K - 1$ and $df_2 = n - K$
   c. For $\alpha = 0.01$, find F_α with the degrees of freedom computed above.
   d. At $\alpha = 0.01$, test hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad (11.E.8)$$
$$vs\ H_a : \text{at least one pair of the population means are not equal}$$

Applications

5. The Mozart effect refers to a boost of average performance on tests for elementary school students if the students listen to Mozart's chamber music for a period of time immediately before the test. In order to attempt to test whether the Mozart effect actually exists, an elementary school teacher conducted an experiment by dividing her third-grade class of 15 students into three groups of 5. The first group was given an end-of-grade test without music; the second group listened to Mozart's chamber music for 10 minutes; and the third groups listened to Mozart's chamber music for 20 minutes before the test. The scores of the 15 students are given below:

Group 1	Group 2	Group 3
80	79	73
63	73	82
74	74	79
71	77	82
70	81	84

   Using the ANOVA F-test at $\alpha = 0.10$, is there sufficient evidence in the data to suggest that the Mozart effect exists?

6. The Mozart effect refers to a boost of average performance on tests for elementary school students if the students listen to Mozart's chamber music for a period of time immediately before the test. Many educators believe that such an effect is not necessarily due to Mozart's music per se but rather a relaxation period before the test. To support this belief, an elementary school teacher conducted an experiment by dividing her third-grade class of 15 students into three groups of 5. Students in the first group were asked to give themselves a self-administered facial massage; students in the second group listened to Mozart's chamber music for 15 minutes; students in the third group listened to Schubert's chamber music for 15 minutes before the test. The scores of the 15 students are given below:

Group 1	Group 2	Group 3
79	82	80
81	84	81
80	86	71
89	91	90
86	82	86

   Test, using the ANOVA F-test at the 10% level of significance, whether the data provide sufficient evidence to conclude that any of the three relaxation method does better than the others.

7. Precision weighing devices are sensitive to environmental conditions. Temperature and humidity in a laboratory room where such a device is installed are tightly controlled to ensure high precision in weighing. A newly designed weighing device is claimed to be more robust against small variations of temperature and humidity. To verify such a claim, a laboratory tests the new device under four settings of temperature-humidity conditions. First, two levels of *high* and *low* temperature and two levels of *high* and *low* humidity are identified. Let T stand for temperature and H for humidity. The four experimental settings are

defined and noted as (T, H): (high, high), (high, low), (low, high), and (low, low). A pre-calibrated standard weight of 1 kg was weighed by the new device four times in each setting. The results in terms of error (in micrograms mcg) are given below:

(high, high)	(high, low)	(low, high)	(low, low)
−1.50	11.47	−14.29	5.54
−6.73	9.28	−18.11	10.34
11.69	5.58	−11.16	15.23
−5.72	10.80	−10.41	−5.69

Test, using the ANOVA F-test at the 1% level of significance, whether the data provide sufficient evidence to conclude that the mean weight readings by the newly designed device vary among the four settings.

8. To investigate the real cost of owning different makes and models of new automobiles, a consumer protection agency followed 16 owners of new vehicles of four popular makes and models, call them TC,HA,NA, and FT, and kept a record of each of the owner's real cost in dollars for the first five years. The five-year costs of the 16 car owners are given below:

TC	HA	NA	FT
8423	7776	8907	10333
7889	7211	9077	9217
8665	6870	8732	10540
	7129	9747	
	7359	8677	

Test, using the ANOVA F-test at the 5% level of significance, whether the data provide sufficient evidence to conclude that there are differences among the mean real costs of ownership for these four models.

