Collaborative Statistics

By: Susan Dean and Barbara Illowsky

Online: <http://cnx.org/content/col10522/1.39>

Collection structure revised: 2011/06/22

Table of Contents

PREFACE

Welcome to *Collaborative Statistics*, presented by Connexions. The initial section below introduces you to Connexions. If you are familiar with Connexions, please skip to **About "Collaborative Statistics."**

About Connexions

Connexions Modular Content

Connexions (**cnx.org (http://cnx.org/)**) is an online, **open access** educational resource dedicated to providing high quality learning materials free online, free in printable PDF format, and at low cost in bound volumes through print-on-demand publishing. The *Collaborative Statistics* textbook is one of many **collections** available to Connexions users. Each **collection** is composed of a number of re-usable learning **modules** written in the Connexions XML markup language. Each module may also be re-used (or 're-purposed') as part of other collections and may be used outside of Connexions. Including *Collaborative Statistics*, Connexions currently offers over 6500 modules and more than 350 collections.

The modules of *Collaborative Statistics* are derived from the original paper version of the textbook under the same title, *Collaborative Statistics*. Each module represents a self-contained concept from the original work. Together, the modules comprise the original textbook.

Re-use and Customization

The **Creative Commons (CC) Attribution license (http://creativecommons.org/licenses/by/2.0/)** applies to all Connexions modules. Under this license, any module in Connexions may be used or modified for any purpose as long as proper attribution to the original author(s) is maintained. Connexions' authoring tools make re-use (or re-purposing) easy. Therefore, instructors anywhere are permitted to create customized versions of the *Collaborative Statistics* textbook by editing modules, deleting unneeded modules, and adding their own supplementary modules. Connexions' authoring tools keep track of these changes and maintain the CC license's required attribution to the original authors. This process creates a new collection that can be viewed online, downloaded as a single PDF file, or ordered in any quantity by instructors and students as a low-cost printed textbook. To start building custom collections, please visit the help page, **“Create a Collection with Existing Modules” (http://cnx.org/help/CreateCollection)** . For a guide to authoring modules, please look at the help page, **“Create a Module in Minutes” (http://cnx.org/help/ModuleInMinutes)** .

Read the book online, print the PDF, or buy a copy of the book.

To browse the *Collaborative Statistics* textbook online, visit the collection home page at **cnx.org/content/col10522/latest (http://cnx.org/content/col10522/latest/)** . You will then have three options.

1. You may obtain a PDF of the entire textbook to print or view offline by clicking on the “Download PDF” link in the “Content Actions” box.
2. You may order a bound copy of the collection by clicking on the “Order Printed Copy” button.
3. You may view the collection modules online by clicking on the “Start >>” link, which takes you to the first module in the collection. You can then navigate through the subsequent modules by using their “Next >>” and “Previous >>” links to move forward and backward in the collection. You can jump to any module in the collection by clicking on that module’s title in the “Collection Contents” box on the left side of the window. If these contents are hidden, make them visible by clicking on “[show table of contents]”.

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit **http://cnx.org/content/m17212/latest/ (http://cnx.org/content/m17212/latest/)** .
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit **http://cnx.org/content/m17211/latest/ (http://cnx.org/content/m17211/latest/)** .

Version Change History and Errata

- For a list of modifications, updates, and corrections, please visit **http://cnx.org/content/m17360/latest/ (http://cnx.org/content/m17360/latest/)** .

Adoption and Usage

- The Collaborative Statistics collection has been adopted and customized by a number of professors and educators for use in their classes. For a list of known versions and adopters, please visit **http://cnx.org/content/m18261/latest/ (http://cnx.org/content/m18261/latest/)** .

About “Collaborative Statistics”

Collaborative Statistics was written by Barbara Illowsky and Susan Dean, faculty members at De Anza College in Cupertino, California. The textbook was developed over several years and has been used in regular and honors-level classroom settings and in distance learning classes. Courses using this textbook have been articulated by the University of California for transfer of credit. The textbook contains full materials for course offerings, including expository text, examples, labs, homework, and projects. A Teacher’s Guide is currently available in print form and on the Connexions site at **http://cnx.org/content/col10547/latest/ (http://cnx.org/content/col10547/latest/)** , and supplemental course materials including additional problem sets and video lectures are available at **http://cnx.org/content/col10586/latest/ (http://cnx.org/content/col10586/latest/)** . The on-line text for each of these collections collections will meet the Section 508 standards for accessibility.

An on-line course based on the textbook was also developed by Illowsky and Dean. It has won an award as the best on-line California community college course. The on-line course will be available at a later date as a collection in Connexions, and each lesson in the on-line course will be linked to the on-line textbook chapter. The on-line course will include, in addition to expository text and examples, videos of course lectures in captioned and non-captioned format.

The original preface to the book as written by professors Illowsky and Dean, now follows:

This book is intended for introductory statistics courses being taken by students at two– and four–year colleges who are majoring in fields other than math or engineering. Intermediate algebra is the only prerequisite. The book focuses on applications of statistical knowledge rather than the theory behind it. The text is named *Collaborative Statistics* because students learn best by **doing**. In fact, they learn best by working in small groups. The old saying “two heads are better than one” truly applies here.

Our emphasis in this text is on four main concepts:

- thinking statistically
- incorporating technology
- working collaboratively
- writing thoughtfully

These concepts are integral to our course. Students learn the best by actively participating, not by just watching and listening. Teaching should be highly interactive. Students need to be thoroughly engaged in the learning process in order to make sense of statistical concepts. *Collaborative Statistics* provides techniques for students to write across the curriculum, to collaborate with their peers, to think statistically, and to incorporate technology.

This book takes students step by step. The text is interactive. Therefore, students can immediately apply what they read. Once students have completed the process of problem solving, they can tackle interesting and challenging problems relevant to today's world. The problems require the students to apply their newly found skills. In addition, technology (TI-83 graphing calculators are highlighted) is incorporated throughout the text and the problems, as well as in the special group activities and projects. The book also contains labs that use real data and practices that lead students step by step through the problem solving process.

At De Anza, along with hundreds of other colleges across the country, the college audience involves a large number of ESL students as well as students from many disciplines. The ESL students, as well as the non-ESL students, have been especially appreciative of this text. They find it extremely readable and understandable. *Collaborative Statistics* has been used in classes that range from 20 to 120 students, and in regular, honor, and distance learning classes.

Susan Dean

Barbara Illowsky

ADDITIONAL RESOURCES

Additional Resources Currently Available

- **Glossary**
- **View or Download This Textbook Online**
- **Collaborative Statistics Teacher's Guide**
- **Supplemental Materials**
- **Video Lectures**
- **Version History**
- **Textbook Adoption and Usage**
- **Additional Technologies and Notes**
- **Accessibility and Section 508 Compliance**

The following section describes some additional resources for learners and educators. These modules and collections are all available on the Connexions website (**http://cnx.org/ (http://cnx.org/)**) and can be viewed online, downloaded, printed, or ordered as appropriate.

Glossary

This module contains the entire glossary for the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The glossary is located at **http://cnx.org/content/m16129/latest/ (http://cnx.org/content/m16129/latest/)** .

View or Download This Textbook Online

The complete contents of this book are available at no cost on the Connexions website at **http://cnx.org/content/col10522/latest/ (http://cnx.org/content/col10522/latest/)** . Anybody can view this content free of charge either as an online e-book or a downloadable PDF file. A low-cost printed version of this textbook is also available **here (http://my.qoop.com/store/7064943342106149/7781159220340)** .

Collaborative Statistics Teacher's Guide

A complementary Teacher's Guide for Collaborative statistics is available through Connexions at **http://cnx.org/content/col10547/latest/ (http://cnx.org/content/col10547/latest/)** . The Teacher's Guide includes suggestions for presenting concepts found throughout the book as well as recommended homework assignments. A low-cost printed version of this textbook is also available **here (http://my.qoop.com/store/7064943342106149/8791310589747)** .

Supplemental Materials

This companion to Collaborative Statistics provides a number of additional resources for use by students and instructors based on the award winning **Elementary Statistics Sofia online course (http://sofia.fhda.edu/gallery/statistics/index.html)** , also by textbook authors Barbara Illowsky and Susan Dean. This content is designed to complement the textbook by providing video tutorials, course management materials, and sample problem sets. The Supplemental Materials collection can be found at **http://cnx.org/content/col10586/latest/ (http://cnx.org/content/col10586/latest/)** .

Video Lectures

- **Video Lecture 1: Sampling and Data (http://cnx.org/content/m17561/latest/)**
- **Video Lecture 2: Descriptive Statistics (http://cnx.org/content/m17562/latest/)**
- **Video Lecture 3: Probability Topics (http://cnx.org/content/m17563/latest/)**
- **Video Lecture 4: Discrete Distributions (http://cnx.org/content/m17565/latest/)**
- **Video Lecture 5: Continuous Random Variables (http://cnx.org/content/m17566/latest/)**
- **Video Lecture 6: The Normal Distribution (http://cnx.org/content/m17567/latest/)**
- **Video Lecture 7: The Central Limit Theorem (http://cnx.org/content/m17568/latest/)**
- **Video Lecture 8: Confidence Intervals (http://cnx.org/content/m17569/latest/)**
- **Video Lecture 9: Hypothesis Testing with a Single Mean (http://cnx.org/content/m17570/latest/)**
- **Video Lecture 10: Hypothesis Testing with Two Means (http://cnx.org/content/m17577/latest/)**
- **Video Lecture 11: The Chi-Square Distribution (http://cnx.org/content/m17571/latest/)**

- **Video Lecture 12: Linear Regression and Correlation (http://cnx.org/content/m17572/latest/)**

Version History

This module contains a listing of changes, updates, and corrections made to the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The Version History is located at **http://cnx.org/content/m17360/latest/ (http://cnx.org/content/m17360/latest/)** .

Textbook Adoption and Usage

This module is designed to track the various derivations of the Collaborative Statistics textbook and its various companion resources, as well as keep track of educators who have adopted various versions for their courses. New adopters are encouraged to provide their contact information and describe how they will use this book for their courses. The goal is to provide a list that will allow educators using this book to collaborate, share ideas, and make suggestions for future development of this text. The Adoption and Usage module is located at **http://cnx.org/content/m18261/latest/ (http://cnx.org/content/m18261/latest/)** .

Additional Technologies

In order to provide the most flexible learning resources possible, we invite collaboration from all instructors wishing to create customized versions of this content for use with other technologies. For instance, you may be interested in creating a set of instructions similar to this collection's calculator notes. If you would like to contribute to this collection, please use the contact the authors with any ideas or materials you have created.

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit **http://cnx.org/content/m17212/latest/ (http://cnx.org/content/m17212/latest/)** .
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit **http://cnx.org/content/m17211/latest/ (http://cnx.org/content/m17211/latest/)** .

AUTHOR ACKNOWLEDGEMENTS

For this second edition, we appreciate the tremendous feedback from De Anza College colleagues and students, as well as from the dozens of faculty around the world who taught out of the first and preliminary editions. We have updated Collaborative Statistics with contributions from many faculty and students. We especially thank Roberta Bloom, who wrote new problems and additional text.

So many students and colleagues have contributed to the text, both the hard copy and open version. We thank the following people for their contributions to the first and/or second editions.

At De Anza College:
Dr. Inna Grushko (deceased), who wrote the glossary; Diane Mathios, who checked every homework problem in the first edition; Kathy Plum, Lenore Desilets, Charles Klein, Janice Hector, Frank Snow, Dr. Lisa Markus, Dr. Vladimir Logvinenko (deceased), Mo Geraghty, Rupinder Sekhon, Javier Rueda, Carol Olmstead; Also, Dr. Jim Lucas and Valerie Hauber of De Anza's Office of Institutional Research, Mary Jo Kane of Health Services; and the thousands of students who have used this text. Many of the students gave us permission to include their outstanding word problems as homework.