9. Helping people to lose weight has become a huge industry in the United States, with annual revenue in the hundreds of billion dollars. Recently each of the three market-leading weight reducing programs claimed to be the most effective. A consumer research company recruited 33 people who wished to lose weight and sent them to the three leading programs. After six months their weight losses were recorded. The results are summarized below:

Statistic	Prog. 1	Prog. 2	Prog. 3
Sample Mean	$\bar{x_1} = 10.65$	$\bar{x_2} = 8.90$	$\bar{x_3} = 9.33$
Sample Variance	$s_1^2 = 27.20$	$s_2^2 = 16.86$	$s_3^2 = 32.40$
Sample Size	$n_1 = 11$	$n_2 = 11$	$n_3 = 11$

The mean weight loss of the combined sample of all 33 people was $\bar{x} = 9.63$. Test, using the ANOVA F-test at the 5% level of significance, whether the data provide sufficient evidence to conclude that some program is more effective than the others.

10. A leading pharmaceutical company in the disposable contact lenses market has always taken for granted that the sales of certain peripheral products such as contact lens solutions would automatically go with the established brands. The long-standing culture in the company has been that lens solutions would not make a significant difference in user experience. Recent market research surveys, however, suggest otherwise. To gain a better understanding of the effects of contact lens solutions on user experience, the company conducted a comparative study in which 63 contact lens users were randomly divided into three groups, each of which received one of three top selling lens solutions on the market, including one of the company's own. After using the assigned solution for two weeks, each participant was asked to rate the solution on the scale of 1 to 5 for satisfaction, with 5 being the highest level of satisfaction. The results of the study are summarized below:

Statistics	Sol. 1	Sol. 2	Sol. 3
Sample Mean	$\bar{x_1} = 3.28$	$\bar{x_2} = 3.96$	$\bar{x_3} = 4.10$
Sample Variance	$s_1^2 = 0.15$	$s_2^2 = 0.32$	$s_3^2 = 0.36$

Statistics	Sol. 1	Sol. 2	Sol. 3
Sample Size	$n_1 = 18$	$n_1 = 23$	$n_1 = 22$

The mean satisfaction level of the combined sample of all 63 participants was $\bar{x} = 3.81$. Test, using the ANOVA F-test at the 5% level of significance, whether the data provide sufficient evidence to conclude that not all three average satisfaction levels are the same

Large Data Set Exercise

Large Data Set not available

11. Large $\text{Data Set } 9$ records the costs of materials (textbook, solution manual, laboratory fees, and so on) in each of ten different courses in each of three different subjects, chemistry, computer science, and mathematics. Test, at the 1% level of significance, whether the data provide sufficient evidence to conclude that the mean costs in the three disciplines are not all the same.

Answers

1. a. $n = 12$
   b. $\bar{x} = 2.8333$
   c. $\bar{x_1} = 3,\ \bar{x_2} = 5,\ \bar{x_3} = 1$
   d. $s_1^2 = 1.5,\ s_2^2 = 4,\ s_3^2 = 0.6667$
   e. $MST = 13.83$
   f. $MSE = 1.78$
   g. $F = 7.7812$
2.
3. a. $K = 3$
   b. $df_1 = 2,\ df_2 = 9$
   c. $F_{0.05} = 4.26$
   d. $F = 5.53$, reject H_0
4.
5. $F = 3.9647,\ F_{0.10} = 2.81$, reject H_0
6.
7. $F = 9.6018,\ F_{0.01} = 5.95$, reject H_0
8.
9. $F = 0.3589,\ F_{0.05} = 5.32$, do not reject H_0
10.
11. $F = 1.418$, $df_1 = 2,\ and\ df_2 = 27$, Rejection Region: $[5.4881, \infty)$, Decision: Fail to reject H_0 of equal means.

This page titled 11.E: Chi-Square Tests and F-Tests (Exercises) is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.E: Chi-Square Tests and F-Tests (Exercises)** has no license indicated.

Index

A

Adding probabilities
3.2: Complements, Intersections, and Unions
alternative hypothesis
8.1: The Elements of Hypothesis Testing
anova
11.4: F-Tests in One-Way ANOVA

B

binomial probability distribution
4.3: The Binomial Distribution
binomial random variable
4.3: The Binomial Distribution
box plots
2.4: Relative Position of Data

C

central limit theorem
6.2: The Sampling Distribution of the Sample Mean
Chebyshev's Theorem
2.5: The Empirical Rule and Chebyshev's Theorem
coefficient of determination
10.6: The Coefficient of Determination
combined sample size
11.4: F-Tests in One-Way ANOVA
Comparing two population means
9.1: Comparison of Two Population Means- Large, Independent Samples
9.2: Comparison of Two Population Means - Small, Independent Samples
Comparing Two Population Proportions
9.4: Comparison of Two Population Proportions
complement
3.2: Complements, Intersections, and Unions
conditional probability
3.3: Conditional Probability and Independent Events
confidence interval for estimating a population mean
7.4: Sample Size Considerations
CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS
9.4: Comparison of Two Population Proportions
confidence interval for the difference in two population means
9.5: Sample Size Considerations
confidence interval formula for estimating a population proportion
7.4: Sample Size Considerations
Confidence Intervals for a Proportion
7.3: Large Sample Estimation of a Population Proportion
7.4: Sample Size Considerations
contingency table
11.1: Chi-Square Tests for Independence
critical value test
8.5: Large Sample Tests for a Population Proportion
Cumulative Normal Probability
5.2: The Standard Normal Distribution
cumulative probability distributions
4.3: The Binomial Distribution

D

data frequency table
1.3: Presentation of Data
DENSITY FUNCTION
5.1: Continuous Random Variables

E

ELEMENT
3.1: Sample Spaces, Events, and Their Probabilities
empirical rule
2.5: The Empirical Rule and Chebyshev's Theorem
7.1: Large Sample Estimation of a Population Mean
error term
10.3: Modelling Linear Relationships with Randomness Present
estimation
10.7: Estimation and Prediction
extrapolation
10.4: The Least Squares Regression Line

G

goodness of fit
10.4: The Least Squares Regression Line

H

histograms
2.1: Three Popular Data Displays
Hypothesis Testing
8.1: The Elements of Hypothesis Testing

I

independent events
3.3: Conditional Probability and Independent Events
INTERSECTIONS
3.2: Complements, Intersections, and Unions

L

least squares criterion
10.4: The Least Squares Regression Line
level of significance
8.1: The Elements of Hypothesis Testing
linear correlation coefficient
10.2: The Linear Correlation Coefficient
LINEAR REGRESSION MODEL
10.3: Modelling Linear Relationships with Randomness Present
10.5: Statistical Inferences About β_1

M

margin of error
7.1: Large Sample Estimation of a Population Mean
mean
4.2: Probability Distributions for Discrete Random Variables
mean of the sample proportion
6.3: The Sample Proportion
mean square for error
11.4: F-Tests in One-Way ANOVA
mean square for treatment
11.4: F-Tests in One-Way ANOVA
Minimum Sample Size for Estimating a Population Mean
7.4: Sample Size Considerations
mode
2.2: Measures of Central Location - Three Kinds of Averages
model
10.3: Modelling Linear Relationships with Randomness Present
most conservative estimate
7.4: Sample Size Considerations
mutually exclusive
3.2: Complements, Intersections, and Unions

N

noise
10.3: Modelling Linear Relationships with Randomness Present
normal distribution
5.1: Continuous Random Variables
null hypothesis
8.1: The Elements of Hypothesis Testing

O

observed significance
8.3: The Observed Significance of a Test
OCCURRENCE
3.1: Sample Spaces, Events, and Their Probabilities