Additional thanks:
Dr. Larry Green of Lake Tahoe Community College, Terrie Teegarden of San Diego Mesa College, Ann Flanigan of Kapiolani Community College, Birgit Aquilonius of West Valley College.

The conversion from a for-profit hard copy text to a free open textbook is the result of many individuals and organizations. We particularly thank Dr. Martha Kanter, Hal Plotkin, Dr. Judy Baker, Dr. Robert Maxfield of Maxfield Foundation, Hewlett Foundation, and Connexions.

Finally, we owe much to Frank, Jeffrey, and Jessica Dean and to Dan, Rachel, Matthew, and Rebecca Illowsky, who encouraged us to continue with our work and who had to hear more than their share of "I'm sorry, I can't" and "Just a minute, I'm working."

STUDENT WELCOME LETTER

Dear Student:

Have you heard others say, "You're taking statistics? That's the hardest course I ever took!" They say that, because they probably spent the entire course confused and struggling. They were probably lectured to and never had the chance to experience the subject. You will not have that problem. Let's find out why.

There is a Chinese Proverb that describes our feelings about the field of statistics:

I HEAR, AND I FORGET

I SEE, AND I REMEMBER

I DO, AND I UNDERSTAND

Statistics is a "do" field. In order to learn it, you must "do" it. We have structured this book so that you will have hands-on experiences. They will enable you to truly understand the concepts instead of merely going through the requirements for the course.

What makes this book different from other texts? First, we have eliminated the drudgery of tedious calculations. You might be using computers or graphing calculators so that you do not need to struggle with algebraic manipulations. Second, this course is taught as a collaborative activity. With others in your class, you will work toward the common goal of learning this material.

Here are some hints for success in your class:

- Work hard and work every night.
- Form a study group and learn together.

- Don’t get discouraged - you can do it!
- As you solve problems, ask yourself, “Does this answer make sense?”
- Many statistics words have the same meaning as in everyday English.
- Go to your teacher for help as soon as you need it.
- Don’t get behind.
- Read the newspaper and ask yourself, “Does this article make sense?”
- Draw pictures - they truly help!

Good luck and don’t give up!

Sincerely,
Susan Dean and Barbara Illowsky

De Anza College
21250 Stevens Creek Blvd.
Cupertino, California 95014

1 SAMPLING AND DATA

1.1 Sampling and Data

Student Learning Outcomes

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

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

Introduction

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

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

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

1.2 Statistics

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

Optional Collaborative Classroom Exercise

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

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

The dot plot for this data would be as follows:

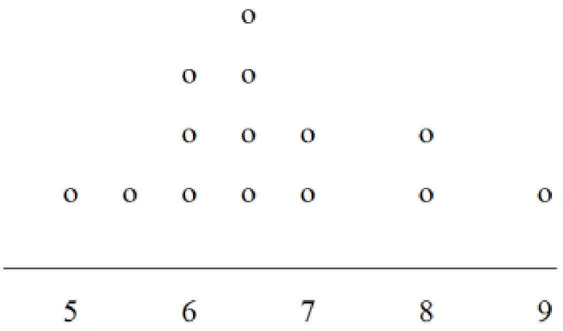


Figure 1.1 Frequency of Average Time (in Hours) Spent Sleeping per Night

Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

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

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

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

1.3 Probability

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

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

1.4 Key Terms

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

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

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

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

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

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

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

Mean and Average

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

Example 1.1

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

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Optional Collaborative Classroom Exercise

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

1.5 Data

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

- Qualitative
- Quantitative

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

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

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

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

Example 1.2 Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example 1.3 Data Sample of Quantitative Continuous Data

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

Example 1.4 Data Sample of Qualitative Data

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

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example 1.5

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

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.
10. Weight of sumo wrestlers.
11. Amount of money (in dollars) won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

1.6 Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

Table 1.1 Class Roster

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

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

.94360; .99832; .14669; .51470; .40581; .73381; .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

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

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

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

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

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

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

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

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

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

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To 4 decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To 4 decimal places, these numbers are not equivalent.

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

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

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

Example 1.6

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

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

1. stratified
2. cluster
3. stratified

4. systematic
5. simple random
6. convenience

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

Example 1.7

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

Suppose we take two different samples.

First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:

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

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:

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

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

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

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

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

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

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

Is the sample biased?

The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

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

Optional Collaborative Classroom Exercise

Exercise 1.7

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

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

1.7 Variation

Variation in Data

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

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

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

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

Variation in Samples

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

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

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

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

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

Optional Collaborative Classroom Exercise

Exercise 1.8

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice.** Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Table 1.2 First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.3 Second Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

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

Critical Evaluation

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

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

1.8 Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round only the final answer. Do not round any intermediate results, if possible. If it becomes necessary to round intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores 4, 6, 9 is 6.3, rounded to the nearest tenth, because the data are whole numbers. Most answers will be rounded in this manner.

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

1.9 Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3

Below is a frequency table listing the different data values in ascending order and their frequencies.

Table 1.4 Frequency Table of Student Work Hours

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

A **frequency** is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the fraction or proportion of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Table 1.5 Frequency Table of Student Work Hours w/ Relative Frequency

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

The sum of the relative frequency column is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

Table 1.6 Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

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

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table 1.7 Frequency Table of Soccer Player Height

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

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

- 59.95 - 61.95 inches
- 61.95 - 63.95 inches
- 63.95 - 65.95 inches
- 65.95 - 67.95 inches
- 67.95 - 69.95 inches
- 69.95 - 71.95 inches
- 71.95 - 73.95 inches
- 73.95 - 75.95 inches

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

In this sample, there are **5** players whose heights are between 59.95 - 61.95 inches, **3** players whose heights fall within the interval 61.95 - 63.95 inches, **15** players whose heights fall within the interval 63.95 - 65.95 inches, **40** players whose heights fall within the interval 65.95 - 67.95 inches, **17** players whose heights fall within the interval 67.95 - 69.95 inches, **12** players whose heights fall within the interval 69.95 - 71.95, 7 players whose height falls within the interval 71.95 - 73.95, and **1** player whose height falls within the interval 73.95 - 75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example 1.8

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

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

Example 1.9

From the table, find the percentage of heights that fall between 61.95 and 65.95 inches.

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Example 1.10

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

1. The percentage of heights that are from 67.95 to 71.95 inches is:
2. The percentage of heights that are from 67.95 to 73.95 inches is:
3. The percentage of heights that are more than 65.95 inches is:
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

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

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Optional Collaborative Classroom Exercise

Exercise 1.12

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

1. What percentage of the students in your class has 0 siblings?
2. What percentage of the students has from 1 to 3 siblings?
3. What percentage of the students has fewer than 3 siblings?

Example 1.11

Nineteen people were asked how many miles, to the nearest mile they commute to work each day. The data are as follows:

2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10

The following table was produced:

Table 1.8 Frequency of Commuting Distances

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

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute 3 miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute 5 or 7 miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (does not include 5 and 13 miles)?

1. No. Frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. Frequency for 3 miles should be 1; for 2 miles (left out), 2. Cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.
3. $\frac{5}{19}$
4. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

1.10 Summary

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
  - Discrete (You count it.)
  - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement**: A member of the population may be chosen more than once
- **Without Replacement**: A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
  - Simple random sample
  - Stratified sample
  - Cluster sample
  - Systematic sample
- Not Random
  - Convenience sample

Frequency (freq. or f)

- The number of times an answer occurs

Relative Frequency (rel. freq. or RF)

- The proportion of times an answer occurs
- Can be interpreted as a fraction, decimal, or percent

Cumulative Relative Frequencies (cum. rel. freq. or cum RF)

- An accumulation of the previous relative frequencies

1.11 Practice: Sampling and Data

Student Learning Outcomes

- The student will construct frequency tables.
- The student will differentiate between key terms.
- The student will compare sampling techniques.

Given

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

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

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

Organize the Data

Complete the tables below using the data provided.

Table 1.9 Researcher A

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

Table 1.10 Researcher B

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

Key Terms

Define the key terms based upon the above example for Researcher A.

Discussion Questions

Discuss the following questions and then answer in complete sentences.

1.12 Homework

Try these multiple choice questions

The next four questions refer to the following: A Lake Tahoe Community College instructor is interested in the average number of days Lake Tahoe Community College math students are absent from class during a quarter.

The next two questions refer to the following relative frequency table on hurricanes that have made direct hits on the U.S between 1851 and 2004. Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. (*http://www.nhc.noaa.gov/gifs/table5.gif* [**url (http://www.nhc.noaa.gov/gifs/table5.gif)**])

Table 1.11 Frequency of Hurricane Direct Hits

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

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

Exercises 28 and 29 are not multiple choice exercises.

1.13 Lab 1: Data Collection

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct Relative Frequency Tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey

Ask five classmates from a different class how many movies they saw last month at the theater. Do not include rented movies.

1. Record the data
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four people's names on the class list. Mark that person's name. Continue doing this until you have marked 12 people's names. You may need to go back to the start of the list. For each marked name record below the five data values. You now have a total of 60 data values.
3. For each name marked, record the data:

Table 1.12

______	______	______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______	______	______

Order the Data

Complete the two relative frequency tables below using your class data.

Table 1.13 Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

Table 1.14 Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1			
2-3			
4-5			
6-7+			

1. Using the tables, find the percent of data that is at most 2. Which table did you use and why?
2. Using the tables, find the percent of data that is at most 3. Which table did you use and why?
3. Using the tables, find the percent of data that is more than 2. Which table did you use and why?
4. Using the tables, find the percent of data that is more than 3. Which table did you use and why?

Discussion Questions

1. Is one of the tables above "more correct" than the other? Why or why not?
2. In general, why would someone group the data in different ways? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

1.14 Lab 2: Sampling Experiment

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain each of the details of each procedure used.

In this lab, you will be asked to pick several random samples. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained

The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

A Simple Random Sample

Pick a **simple random sample** of 15 restaurants.

1. Describe the procedure:

2.

Table 1.15

1. ___________	6. ___________	11. ___________
2. ___________	7. ___________	12. ___________
3. ___________	8. ___________	13. ___________
4. ___________	9. ___________	14. ___________
5. ___________	10. ___________	15. ___________

A Systematic Sample

Pick a **systematic sample** of 15 restaurants.

1. Describe the procedure:

2.

Table 1.16

1. ___________	6. ___________	11. ___________
2. ___________	7. ___________	12. ___________
3. ___________	8. ___________	13. ___________
4. ___________	9. ___________	14. ___________
5. ___________	10. ___________	15. ___________

A Stratified Sample

Pick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe the procedure:

2.

Table 1.17

1. ___________	6. ___________	11. ___________	16. ___________
2. ___________	7. ___________	12. ___________	17. ___________
3. ___________	8. ___________	13. ___________	18. ___________
4. ___________	9. ___________	14. ___________	19. ___________
5. ___________	10. ___________	15. ___________	20. ___________

A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe the procedure:

2.

Table 1.18

1. ___________	6. ___________	11. ___________	16. ___________
2. ___________	7. ___________	12. ___________	17. ___________
3. ___________	8. ___________	13. ___________	18. ___________
4. ___________	9. ___________	14. ___________	19. ___________
5. ___________	10. ___________	15. ___________	20. ___________
			21. ___________

A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe the procedure:

2.

Table 1.19

1. ___________	6. ___________	11. ___________	16. ___________	21. ___________
2. ___________	7. ___________	12. ___________	17. ___________	22. ___________
3. ___________	8. ___________	13. ___________	18. ___________	23. ___________
4. ___________	9. ___________	14. ___________	19. ___________	24. ___________
5. ___________	10. ___________	15. ___________	20. ___________	25. ___________

Restaurants Stratified by City and Entree Cost

Table 1.20 Restaurants Used in Sample

Entree Cost →	Under $10	$10 to under $15	$15 to under $20	Over $20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, Panda Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

The original lab was designed and contributed by Carol Olmstead.