P

paired difference samples
9.3: Comparison of Two Population Means - Paired Samples
paired samples
9.3: Comparison of Two Population Means - Paired Samples
percentiles
2.4: Relative Position of Data
pooled variance
9.2: Comparison of Two Population Means - Small, Independent Samples
population mean
2.2: Measures of Central Location - Three Kinds of Averages
population median
2.2: Measures of Central Location - Three Kinds of Averages
population mode
2.2: Measures of Central Location - Three Kinds of Averages
population regression line
10.3: Modelling Linear Relationships with Randomness Present
prediction
10.7: Estimation and Prediction
probability distribution function
4.2: Probability Distributions for Discrete Random Variables
Probability Rule for Complements
3.2: Complements, Intersections, and Unions

Q

quartiles
2.4: Relative Position of Data

R

random experiment
3.1: Sample Spaces, Events, and Their Probabilities

range
2.3: Measures of Variability
rare events
8.3: The Observed Significance of a Test
relative frequency histograms
2.1: Three Popular Data Displays

S

sample mean
2.2: Measures of Central Location - Three Kinds of Averages
6.1: The Mean and Standard Deviation of the Sample Mean
sample median
2.2: Measures of Central Location - Three Kinds of Averages
sample mode
2.2: Measures of Central Location - Three Kinds of Averages
sample proportion
6.3: The Sample Proportion
sample size
9.5: Sample Size Considerations
sample space
3.1: Sample Spaces, Events, and Their Probabilities
sample standard deviation
6.1: The Mean and Standard Deviation of the Sample Mean
sampling distribution
6.3: The Sample Proportion
scatter diagram
10.1: Linear Relationships Between Variables
scatter plot
10.1: Linear Relationships Between Variables
set notation
1.3: Presentation of Data
Skewed
2.2: Measures of Central Location - Three Kinds of Averages
slope
10.5: Statistical Inferences About β_1
SPECIFICITY OF A DIAGNOSTIC TEST
3.3: Conditional Probability and Independent Events
standard deviation
2.3: Measures of Variability
4.2: Probability Distributions for Discrete Random Variables
standard deviation of the sample proportion
6.3: The Sample Proportion
standard normal random variable
5.2: The Standard Normal Distribution
Sum of the Squared Errors
10.4: The Least Squares Regression Line

T

tails
5.4: Areas of Tails of Distributions
Tests for Independence
11.1: Chi-Square Tests for Independence
tree diagram
3.1: Sample Spaces, Events, and Their Probabilities
Type I error
8.1: The Elements of Hypothesis Testing
Type II error
8.1: The Elements of Hypothesis Testing

U

unions
3.2: Complements, Intersections, and Unions

V

variance
2.3: Measures of Variability
Venn diagram
3.1: Sample Spaces, Events, and Their Probabilities

Glossary

***F*-distribution.**

alternative hypothesis | A statement that is accepted as true only if there is convincing evidence in favor of it.

ANOVA *F*-test | a test based on an F statistic to check whether several population means are equal.

binomial random variable | A random variable that counts successes in a fixed number of independent, identical trials of a success/failure experiment.

box plot | For a data set, a diagram constructed using the five-number summary, which graphically summarizes the distribution of the data.

chi-square goodness-of-fit test | A test based on a chi-square statistic to check whether a sample is taken from a population with a hypothesized probability distribution.

chi-square random variable | A random variable that follows a chi-square distribution.

chi-square test | A test based on a chi-square statistic to check whether two factors are independent.

coefficient of determination | A number that measures the proportion of the variability in y that is explained by x.

complement of an event | The event does not occur.

conditional probability | The probability of the event A taking into account the fact that event B is known to have occurred.

confidence interval | An interval with endpoints $\bar{x} \pm E$, computed from the sample data in such a way that a specified proportion of all intervals constructed by this process will contain the parameter of interest.

continuous | A random variable whose possible values contain an interval of decimal numbers.

critical value | The number or one of a pair of numbers that determines the rejection region.

data list | An explicit listing of all the individual measurements made on a sample.

degrees of freedom | A number that specifies a particular t-distribution and that is computed based on the sample size.

density function | The function $f(x)$ such that probabilities of a continuous random variable X are areas of regions under the graph of $y = f(x)$.

Descriptive statistics | The organization, display, and description of data.

error | Using $y - \hat{y}$, the actual y-value of a data point minus the y-value that is computed from the equation of the line fitting the data.

event | Any set of outcomes.

expected value | Its mean.

extrapolation | The process of using the least squares regression equation to estimate the value of y at an x value not in the proper range.

***F*-distribution** | A particular probability distribution specified by two degrees of freedom, df1 and df2.

factors | A variable with several qualitative levels.

five-number summary | Of a data set, the list {msubxmin, Q1, Q2, Q3, xmax}.

frequency | Of a class of measurements, the number of measurements in the data set that are in the class.

hypothesis | A statement about a population parameter.

Hypothesis testing | A statistical procedure in which a choice is made between a null hypothesis and a specific alternative hypothesis based on information in a sample.

independent | Events whose probability of occurring together is the product of their individual probabilities.

Inferential statistics | Drawing conclusions about a population based on a sample.

interquartile range (IQR) | Of a data set, the difference between the first and third quartiles.

intersection of events | Both events occur.

least squares regression equation | The equation $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ of the least squares regression line.

least squares regression line | The line that best fits a set of sample data in the sense of minimizing the sum of the squared errors.

level of confidence | The proportion of confidence intervals which, if under repeated random sampling were always constructed according to the formula of the text, would contain the parameter of interest.

level of significance of the test | The probability α that defines an event as "rare;" the probability that the test procedure will lead to a Type I error.

linear correlation coefficient | A number computed directly from the data that measures the strength of the linear relationship between the two variables x and y.

most conservative estimate | The estimate obtained using $\hat{p} = 0.5$, which gives the largest estimate of n.

mutually exclusive | Events that cannot both occur at once.

normal distribution | Assignment of probabilities to a continuous random variable using a bell curve for the density function.

normal random variable | A continuous random variable whose probabilities are determined by a bell curve.

NOVA | Analysis of variance.

null hypothesis | The statement that is assumed to be true unless there is convincing evidence to the contrary.

observed significance *or* *p*-value | The probability, if H_0 is true, of obtaining a result as contrary to H_0 and in favor of H_a as the result observed in the sample data.

percentile rank | Of a measurement x, the percentage of the data that are less than or equal to x.

percentile rank | Of a measurement x, the percentage of the data that are less than or equal to x.

population mean | The familiar average of a population data set.

population regression line | The line with equation y _β1x+β0 that gives the mean of the variable y over the sub-population determined by x.

population standard deviation | The variability of population data as measured by the number $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$.

probability distribution | A list of each possible value and its probability.

probability of an event | A number that measures the likelihood of the event.

probability of an outcome | A number that measures the likelihood of the outcome.

Qualitative data | Measurements for which there is no natural numerical scale.

Quantitative data | Numerical measurements that arise from a natural numerical scale.

quartiles | Of a data set, the three numbers Q_1, Q_2, Q_3 that divide the data approximately into fourths.

random variable | A numerical value generated by a random experiment.

range | The variability of a data set as measured by the number $R = x_{\max} - x_{\min}$.

rejection region | An interval or union of intervals such that the null hypothesis is rejected if and only if the statistic of interest lies in this region.

relative frequency histogram | A graphical device showing how data are distributed across the range of their values by collecting them into classes and indicating the proportion of measurements in each class.

sample | The objects examined.

sample data | The measurements from a sample.

sample mean | The familiar average of a sample data set.

sample median | The middle value when data are listed in numerical order.

sample mode | The most frequent value in a data set.

sample standard deviation | The variability of sample data as measured by the number $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$.

sample standard deviation of errors | The statistic s_ε.

sampling distribution | The probability distribution of a sample statistic when the statistic is viewed as a random variable.

standard deviation | The number $\sqrt{\Sigma(x-\mu)^2 P(x)}$ (also computed using $\sqrt{[\Sigma x2P(x)\,]-\mu 2}$), measuring its variability under repeated trials.

standard normal random variable | The normal random variable with mean 0 and standard deviation 1.

standardized test statistic | The standardized statistic used in performing the test.

statistic | A number computed from the sample data.