Glossary

Average: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Continuous Random Variable: A random variable (RV) whose outcomes are measured.

Example .

The height of trees in the forest is a continuous RV.

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

Data: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Data: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Data: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Discrete Random Variable: A random variable (RV) whose outcomes are counted.

Frequency: The number of times a value of the data occurs.

Population: The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Probability: A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1;$.
- If A and B are any two mutually exclusive events, then $P(A\text{or}B) = P(A) + P(B)$.
- $P(S) = 1$.

Proportion: As a number: A proportion is the number of successes divided by the total number in the sample.

- As a probability distribution: Given a binomial random variable (RV), $X \sim B(n, p)$, consider the ratio of the number X of successes in n Bernouli trials to the number n of trials. $P' = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, $P' \sim N\left(p, \frac{pq}{n}\right)$.

Qualitative Data: See **Data**.

Quantitative Data: See **Data**.

Relative Frequency: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Sample: A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Statistic: A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

2 DESCRIPTIVE STATISTICS

2.1 Descriptive Statistics

Student Learning Outcomes

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

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Introduction

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

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

2.2 Displaying Data

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

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

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

2.3 Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis.It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.1

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

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

Table 2.1 Stem-and-Leaf Diagram

Stem	Leaf
3	3
4	299
5	355
6	1378899
7	2348
8	03888
9	0244446
10	0

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stemplot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example 2.2

Create a stem plot using the data:

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

The data are the distance (in kilometers) from a home to the nearest supermarket.

1. Are there any outliers?
2. Do the data seem to have any concentration of values?

Hint

The leaves are to the right of the decimal.

The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

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

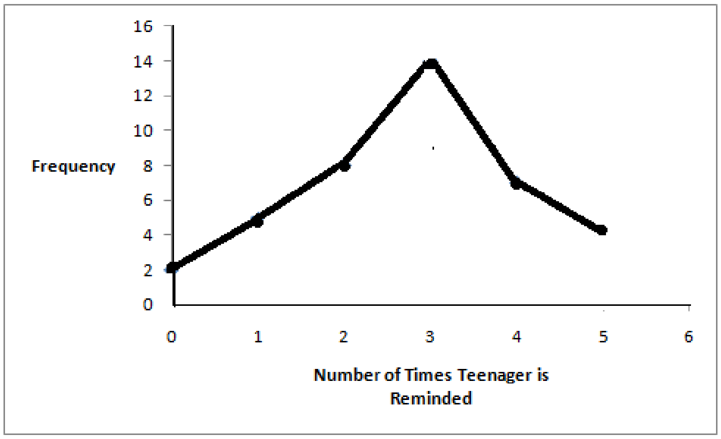
Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequency points**. The frequency points are connected.

Example 2.3

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

Table 2.2

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal.

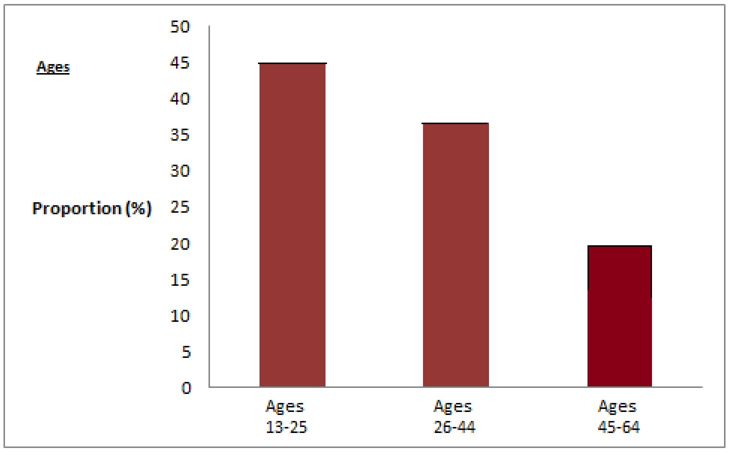
The **bar graph** shown in **Example 4** has age groups represented on the **x-axis** and proportions on the **y-axis**.

Example 2.4

By the end of 2011, in the United States, Facebook had over 146 million users. The table shows three age groups, the number of users in each age group and the proportion (%) of users in each age group. **Source:** ***http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/***

Table 2.3

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

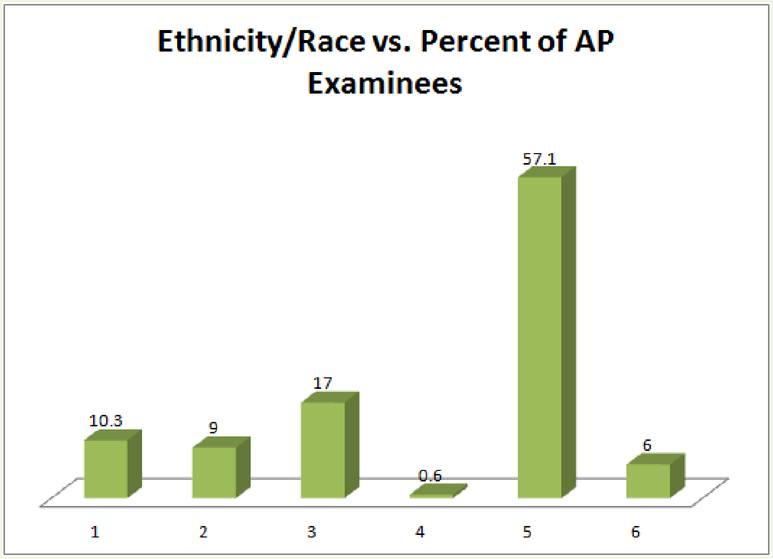


Example 2.5

The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2011, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools (qualitative data) on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**. (**Source: http://www.collegeboard.com** and **Source: http://apreport.collegeboard.org/goals-and-findings/promoting-equity**)

Table 2.4

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



Go to **Outcomes of Education Figure 22 (http://nces.ed.gov/pubs2011/2011015_5.pdf)** for an example of a bar graph that shows unemployment rates of persons 25 years and older for 2009.

This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the **Texas Instruments (TI) website (http://education.ti.com/educationportal/sites/US/sectionHome/support.html)** .

2.4 Histograms

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

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

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on **Sampling and Data**, we defined frequency as the number of times an answer occurs.) If:

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

then:

$$RF = \frac{f}{n} \tag{2.1}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100 % are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 - 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 - 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 - .0005 = 0.9995). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 (2 - 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example 2.6

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

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

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

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

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

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

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

74

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

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

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

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.

$$\frac{74.05 - 59.95}{8} = 1.76 \tag{2.2}$$

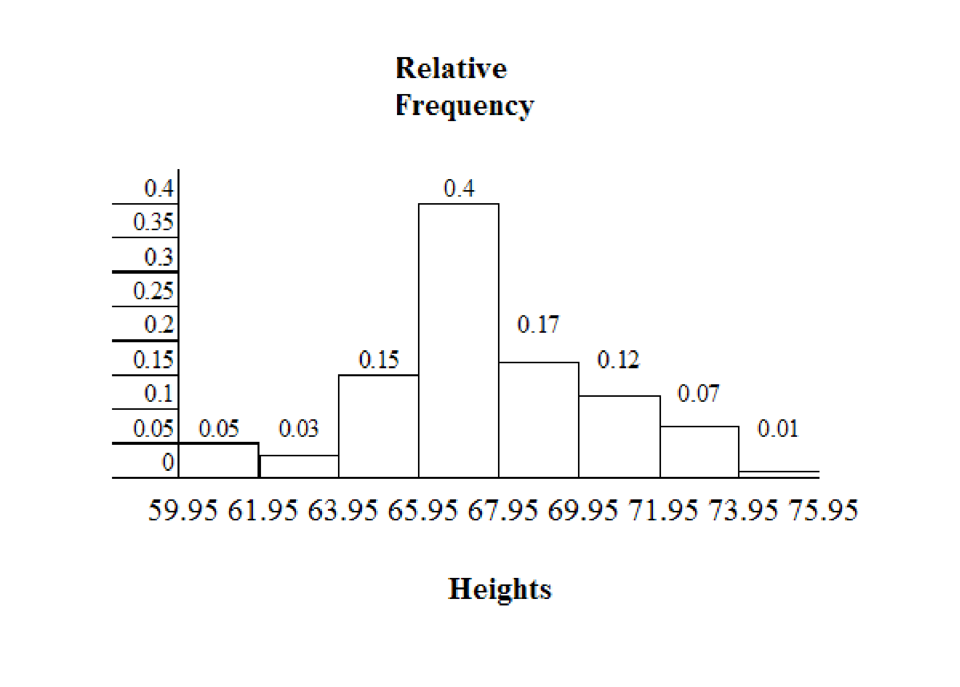
We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

The boundaries are:

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

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

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



Example 2.7

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

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

2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4; 4

5; 5; 5; 5; 5

6; 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.

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

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

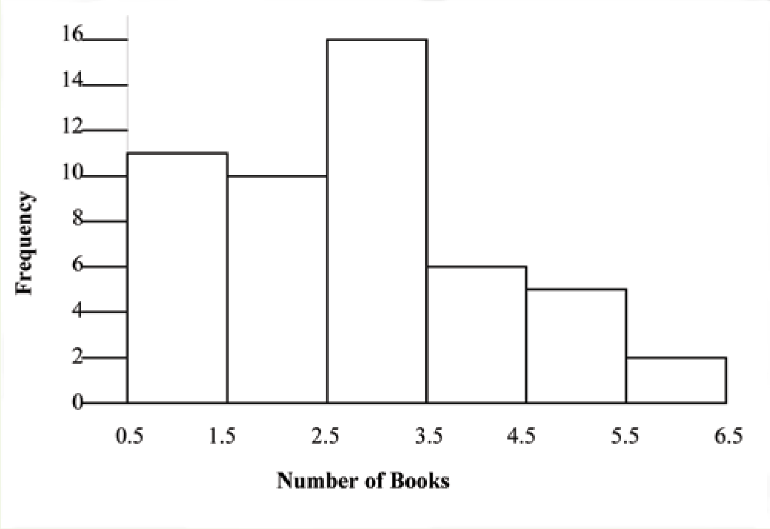
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{bars}} = 1 \quad (2.3)$$

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

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



Optional Collaborative Exercise

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

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

2.5 Box Plots

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

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

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

Ordered from smallest to largest:

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

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.

$$\frac{6.8 + 7.2}{2} = 7 \tag{2.4}$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

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

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

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

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

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

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

7.2; 8; 8.3; **9**; 10; 10; 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

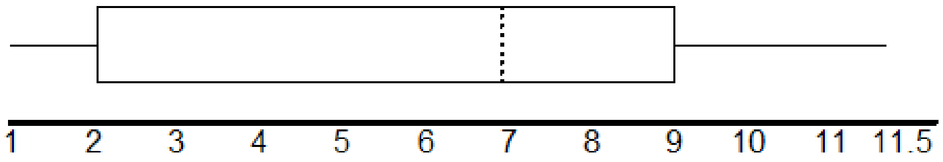
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

You may encounter box and whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider the following data:

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

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the **TI web site (http://education.ti.com/educationportal/sites/US/sectionHome/support.html)**):



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

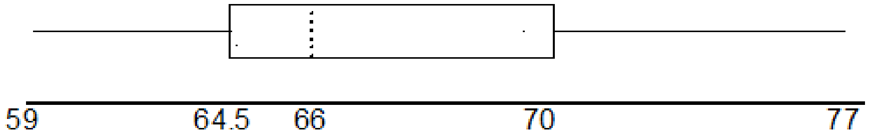
Example 2.8

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

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties:

- Smallest value = 59
- Largest value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median= 66
- Q3: Third quartile = 70



a. Each quarter has 25% of the data.

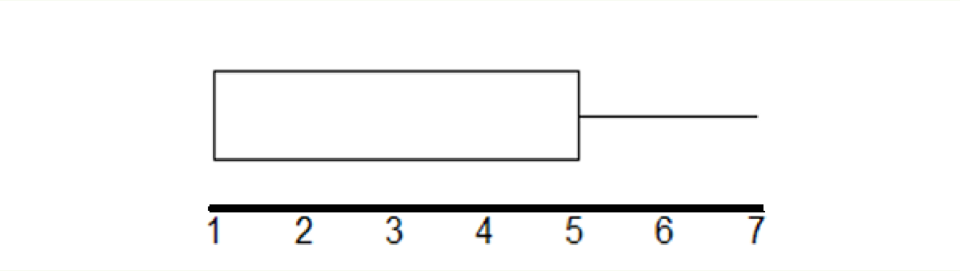
b. The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.

c. Interquartile Range: IQR = Q3 − Q1 = 70 − 64.5 = 5.5.

d. The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.

e. The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example 2.9

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

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

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

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

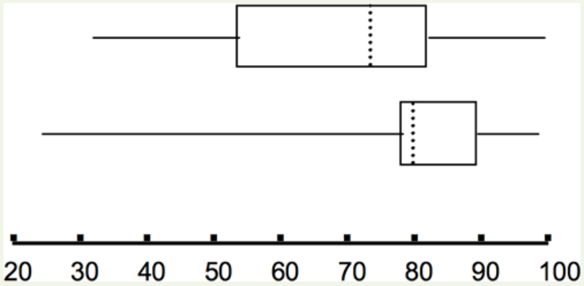
- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

First Data Set

- Xmin = 32
- Q1 = 56
- *M* = 74.5
- Q3 = 82.5
- Xmax = 99

Second Data Set

- Xmin = 25.5
- Q1 = 78
- *M* = 81
- Q3 = 89
- Xmax = 98



The first data set (the top box plot) has the widest spread for the middle 50% of the data. IQR = Q3 − Q1 is 82.5 − 56 = 26.5 for the first data set and 89 − 78 = 11 for the second data set. So, the first set of data has its middle 50% of scores more spread out.

25% of the data is between *M* and Q3 and 25% is between Q3 and Xmax.

2.6 Measures of the Location of the Data

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

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

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

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

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

$$\text{IQR} = Q_3 - Q_1 \tag{2.5}$$

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

Example 2.10

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

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

Order the data from smallest to largest.

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

M = 488,800

$$Q_1 = \frac{230500 + 387000}{2} = 308750$$

$$Q_3 = \frac{639000 + 659000}{2} = 649000$$

IQR = 649000 − 308750 = 340250

(1.5)(IQR) = (1.5)(340250) = 510375

$$Q_1 - \left(1.5\right)\left(\text{IQR}\right) = 308750 - 510375 = -201625$$

$$Q_3 + \left(1.5\right)\left(\text{IQR}\right) = 649000 + 510375 = 1159375$$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2.11

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

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.
c. The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

For the IQRs, see the **answer to the test scores example**. The first data set has the larger IQR, so the scores between Q3 and Q1 (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right) = \left(\frac{3}{2}\right) \cdot \left(26.5\right) = 39.75$
- Xmax - Q3 = 99 - 82.5 = 16.5
- Q1 - Xmin = 56 - 32 = 24

$\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right)$ = 39.75 is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right) = \left(\frac{3}{2}\right) \cdot \left(11\right) = 16.5$
- Xmax − Q3 = 98 − 89 = 9
- Q1 − Xmin = 78 − 25.5 = 52.5

$\left(\frac{3}{2}\right) \cdot \left(\text{IQR}\right)$ = 16.5 is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see **"Frequency" from the Sampling and Data Chapter**). Get the percentiles from that chart.

First Data Set

- 30th %ile (between the 6th and 7th values) = $\frac{(56 + 59)}{2}$ = 57.5
- 80th %ile (between the 16th and 17th values) = $\frac{(84 + 84.5)}{2}$ = 84.25

Second Data Set

- 30th %ile (7th value) = 78
- 80th %ile (18th value) = 90

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Example 2.12 Finding Quartiles and Percentiles Using a Table

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

Table 2.5

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

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

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example 2.13

Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile. What is another name for the first quartile?
4. Construct a box plot of the data.

1. $\frac{(8+9)}{2} = 8.5$
2. 9
3. 6
4. First Quartile = 25th %ile

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

1. How many students were surveyed?
2. What kind of sampling did you do?

3. Find the mean and standard deviation.
4. Find the mode.
5. Construct 2 different histograms. For each, starting value = _____ ending value = ____.
6. Find the median, first quartile, and third quartile.
7. Construct a box plot.
8. Construct a table of the data to find the following:
    - The 10th percentile
    - The 70th percentile
    - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

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

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

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

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

Guideline:

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

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example 2.14

On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 2.15

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example 2.16

At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles

Exercise 2.7

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?

b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

c. A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Solution

a. For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.

b. INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.

c. He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Exercise 2.8

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?

b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.

b. INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise 2.9

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

Solution

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Exercise 2.10

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

Solution

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

Exercise 2.11

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

Solution

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Exercise 2.12

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Solution

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

Exercise 2.13

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

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?

b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Solution

a. The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.

b. The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Exercise 2.14

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

Solution

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

**With contributions from Roberta Bloom

2.7 Measures of the Center of the Data

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

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

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an *x* with a bar over it (pronounced "*x* bar"): $\bar{x}$.

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

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

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

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{2.6}$$

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

In the second example, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter *n* is the total number of data values in the sample. If *n* is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If *n* is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.17

AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

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

Calculate the mean and the median.

The calculation for the mean is:

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

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

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

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

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

$M = \frac{24+24}{2} = 24$

The median is 24.

Example 2.18

Suppose that, in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center," the mean or the median?

$\overline{x} = \frac{5000000 + 49 \times 30000}{50} = 129400$

$M = 30000$

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

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

Another measure of the center is the mode. The **mode** is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example 2.19 Statistics exam scores for 20 students are as follows

Statistics exam scores for 20 students are as follows:

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

Find the mode.

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

Example 2.20

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

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

The mode can be calculated for qualitative data as well as for quantitative data.

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

The Law of Large Numbers and the Mean

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

The formula for the mean is located in the **Summary of Formulas** section course.

Sampling Distributions and Statistic of a Sampling Distribution

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

Table 2.6

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

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

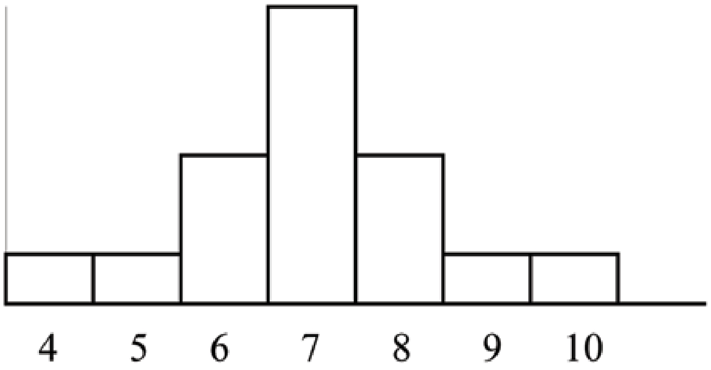
A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

2.8 Skewness and the Mean, Median, and Mode

Consider the following data set:

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

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.

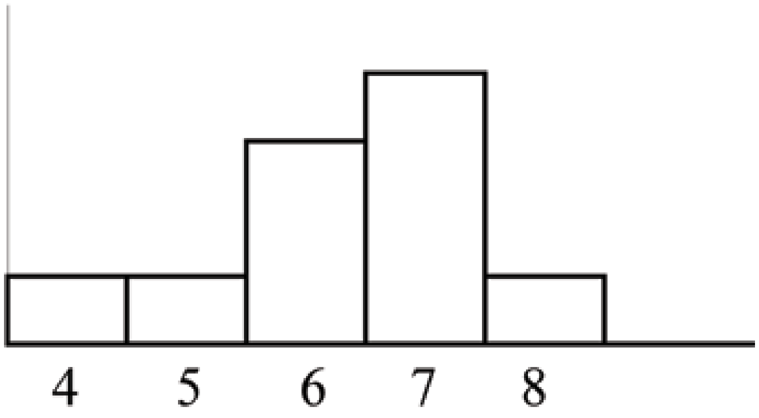


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

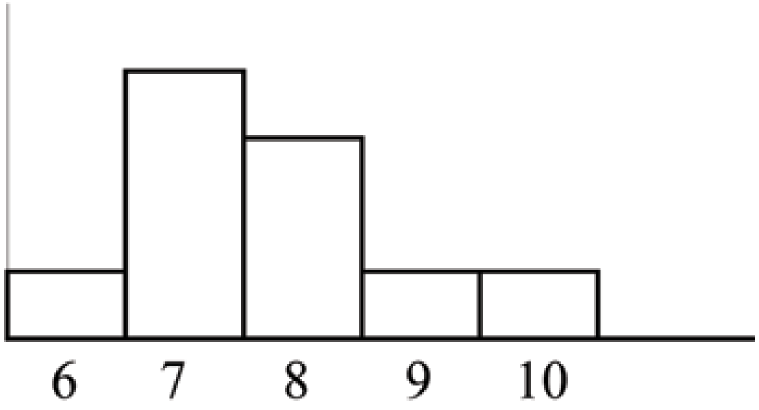


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

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

2.9 Measures of the Spread of the Data

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

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

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

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

The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

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

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

Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

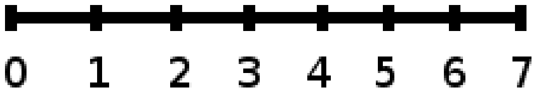
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

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

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



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7=5+**(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1=5+**(−2)**(2)

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

- **sample:** $x = \bar{x}$ + (#ofSTDEV)(s)
- **Population:** $x = \mu$ + (#ofSTDEV)(σ)

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

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

Calculating the Standard Deviation

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

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

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

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

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x - \bar{x})^2}{n - 1}}$

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

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma\left(x - \mu\right)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot \left(x - \mu\right)^2}{N}}$
- For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is 1. If a value appears three times in the data set or population, *f* is 3.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and *n* is the size of the sample.

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example 2.21

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of *n* = 20 fifth grade students. The ages are rounded to the nearest half year:

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

$$\bar{x} = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525 \tag{2.8}$$

The average age is 10.53 years, rounded to 2 places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Table 2.7

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
x	f	$\left(x - \bar{x}\right)$	$\left(x - \bar{x}\right)^2$	$(f)\left(x - \bar{x}\right)^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1):

$s^2 = \frac{9.7375}{20 - 1} = 0.5125$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = .0715891$ Rounded to two decimal places, $s = 0.72$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Verify the mean and standard deviation calculated above on your calculator or computer.

For the TI-83,83+,84+, enter data into the list editor.

Put the data values in list L1 and the frequencies in list L2.
STAT CALC 1-VarStats L1, L2
$\bar{x}$=10.525
Use Sx because this is sample data (not a population): Sx=.715891

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)(σ)
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

Find the value that is 1 standard deviation above the mean. Find $\left(\bar{x} + 1s\right)$.

$\left(\bar{x} + 1s\right) = 10.53 + \left(1\right)\left(0.72\right) = 11.25$

Find the value that is two standard deviations below the mean. Find $\left(\bar{x} - 2s\right)$.

$\left(\bar{x} - 2s\right) = 10.53 - \left(2\right)\left(0.72\right) = 9.09$

Find the values that are 1.5 standard deviations **from** (below and above) the mean.