Statistics | Collection, display, analysis, and inference from data.

tail | The region under a density curve whose area is either $P(X < x^*)$ or $P(X > x^*)$ for some number x^*.

Type II error | Failure to reject a false null hypothesis.

union of events | One or the other event occurs.

***z*-score** | Of a measurement x, the distance of x from the mean in units of standard deviation.

Detailed Licensing

Overview

Title: Introductory Statistics (Shafer and Zhang)

Webpages: 80

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 3.0: 95% (76 pages)
- Undeclared: 5% (4 pages)

By Page

- Introductory Statistics (Shafer and Zhang) - *CC BY-NC-SA 3.0*
  - Front Matter - *CC BY-NC-SA 3.0*
    - TitlePage - *CC BY-NC-SA 3.0*
    - InfoPage - *CC BY-NC-SA 3.0*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Introduction to Statistics - *CC BY-NC-SA 3.0*
    - 1.1: Basic Definitions and Concepts - *CC BY-NC-SA 3.0*
    - 1.2: Overview - *CC BY-NC-SA 3.0*
    - 1.3: Presentation of Data - *CC BY-NC-SA 3.0*
    - 1.E: Introduction to Statistics (Exercises) - *CC BY-NC-SA 3.0*
  - 2: Descriptive Statistics - *CC BY-NC-SA 3.0*
    - 2.1: Three Popular Data Displays - *CC BY-NC-SA 3.0*
    - 2.2: Measures of Central Location - Three Kinds of Averages - *CC BY-NC-SA 3.0*
    - 2.3: Measures of Variability - *CC BY-NC-SA 3.0*
    - 2.4: Relative Position of Data - *CC BY-NC-SA 3.0*
    - 2.5: The Empirical Rule and Chebyshev's Theorem - *CC BY-NC-SA 3.0*
    - 2.E: Descriptive Statistics (Exercises) - *CC BY-NC-SA 3.0*
  - 3: Basic Concepts of Probability - *CC BY-NC-SA 3.0*
    - 3.1: Sample Spaces, Events, and Their Probabilities - *CC BY-NC-SA 3.0*
    - 3.2: Complements, Intersections, and Unions - *CC BY-NC-SA 3.0*
    - 3.3: Conditional Probability and Independent Events - *CC BY-NC-SA 3.0*
    - 3.E: Basic Concepts of Probability (Exercises) - *CC BY-NC-SA 3.0*
  - 4: Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 4.1: Random Variables - *CC BY-NC-SA 3.0*
    - 4.2: Probability Distributions for Discrete Random Variables - *CC BY-NC-SA 3.0*
    - 4.3: The Binomial Distribution - *CC BY-NC-SA 3.0*
    - 4.E: Discrete Random Variables (Exercises) - *CC BY-NC-SA 3.0*
  - 5: Continuous Random Variables - *CC BY-NC-SA 3.0*
    - 5.1: Continuous Random Variables - *CC BY-NC-SA 3.0*
    - 5.2: The Standard Normal Distribution - *CC BY-NC-SA 3.0*
    - 5.3: Probability Computations for General Normal Random Variables - *CC BY-NC-SA 3.0*
    - 5.4: Areas of Tails of Distributions - *CC BY-NC-SA 3.0*
    - 5.E: Continuous Random Variables (Exercises) - *CC BY-NC-SA 3.0*
  - 6: Sampling Distributions - *CC BY-NC-SA 3.0*
    - 6.1: The Mean and Standard Deviation of the Sample Mean - *CC BY-NC-SA 3.0*
    - 6.2: The Sampling Distribution of the Sample Mean - *CC BY-NC-SA 3.0*
    - 6.3: The Sample Proportion - *CC BY-NC-SA 3.0*
    - 6.E: Sampling Distributions (Exercises) - *CC BY-NC-SA 3.0*
  - 7: Estimation - *CC BY-NC-SA 3.0*
    - 7.1: Large Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*
    - 7.2: Small Sample Estimation of a Population Mean - *CC BY-NC-SA 3.0*
    - 7.3: Large Sample Estimation of a Population Proportion - *CC BY-NC-SA 3.0*
    - 7.4: Sample Size Considerations - *CC BY-NC-SA 3.0*
    - 7.E: Estimation (Exercises) - *CC BY-NC-SA 3.0*
  - 8: Testing Hypotheses - *CC BY-NC-SA 3.0*
    - 8.1: The Elements of Hypothesis Testing - *CC BY-NC-SA 3.0*