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

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero**. (For this example, there are n=20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n=20, the calculation divided by n-1=20-1=19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n-1). Why not divide by *n*? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n-1) gives a better estimate of the population variance.

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, *s* or σ, is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make *s* or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**.

The formula for the standard deviation is at the end of the chapter.

Example 2.22

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

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

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
i. The sample mean
ii. The sample standard deviation
iii. The median
iv. The first quartile
v. The third quartile
vi. IQR

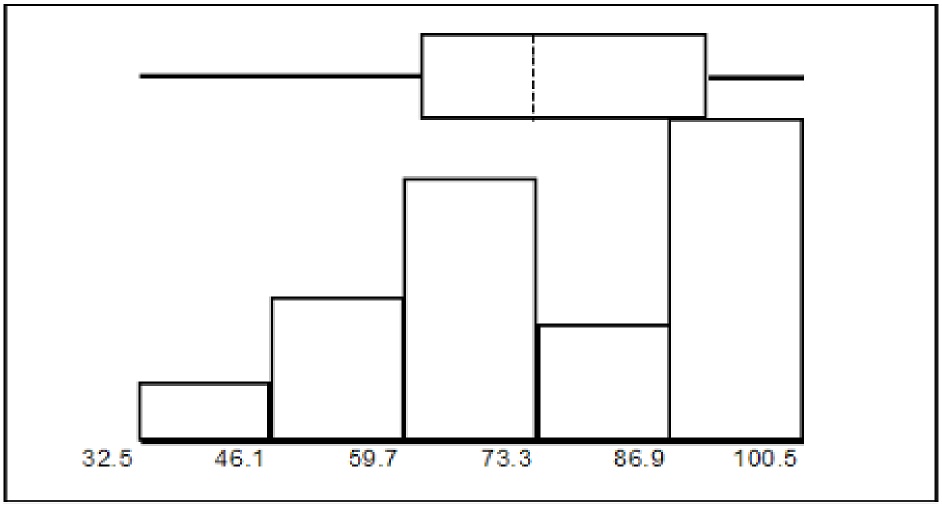
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

a.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	**0.998** (Why isn't this value 1?)

b.
i. The sample mean = 73.5
ii. The sample standard deviation = 17.9
iii. The median = 73
iv. The first quartile = 61
v. The third quartile = 90
vi. IQR = 90 - 61 = 29

c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 - 33 = 40) than the spread in the upper 50% (100 - 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, it can be misleading to compare the data values directly.

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\# \text{ ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Table 2.8

Sample	$x = \bar{x} + z\,s$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\,\sigma$	$z = \frac{x - \mu}{\sigma}$

Example 2.23

Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$\#\text{ ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}\text{ ; } z = \frac{x - \mu}{\sigma}$$

For John, $z = \#\text{ ofSTDEVs} = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ ofSTDEVs} = \frac{77 - 80}{10} = -0.3$

John has the better G.P.A. when compared to his school because his G.P.A. is 0.21 standard deviations **below** his school's mean while Ali's G.P.A. is 0.3 standard deviations **below** his school's mean.

John's z-score of −0.21 is higher than Ali's z-score of −0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

2.10 Summary of Formulas

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- $\overline{x}$ = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x * f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x * f$ = The sum of values multiplied by their respective frequencies
- $\left(x - \overline{x}\right)$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean)
- $\left(x - \overline{x}\right)^2$ or $(x - \mu)^2$ = Deviations squared
- $f\left(x - \overline{x}\right)^2$ or $f(x - \mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $\overline{x} = \frac{\sum x}{n}$ or $\overline{x} = \frac{\sum f \cdot x}{n}$

- $\mu = \frac{\sum x}{N}$ or $\mu = \frac{\sum f \cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot \left(x - \bar{x}\right)^2}{n-1}}$
- $\sigma = \sqrt{\frac{\Sigma\left(x - \mu\right)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot \left(x - \mu\right)^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- $X = \bar{x}$+ (#ofSTDEVs)(s)
- $X = \mu$ + (#ofSTDEVs)(σ)

2.11 Practice 1: Center of the Data

Student Learning Outcomes

- The student will calculate and interpret the center, spread, and location of the data.
- The student will construct and interpret histograms an box plots.

Given

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

Complete the Table

Table 2.9

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Discussion Questions

Enter the Data

Enter your data into your calculator or computer.

Construct a Histogram

Determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram below. Label the horizontal and vertical axes with words. Include numerical scaling.

Data Statistics

Calculate the following values:

Calculations

Use the table in section 2.11.3 to calculate the following values:

Box Plot

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

Interpretation

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

2.12 Practice 2: Spread of the Data

Student Learning Outcomes

- The student will calculate measures of the center of the data.
- The student will calculate the spread of the data.

Given

The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976-77 through 2004-2005. (*Source: Graphically Speaking by Bill King, LTCC Institutional Research, December 2005*).

Use these values to answer the following questions:

- μ = 1000 FTES
- Median - 1014 FTES
- σ = 474 FTES
- First quartile = 528.5 FTES
- Third quartile = 1447.5 FTES
- n = 29 years

Calculate the Values

2.13 Homework

Try these multiple choice questions (Exercises 24 - 30).

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

Table 2.10

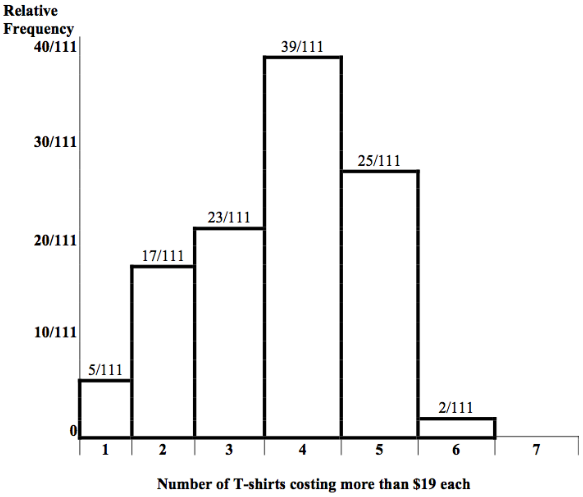
Number of years	Frequency
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

The next two questions refer to the following table. *X* = the number of days per week that 100 clients use a particular exercise facility.

Table 2.11

X	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

The next two questions refer to the following histogram. Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



**Exercises 32 and 33 contributed by Roberta Bloom

2.14 Lab: Descriptive Statistics

Class Time:

Names:

Student Learning Outcomes

- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data

Record the number of pairs of shoes you own:

1. Randomly survey 30 classmates. Record their values.

Table 2.12 Survey Results

______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______

2. Construct a histogram. Make 5-6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

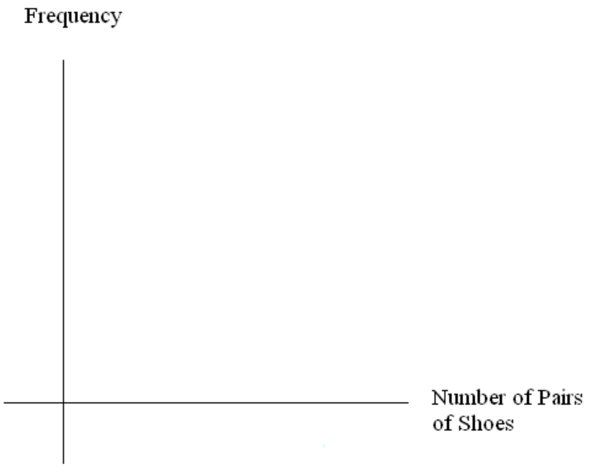


Figure 2.1

3. Calculate the following:
   - $\overline{x}$ =
   - s =
4. Are the data discrete or continuous? How do you know?
5. Describe the shape of the histogram. Use complete sentences.
6. Are there any potential outliers? Which value(s) is (are) it (they)? Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following:
   - Minimum value =
   - Median =
   - Maximum value =
   - First quartile =
   - Third quartile =
   - IQR =
2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:

a. Above the mean:

b. Below the mean:

Glossary

Frequency: The number of times a value of the data occurs.

Interquartile Range (IRQ): The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

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

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

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

Mode: The value that appears most frequently in a set of data.

Outlier: An observation that does not fit the rest of the data.

Outlier: An observation that does not fit the rest of the data.

Percentile: A number that divides ordered data into hundredths.

Example .

Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0...\}$. Then the first percentile is $\frac{(2.7 + 2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9 + 2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

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

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

Relative Frequency: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variance: Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

3 PROBABILITY TOPICS

3.1 Probability Topics

Student Learning Outcomes

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

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams (optional).
- Construct and interpret Tree Diagrams (optional).

Introduction

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn to solve probability problems using a systematic approach.

Optional Collaborative Classroom Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. P(change) means the probability that a randomly chosen person in your class has change in his/her pocket or purse. P(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find P(change).
- Find P(bus).
- Find P(change and bus) Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find P(change| bus) Find the probability that a randomly chosen student has change given that he/she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

3.2 Terminology

Probability measures the uncertainty that is associated with the outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin is an example of an experiment.

The result of an experiment is called an **outcome**. A **sample space** is a set of all possible outcomes. Three ways to represent a sample space are to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is used to denote the sample space. For example, if you flip one fair coin, S = {H, T} where *H* = heads and *T* = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like *A* and *B* represent events. For example, if the experiment is to flip one fair coin, event *A* might be getting at most one head. The probability of an event *A* is written P(A).

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between 0 and 1, inclusive** (includes 0 and 1 and all numbers between these values). P(A) = 0 means the event *A* can never happen. P(A) = 1 means the event *A* always happens. P(A) = 0.5 means the event *A* is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative fequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head(H) and a Tail(T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is {HH, TH, HT, TT} where *T* = tails and *H* = heads. The sample space has four outcomes. *A* = getting one head. There are two outcomes {HT, TH}. $P(A) = \frac{2}{4}$.

Suppose you roll one fair six-sided die, with the numbers {1,2,3,4,5,6} on its faces. Let event *E* = rolling a number that is at least 5. There are two outcomes {5, 6}. P(E) =$\frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, 2/6 of the rolls would result an outcome of "at least 5". The long-term relative frequency of obtaining this result would approach the theoretical probability of 2/6 as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is the known as the **Law of Large Numbers**: as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes don't happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.) The Law of Large Numbers will be discussed again in Chapter 7.

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased** . Two math professors in Europe had their statistics students test the Belgian 1 Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos have a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later in this chapter we will learn techniques to use to work with probabilities for events that are not equally likely.

"OR" Event:

An outcome is in the event *A* OR *B* if the outcome is in *A* or is in *B* or is in both *A* and *B*. For example, let A = {1, 2, 3, 4, 5} and B = {4, 5, 6, 7, 8}. *A* OR B = {1, 2, 3, 4, 5, 6, 7, 8}. Notice that 4 and 5 are NOT listed twice.

"AND" Event:

An outcome is in the event A AND B if the outcome is in both *A* and *B* at the same time. For example, let *A* and *B* be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then A AND B = {4, 5}.

The **complement** of event *A* is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in *A*. Notice that P(A) + P(A') = 1. For example, let S = {1, 2, 3, 4, 5, 6} and let A = {1, 2, 3, 4}. Then, A' = {5, 6}. P(A) =$\frac{4}{6}$, P(A') =$\frac{2}{6}$, and P(A) + P(A') =$\frac{4}{6}+\frac{2}{6}= 1$

The **conditional probability** of *A* given *B* is written P(A|B). P(A|B) is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate P(A|B) is

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

where P(B) is greater than 0.

For example, suppose we toss one fair, six-sided die. The sample space S = {1, 2, 3, 4, 5, 6}. Let *A* = face is 2 or 3 and *B* = face is even (2, 4, 6). To calculate P(A|B), we count the number of outcomes 2 or 3 in the sample space B = {2, 4, 6}. Then we divide that by the number of outcomes in *B* (and not *S*).

We get the same result by using the formula. Remember that *S* has 6 outcomes.

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

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Exercise 3.1

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events f or parts (a) through (j) below. (Note that you can't find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.

b. The probability that a student is male or has short hair.

c. The probability that a student is a female and has long hair.

d. The probability that a student is male, given that the student has long hair.

e. The probability that a student is has long hair, given that the student is male.

f. Of all the female students, the probability that a student has short hair.

g. Of all students with long hair, the probability that a student is female.

h. The probability that a student is female or has long hair.

i. The probability that a randomly selected student is a male student with short hair.

j. The probability that a student is female.

Solution

a. P(L')=P(S)
b. P(M or S)
c. P(F and L)
d. P(M|L)
e. P(L|M)
f. P(S|F)
g. P(F|L)
h. P(F or L)
i. P(M and S)
j. P(F)

**With contributions from Roberta Bloom

3.3 Independent and Mutually Exclusive Events

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- P(A|B) = P(A)
- P(B|A) = P(B)
- P(A AND B) = P(A) · P(B)

Two events *A* and *B* are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:**: When sampling is done without replacement, then each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

Mutually Exclusive Events

A and *B* are **mutually exclusive** events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and P(A AND B) = 0.

For example, suppose the sample space S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let A = {1, 2, 3, 4, 5}, B = {4, 5, 6, 7, 8}, and C = {7, 9}. A AND B = {4, 5}. P(A AND B) = $\frac{2}{10}$and is not equal to zero. Therefore, *A* and *B* are not mutually exclusive. *A* and *C* do not have any numbers in common so P(A AND C) = 0. Therefore, *A* and *C* are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**.

The following examples illustrate these definitions and terms.

Example 3.1

Flip two fair coins. (This is an experiment.)

The sample space is {HH, HT, TH, TT} where *T* = tails and *H* = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let *A* = the event of getting **at most one tail**. (At most one tail means 0 or 1 tail.) Then *A* can be written as {HH, HT, TH}. The outcome HH shows 0 tails. HT and TH each show 1 tail.
- Let *B* = the event of getting all tails. *B* can be written as {TT}. *B* is the **complement** of *A*. So, B = A'. Also, P(A) + P(B) = P(A) + P(A') = 1.
- The probabilities for *A* and for *B* are P(A) = $\frac{3}{4}$ and P(B) = $\frac{1}{4}$.
- Let *C* = the event of getting all heads. C = {HH}. Since B = {TT}, P(B AND C) = 0. *B* and *C* are mutually exclusive. (*B* and *C* have no members in common because you cannot have all tails and all heads at the same time.)
- Let *D* = event of getting **more than one** tail. *D* = {TT}. P(D) = $\frac{1}{4}$.
- Let *E* = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) *E* = {HT, HH}. P(E) = $\frac{2}{4}$.

- Find the probability of getting **at least one** (1 or 2) tail in two flips. Let *F* = event of getting at least one tail in two flips. *F* = {HT, TH, TT}. $P(F) = \frac{3}{4}$

Example 3.2

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

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

Are *C* and *E* mutually exclusive events? (Answer yes or no.) Why or why not?

No. *C* = {3, 5} and *E* = {1, 2, 3, 4}. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, P(C AND E) must be 0.

- Find P(C|A). This is a conditional. Recall that the event *C* is {3, 5} and event *A* is {1, 3, 5}. To find P(C|A), find the probability of *C* using the sample space *A*. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$

Example 3.3

Let event *G* = taking a math class. Let event *H* = taking a science class. Then, G AND H = taking a math class and a science class. Suppose P(G) = 0.6, P(H) = 0.5, and P(G AND H) = 0.3. Are *G* and *H* independent?

If *G* and *H* are independent, then you must show **ONE** of the following:

- P(G|H) = P(G)
- P(H|G) = P(H)
- P(G AND H) = P(G) · P(H)

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

Show that P(G|H) = P(G).

$$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$$

Show P(G AND H) = P(G) · P(H).

P(G) · P(H) = 0.6 · 0.5 = 0.3 = P(G AND H)

Since *G* and *H* are independent, then, knowing that a person is taking a science class does not change the chance that he/she is taking math. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he/she is taking math. For practice, show that P(H|G) = P(H) to show that *G* and *H* are independent events.

Example 3.4

In a box there are 3 red cards and 5 blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let *R* = red card is drawn, *B* = blue card is drawn, *E* = even-numbered card is drawn.

The sample space *S* = R1, R2, R3, B1, B2, B3, B4, B5. *S* has 8 outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. P(R AND B) = 0. (You cannot draw one card that is both red and blue.)
- $P(E) = \frac{3}{8}$. (There are 3 even-numbered cards, R2, B2, and B4.)

- P(E|B) = $\frac{2}{5}$. (There are 5 blue cards: B1, B2, B3, B4, and B5. Out of the blue cards, there are 2 even cards: B2 and B4.)
- P(B|E) = $\frac{2}{3}$. (There are 3 even-numbered cards: R2, B2, and B4. Out of the even-numbered cards, 2 are blue: B2 and B4.)
- The events *R* and *B* are mutually exclusive because P(R AND B) = 0.
- Let *G* = card with a number greater than 3. *G* = {B4, B5}. P(G) = $\frac{2}{8}$. Let *H* = blue card numbered between 1 and 4, inclusive. *H* = {B1, B2, B3, B4}. P(G|H) = $\frac{1}{4}$. (The only card in H that has a number greater than 3 is B4.) Since $\frac{2}{8}$ = $\frac{1}{4}$, P(G) = P(G|H) which means that *G* and *H* are independent.

Example 3.5

In a particular college class, 60% of the students are female. 50 % of all students in the class have long hair. 45% of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that the student is female. Let L be the event that the student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- P(F) = 0.60 ; P(L) = 0.50
- P(F AND L) = 0.45
- P(L|F) = 0.75

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know P(F|L) yet, so you can not use the second condition.

Solution 1

Check whether P(F and L) = P(F)P(L): We are given that P(F and L) = 0.45 ; but P(F)P(L) = (0.60)(0.50)= 0.30 The events of being female and having long hair are not independent because P(F and L) does not equal P(F)P(L).

Solution 2

check whether P(L|F) equals P(L): We are given that P(L|F) = 0.75 but P(L) = 0.50; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

**Example 5 contributed by Roberta Bloom

3.4 Two Basic Rules of Probability

The Multiplication Rule

If *A* and *B* are two events defined on a **sample space**, then: P(A AND B) = P(B) · P(A|B).

This rule may also be written as : P(A|B)= $\frac{\text{P(A AND B)}}{\text{P(B)}}$

(The probability of *A* given *B* equals the probability of *A* and *B* divided by the probability of *B*.)

If A and B are **independent**, then P(A|B) = P(A). Then P(A AND B) = P(A|B) P(B) becomes P(A AND B) = P(A) P(B).

The Addition Rule

If *A* and *B* are defined on a sample space, then: P(A OR B) = P(A) + P(B) − P(A AND B).

If *A* and *B* are **mutually exclusive**, then P(A AND B) = 0. Then P(A OR B) = P(A) + P(B) − P(A AND B) becomes P(A OR B) = P(A) + P(B).

Example 3.6

Klaus is trying to choose where to go on vacation. His two choices are: *A* = New Zealand and *B* = Alaska

- Klaus can only afford one vacation. The probability that he chooses *A* is P(A) = 0.6 and the probability that he chooses *B* is P(B) = 0.35.
- P(A and B) = 0 because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is P(A OR B) = P(A) + P(B) = 0.6 + 0.35 = 0.95. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example 3.7

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game.

A = the event Carlos is successful on his first attempt. P(A) = 0.65. *B* = the event Carlos is successful on his second attempt. P(B) = 0.65. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

What is the probability that he makes both goals?

The problem is asking you to find P(A AND B) = P(B AND A). Since P(B|A) = 0.90:

$$\text{P(B AND A)} = \text{P(B|A) P(A)} = 0.90 * 0.65 = 0.585$$

Carlos makes the first and second goals with probability 0.585.

What is the probability that Carlos makes either the first goal or the second goal?

The problem is asking you to find P(A OR B).

$$\text{P(A OR B)} = \text{P(A)} + \text{P(B)} - \text{P(A AND B)} = 0.65 + 0.65 - 0.585 = 0.715$$

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

Are *A* and *B* independent?

No, they are not, because P(B AND A) = 0.585.

$$\text{P(B)} \cdot \text{P(A)} = (0.65) \cdot (0.65) = 0.423$$
$$0.423 \neq 0.585 = \text{P(B AND A)}$$

So, P(B AND A) is **not** equal to P(B) · P(A).

Are *A* and *B* mutually exclusive?

No, they are not because P(A and B) = 0.585.

To be mutually exclusive, P(A AND B) must equal 0.

Example 3.8

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice 4 times a week. **Thirty** of the intermediate swimmers practice 4 times a week. **Ten** of the novice swimmers practice 4 times a week. Suppose one member of the swim team is randomly chosen. Answer the questions (Verify the answers):

What is the probability that the member is a novice swimmer?

$\frac{28}{150}$

What is the probability that the member practices 4 times a week?

$\frac{80}{150}$

What is the probability that the member is an advanced swimmer and practices 4 times a week?

$\frac{40}{150}$

What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

P(advanced AND intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Are being a novice swimmer and practicing 4 times a week independent events? Why or why not?

No, these are not independent events.

$$\begin{gathered}\text{P(novice AND practices 4 times per week)} = 0.0667 \\ \text{P(novice)} \cdot \text{P(practices 4 times per week)} = 0.0996 \\ 0.0667 \neq 0.0996\end{gathered}$$

Example 3.9

Studies show that, if she lives to be 90, about 1 woman in 7 (approximately 14.3%) will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is believed to be negative about 85% of the time. Let *B* = woman develops breast cancer and let *N* = tests negative. Suppose one woman is selected at random.

What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

P(B) = 0.143 ; P(N) = 0.85

Given that the woman has breast cancer, what is the probability that she tests negative?

P(N|B) = 0.02

What is the probability that the woman has breast cancer AND tests negative?

P(B AND N) = P(B) · P(N|B) = (0.143) · (0.02) = 0.0029

What is the probability that the woman has breast cancer or tests negative?

P(B OR N) = P(B) + P(N) − P(B AND N) = 0.143 + 0.85 − 0.0029 = 0.9901

Are having breast cancer and testing negative independent events?

No. P(N) = 0.85; P(N|B) = 0.02. So, P(N|B) does not equal P(N)

Are having breast cancer and testing negative mutually exclusive?

No. P(B AND N) = 0.0020. For *B* and *N* to be mutually exclusive, P(B AND N) must be 0.

3.5 Contingency Tables

A **contingency table** provides a different way of calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 3.10

Suppose a study of speeding violations and drivers who use car phones produced the following fictional data:

Table 3.1

	Speeding violation in the last year	No speeding violation in the last year	Total
Car phone user	25	280	305
Not a car phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table

P(person is a car phone user) =

$$\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$$

P(person had no violation in the last year) =

$$\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$$

P(person had no violation in the last year AND was a car phone user) =

$$\frac{280}{755}$$

P(person is a car phone user OR person had no violation in the last year) =

$$\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$$

P(person is a car phone user GIVEN person had a violation in the last year) =

$\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

P(person had no violation last year GIVEN person was not a car phone user) =

$\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Example 3.11

The following table shows a random sample of 100 hikers and the areas of hiking preferred:

Table 3.2 Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Complete the table.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Are the events "being female" and "preferring the coastline" independent events?