- 8.2: Large Sample Tests for a Population Mean - *CC BY-NC-SA 3.0*
- 8.3: The Observed Significance of a Test - *CC BY-NC-SA 3.0*
- 8.4: Small Sample Tests for a Population Mean - *CC BY-NC-SA 3.0*
- 8.5: Large Sample Tests for a Population Proportion - *CC BY-NC-SA 3.0*
- 8.E: Testing Hypotheses (Exercises) - *CC BY-NC-SA 3.0*

- 9: Two-Sample Problems - *CC BY-NC-SA 3.0*
  - 9.1: Comparison of Two Population Means- Large, Independent Samples - *CC BY-NC-SA 3.0*
  - 9.2: Comparison of Two Population Means - Small, Independent Samples - *CC BY-NC-SA 3.0*
  - 9.3: Comparison of Two Population Means - Paired Samples - *CC BY-NC-SA 3.0*
  - 9.4: Comparison of Two Population Proportions - *CC BY-NC-SA 3.0*
  - 9.5: Sample Size Considerations - *CC BY-NC-SA 3.0*
  - 9.E: Two-Sample Problems (Exercises) - *CC BY-NC-SA 3.0*
- 10: Correlation and Regression - *CC BY-NC-SA 3.0*
  - 10.1: Linear Relationships Between Variables - *CC BY-NC-SA 3.0*
  - 10.2: The Linear Correlation Coefficient - *CC BY-NC-SA 3.0*
  - 10.3: Modelling Linear Relationships with Randomness Present - *CC BY-NC-SA 3.0*
  - 10.4: The Least Squares Regression Line - *CC BY-NC-SA 3.0*
  - 10.5: Statistical Inferences About β_1 - *CC BY-NC-SA 3.0*
  - 10.6: The Coefficient of Determination - *CC BY-NC-SA 3.0*
  - 10.7: Estimation and Prediction - *CC BY-NC-SA 3.0*
  - 10.8: A Complete Example - *CC BY-NC-SA 3.0*
  - 10.9: Formula List - *CC BY-NC-SA 3.0*
  - 10.E: Correlation and Regression (Exercises) - *CC BY-NC-SA 3.0*
- 11: Chi-Square Tests and F-Tests - *CC BY-NC-SA 3.0*
  - 11.1: Chi-Square Tests for Independence - *CC BY-NC-SA 3.0*
  - 11.2: Chi-Square One-Sample Goodness-of-Fit Tests - *CC BY-NC-SA 3.0*
  - 11.3: F-tests for Equality of Two Variances - *CC BY-NC-SA 3.0*
  - 11.4: F-Tests in One-Way ANOVA - *CC BY-NC-SA 3.0*
  - 11.E: Chi-Square Tests and F-Tests (Exercises) - *CC BY-NC-SA 3.0*
- Back Matter - *CC BY-NC-SA 3.0*
  - Index - *Undeclared*
  - Glossary - *CC BY-NC-SA 3.0*
  - Detailed Licensing - *Undeclared*