Let *F* = being female and let *C* = preferring the coastline.

a. P(F AND C) =
b. P(F) · P(C) =

Are these two numbers the same? If they are, then *F* and *C* are independent. If they are not, then *F* and *C* are not independent.

a. P(F AND C) = $\frac{18}{100}$ = 0.18

b. P(F) · P(C) = $\frac{45}{100} \cdot \frac{34}{100}$ = 0.45 · 0.34 = 0.153

P(F AND C) ≠ P(F) · P(C), so the events *F* and *C* are not independent.

Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male and let L = prefers hiking near lakes and streams.

a. What word tells you this is a conditional?
b. Fill in the blanks and calculate the probability: P(___|___) = ___.
c. Is the sample space for this problem all 100 hikers? If not, what is it?

a. The word 'given' tells you that this is a conditional.

b. P(M|L) = $\frac{25}{41}$

c. No, the sample space for this problem is 41.

Find the probability that a person is female or prefers hiking on mountain peaks. Let *F* = being female and let *P* = prefers mountain peaks.

a. P(F) =
b. P(P) =
c. P(F AND P) =
d. Therefore, P(F OR P) =

a. P(F) = $\frac{45}{100}$

b. P(P) = $\frac{25}{100}$

c. P(F AND P) = $\frac{11}{100}$

d. P(F OR P) = $\frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Example 3.12

Muddy Mouse lives in a cage with 3 doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Table 3.3 Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Caught).
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Not Caught).

Verify the remaining entries.

Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\mathbf{\frac{19}{60}}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\mathbf{\frac{41}{60}}$
Total	$\mathbf{\frac{5}{15}}$	$\mathbf{\frac{4}{12}}$	$\mathbf{\frac{2}{6}}$	1

What is the probability that Alissa does not catch Muddy?

$\frac{41}{60}$

What is the probability that Muddy chooses Door One **OR** Door Two given that Muddy is caught by Alissa?

$\frac{9}{19}$

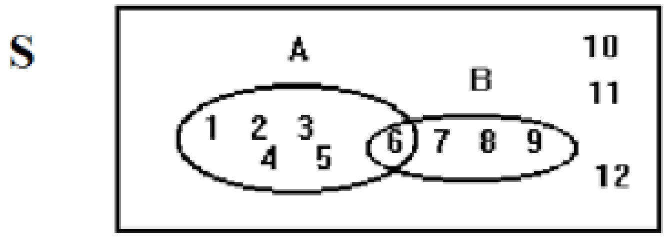
You could also do this problem by using a probability tree. See the **Tree Diagrams (Optional)** section of this chapter for examples.

3.6 Venn Diagrams (optional)

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example 3.13

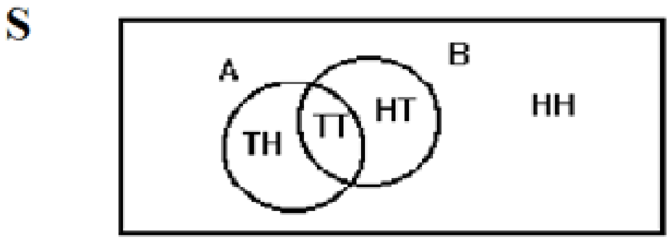
Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event *A* = {1, 2, 3, 4, 5, 6} and event *B* = {6, 7, 8, 9}. Then A AND B = {6} and A OR B = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:



Example 3.14

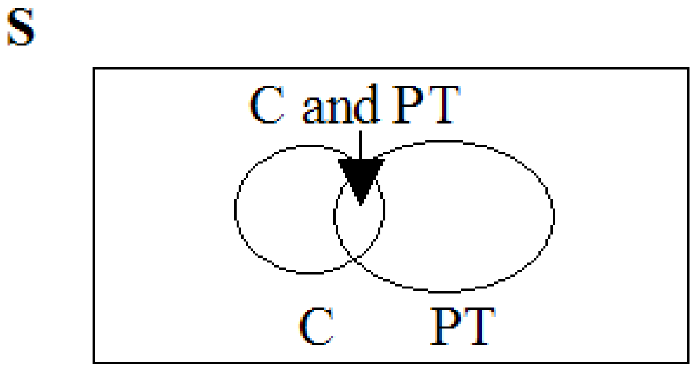
Flip 2 fair coins. Let *A* = tails on the first coin. Let *B* = tails on the second coin. Then *A* = {TT, TH} and *B* = {TT, HT}. Therefore, A AND B = {TT}. A OR B = {TH, TT, HT}.

The sample space when you flip two fair coins is *S* = {HH, HT, TH, TT}. The outcome HH is in neither *A* nor *B*. The Venn diagram is as follows:



Example 3.15

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let *C* = student belongs to a club and PT = student works part time.



If a student is selected at random find

- The probability that the student belongs to a club. P(C) = 0.40.
- The probability that the student works part time. P(PT) = 0.50.

- The probability that the student belongs to a club AND works part time. P(C AND PT) = 0.05.
- The probability that the student belongs to a club **given** that the student works part time.

$$P(C|PT) = \frac{P(C\text{ AND }PT)}{P(PT)} = \frac{0.05}{0.50} = 0.1 \quad (3.1)$$

- The probability that the student belongs to a club **OR** works part time.

$$P(C\text{ OR }PT) = P(C) + P(PT) - P(C\text{ AND }PT) = 0.40 + 0.50 - 0.05 = 0.85 \quad (3.2)$$

3.7 Tree Diagrams (optional)

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 3.16

In an urn, there are 11 balls. Three balls are red (*R*) and 8 balls are blue (*B*). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.

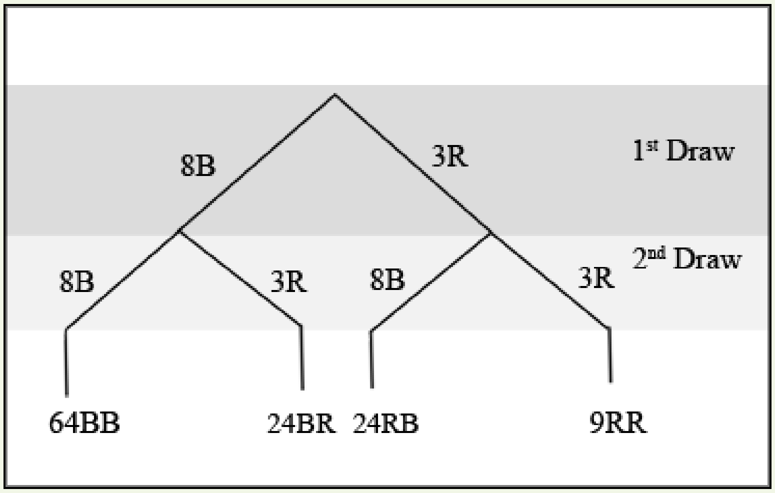


Figure 3.1 Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the 9 RR outcomes can be written as:

R1R1; R1R2; R1R3; R2R1; R2R2; R2R3; R3R1; R3R2; R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, and with replacement. There are 11 · 11 = 121 outcomes, the size of the **sample space**.

List the 24 BR outcomes: B1R1, B1R2, B1R3, ...

B1R1; B1R2; B1R3; B2R1; B2R2; B2R3; B3R1; B3R2; B3R3; B4R1; B4R2; B4R3; B5R1; B5R2; B5R3; B6R1; B6R2; B6R3; B7R1; B7R2; B7R3; B8R1; B8R2; B8R3

Using the tree diagram, calculate P(RR).

$$P(RR) = \frac{3}{11} \cdot \frac{3}{11} = \frac{9}{121}$$

Using the tree diagram, calculate P(RB OR BR).

$$P(RB\text{ OR }BR) = \frac{3}{11} \cdot \frac{8}{11} + \frac{8}{11} \cdot \frac{3}{11} = \frac{48}{121}$$

Using the tree diagram, calculate P(R on 1st draw AND B on 2nd draw).

P(R on 1st draw AND B on 2nd draw) = P(RB) = $\frac{3}{11} \cdot \frac{8}{11} = \frac{24}{121}$

Using the tree diagram, calculate P(R on 2nd draw given B on 1st draw).

P(R on 2nd draw given B on 1st draw) = P(R on 2nd | B on 1st) = $\frac{24}{88} = \frac{3}{11}$

This problem is a conditional. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

Using the tree diagram, calculate P(BB).

P(BB) = $\frac{64}{121}$

Using the tree diagram, calculate P(B on the 2nd draw given R on the first draw).

P(B on 2nd draw | R on 1st draw) = $\frac{8}{11}$

There are 9 + 24 outcomes that have *R* on the first draw (9 RR and 24 RB). The sample space is then 9 + 24 = 33. Twenty-four of the 33 outcomes have *B* on the second draw. The probability is then $\frac{24}{33}$.

Example 3.17

An urn has 3 red marbles and 8 blue marbles in it. Draw two marbles, one at a time, this time without replacement from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second ball. Below is a tree diagram. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\frac{3}{11} \cdot \frac{2}{10} = \frac{6}{110}$.

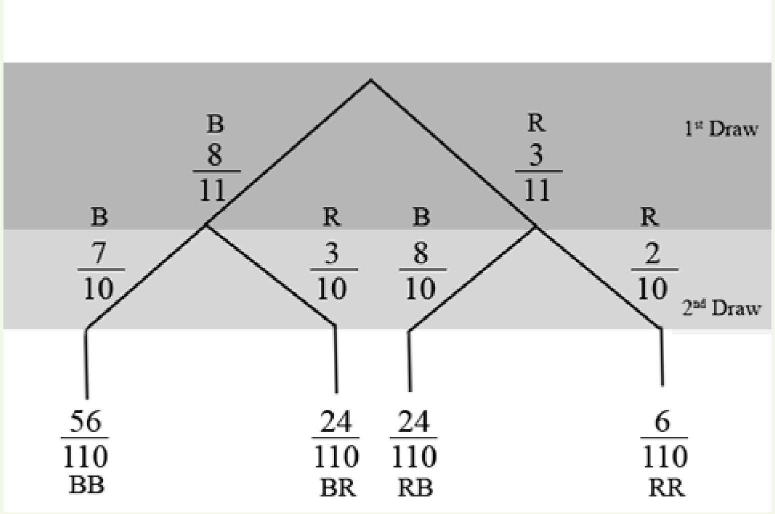


Figure 3.2 Total = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$

If you draw a red on the first draw from the 3 red possibilities, there are 2 red left to draw on the second draw. You do not put back or replace the first ball after you have drawn it. You draw **without replacement**, so that on the second draw there are 10 marbles left in the urn.

Calculate the following probabilities using the tree diagram.

P(RR) =

P(RR) = $\frac{3}{11} \cdot \frac{2}{10} = \frac{6}{110}$

Fill in the blanks:

P(RB OR BR) = $\frac{3}{11} \cdot \frac{8}{10} + (\_\_\_)(\_\_\_) = \frac{48}{110}$

P(RB or BR) = $\frac{3}{11} \cdot \frac{8}{10} + \left(\mathbf{\frac{8}{11}}\right)\left(\mathbf{\frac{3}{10}}\right) = \frac{48}{110}$

P(R on 2d | B on 1st) =

P(R on 2d | B on 1st) = $\frac{3}{10}$

Fill in the blanks:

P(R on 1st and B on 2nd) = P(RB) = $(\_\_\_)(\_\_\_) = \frac{24}{110}$

P(R on 1st and B on 2nd) = P(RB) = $\left(\mathbf{\frac{3}{11}}\right)\left(\mathbf{\frac{8}{10}}\right) = \frac{24}{110}$

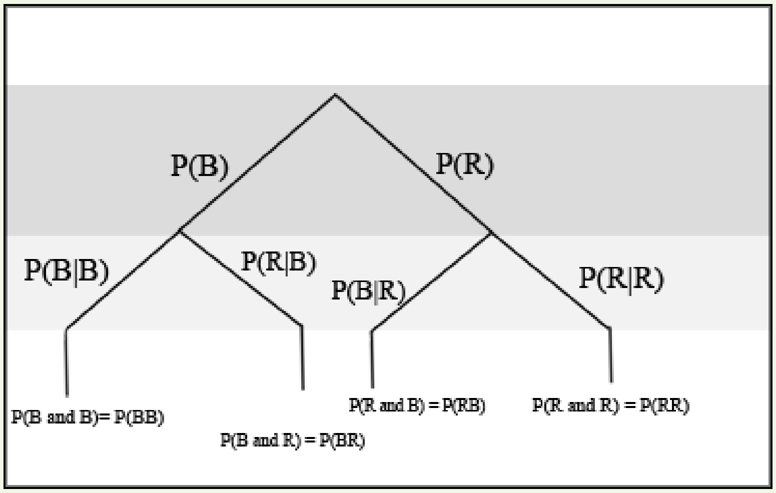
P(BB) =

P(BB) = $\frac{8}{11} \cdot \frac{7}{10}$

P(B on 2nd | R on 1st) =

There are 6 + 24 outcomes that have *R* on the first draw (6 RR and 24 RB). The 6 and the 24 are frequencies. They are also the numerators of the fractions $\frac{6}{110}$ and $\frac{24}{110}$. The sample space is no longer 110 but 6 + 24 = 30. Twenty-four of the 30 outcomes have *B* on the second draw. The probability is then $\frac{24}{30}$. Did you get this answer?

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



- P(R|R) here means P(R on 2nd | R on 1st)
- P(B|R) here means P(B on 2nd | R on 1st)
- P(R|B) here means P(R on 2nd | B on 1st)
- P(B|B) here means P(B on 2nd | B on 1st)

3.8 Summary of Formulas

Formula

If *A* and A' are complements then P(A) + P(A') = 1

Formula

P(A OR B) = P(A) + P(B) − P(A AND B)

Formula

If *A* and *B* are mutually exclusive then P(A AND B) = 0 ; so P(A OR B) = P(A) + P(B).

Formula

- P(A AND B) = P(B)P(A|B)
- P(A AND B) = P(A)P(B|A)

Formula

If *A* and *B* are independent then:

- P(A|B) = P(A)
- P(B|A) = P(B)
- P(A AND B) = P(A)P(B)

3.9 Practice 1: Contingency Tables

Student Learning Outcomes

- The student will construct and interpret contingency tables.

Given

An article in the *New England Journal of Medicine* , reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most 10 cigarettes per day, there were 9886 African Americans, 2745 Native Hawaiians, 12,831 Latinos, 8378 Japanese Americans, and 7650 Whites. Of the people smoking 11-20 cigarettes per day, there were 6514 African Americans, 3062 Native Hawaiians, 4932 Latinos, 10,680 Japanese Americans, and 9877 Whites. Of the people smoking 21-30 cigarettes per day, there were 1671 African Americans, 1419 Native Hawaiians, 1406 Latinos, 4715 Japanese Americans, and 6062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2305 Japanese Americans, and 3970 Whites. (*(Source: http://www.nejm.org/doi/full/10.1056/NEJMoa033250)*)

Complete the Table

Complete the table below using the data provided.

Table 3.4 Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Analyze the Data

Suppose that one person from the study is randomly selected.

Discussion Questions

3.10 Practice 2: Calculating Probabilities

Student Learning Outcomes

- Students will define basic probability terms.
- Students will calculate probabilities.
- Students will determine whether two events are mutually exclusive or whether two events are independent.

Use probability rules to solve the problems below. Show your work.

Given

48% of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. (*Source: http://field.com/fieldpollonline/subscribers/Rls2393.pdf*).
37.6% of all Californians are Latino (*Source: U.S. Census Bureau*).

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder
- L = Latino Californians

Suppose that one Californian is randomly selected.

Analyze the Data

3.11 Homework

The next two questions refer to the following: The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20 - 64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20 - 64; 13.53% are age 65 or over. (Source: Federal Highway Administration, U.S. Dept. of Transportation)

Try these multiple choice questions.

The next two questions refer to the following probability tree diagram which shows tossing an unfair coin **FOLLOWED BY** drawing one bead from a cup containing 3 red (R), 4 yellow (Y) and 5 blue (B) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where H = "heads" and T = "tails".

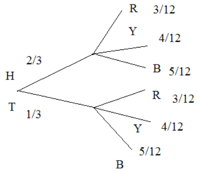


Figure 3.3

The next three questions refer to the following table of data obtained from *www.baseball-almanac.com (http://cnx.org/content/m16836/1.20/www.baseball-almanac.com)* showing hit information for 4 well known baseball players. Suppose that one hit from the table is randomly selected.

Table 3.5

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1517	506	136	714	2873
Jackie Robinson	1054	273	54	137	1518
Ty Cobb	3603	174	295	114	4189
Hank Aaron	2294	624	98	755	3771
TOTAL	8471	1577	583	1720	12351

**Exercises 33 - 40 contributed by Roberta Bloom

3.12 Review

The first six exercises refer to the following study: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks. Answer the following:

The next two questions refer to the following study: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Table 3.6

Weight Gain	Frequency
-2	3
-1	5
0	2
1	4
4	13
6	2
11	1

3.13 Lab: Probability Topics

Class time:

Names:

Student Learning Outcomes:

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment:

Count out 40 mixed-color M&M’s® which is approximately 1 small bag’s worth (distance learning classes using the virtual lab would want to count out 25 M&M’s®). Record the number of each color in the "Population" table. Use the information from this table to complete the theoretical probability questions. Next, put the M&M’s in a cup. The experiment is to pick 2 M&M’s, one at a time. Do **not** look at them as you pick them. The first time through, replace the first M&M before picking the second one. Record the results in the “With Replacement” column of the empirical table. Do this 24 times. The second time through, after picking the first M&M, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the “Without Replacement” column section of the "Empirical Results" table. After you record the pick, put **both** M&M’s back. Do this a total of 24 times, also. Use the data from the "Empirical Results" table to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Table 3.7 Population

Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Table 3.8 Theoretical Probabilities Note: G_2 = green on second pick; R_1 = red on first pick; B_1 = brown on first pick; B_2 = brown on second pick; doubles = both picks are the same colour.

	With Replacement	Without Replacement
P(2 reds)		
$P\left(R_1 B_2 \text{OR} B_1 R_2\right)$		
$P\left(R_1 \text{AND} G_2\right)$		
$P\left(G_2 \mid R_1\right)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Table 3.9 Empirical Results

With Replacement	Without Replacement
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)

Table 3.10 Empirical Probabilities Note:

	With Replacement	Without Replacement
P(2 reds)		
$P\left(R_1B_2 \text{ OR } B_1R_2\right)$		
$P\left(R_1 \text{ AND } G_2\right)$		
$P\left(G_2 \mid R_1\right)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Discussion Questions

1. Why are the “With Replacement” and “Without Replacement” probabilities different?
2. Convert P(no yellows) to decimal format for both Theoretical “With Replacement” and for Empirical “With Replacement”. Round to 4 decimal places.
   a. Theoretical “With Replacement”: P(no yellows) =
   b. Empirical “With Replacement”: P(no yellows) =
   c. Are the decimal values “close”? Did you expect them to be closer together or farther apart? Why?
3. If you increased the number of times you picked 2 M&M’s to 240 times, why would empirical probability values change?
4. Would this change (see (3) above) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
5. Explain the differences in what $P\left(G_1 \text{AND } R_2\right)$ and $P\left(R_1 \mid G_2\right)$ represent. Hint: Think about the sample space for each probability.

Glossary

Conditional Probability: The likelihood that an event will occur given that another event has already occurred.

Contingency Table: The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

Equally Likely: Each outcome of an experiment has the same probability.

Event: A subset in the set of all outcomes of an experiment. The set of all outcomes of an experiment is called a **sample space** and denoted usually by S. An event is any arbitrary subset in **S**. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, etc. Standard notations for events are capital letters such as A, B, C, etc.

Experiment: A planned activity carried out under controlled conditions.

Independent Events: The occurrence of one event has no effect on the probability of the occurrence of any other event. Events A and B are independent if one of the following is true: (1). $P(A|B) = P(A)$; (2) $P(B|A) = P(B)$; (3) $P(A\text{and}B) = P(A)P(B)$.

Independent Events: The occurrence of one event has no effect on the probability of the occurrence of any other event. Events A and B are independent if one of the following is true: (1). $P(A|B) = P(A)$; (2) $P(B|A) = P(B)$; (3) $P(A\text{and}B) = P(A)P(B)$.

Mutually Exclusive: An observation cannot fall into more than one class (category). Being in more than one category prevents being in a mutually exclusive category.

Mutually Exclusive: An observation cannot fall into more than one class (category). Being in more than one category prevents being in a mutually exclusive category.

Outcome (observation): A particular result of an experiment.

Probability: A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let *S* denote the sample space and *A* and *B* are two events in *S* . Then:

- $0 \le P(A) \le 1;$.
- If *A* and *B* are any two mutually exclusive events, then $P(A\text{or}B) = P(A) + P(B)$.
- $P(S) = 1$.

Sample Space: The set of all possible outcomes of an experiment.

Sample Space: The set of all possible outcomes of an experiment.

Sample Space: The set of all possible outcomes of an experiment.

Tree Diagram: The useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes simultaneously with associated probabilities (frequencies, relative frequencies).

Venn Diagram: The visual representation of a sample space and events in the form of circles or ovals showing their intersections.

4 DISCRETE RANDOM VARIABLES

4.1 Discrete Random Variables

Student Learning Outcomes

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

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately (optional).
- Recognize the geometric probability distribution and apply it appropriately (optional).
- Recognize the hypergeometric probability distribution and apply it appropriately (optional).
- Classify discrete word problems by their distributions.

Introduction

A student takes a 10 question true-false quiz. Because the student had such a busy schedule, he or she could not study and randomly guesses at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

In this chapter, you will study probability problems involving discrete random distributions. You will also study long-term averages associated with them.

Random Variable Notation

Upper case letters like *X* or *Y* denote a random variable. Lower case letters like *x* or *y* denote the value of a random variable. If ***X* is a random variable, then *X* is written in words.** and ***x* is given as a number.**

For example, let *X* = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is TTT; THH; HTH; HHT; HTT; THT; TTH; HHH. Then, *x* = 0, 1, 2, 3. *X* is in words and *x* is a number. Notice that for this example, the *x* values are countable outcomes. Because you can count the possible values that *X* can take on and the outcomes are random (the *x* values 0, 1, 2, 3), *X* is a discrete random variable.

Optional Collaborative Classroom Activity

Toss a coin 10 times and record the number of heads. After all members of the class have completed the experiment (tossed a coin 10 times and counted the number of heads), fill in the chart using a heading like the one below. Let *X* = the number of heads in 10 tosses of the coin.

Table 4.1

x	Frequency of *x*	Relative Frequency of *x*

- Which value(s) of *x* occurred most frequently?
- If you tossed the coin 1,000 times, what values could *x* take on? Which value(s) of *x* do you think would occur most frequently?
- What does the relative frequency column sum to?

4.2 Probability Distribution Function (PDF) for a Discrete Random Variable

A discrete **probability distribution function** has two characteristics:

- Each probability is between 0 and 1, inclusive.
- The sum of the probabilities is 1.

Example 4.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let *X* = the number of times a newborn wakes its mother after midnight. For this example, *x* = 0, 1, 2, 3, 4, 5.

P(x) = probability that *X* takes on a value *x*.

Table 4.2

x	P(x)
0	$P(x=0) = \frac{2}{50}$
1	$P(x=1) = \frac{11}{50}$
2	$P(x=2) = \frac{23}{50}$
3	$P(x=3) = \frac{9}{50}$
4	$P(x=4) = \frac{4}{50}$
5	$P(x=5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because

1. Each P(x) is between 0 and 1, inclusive.
2. The sum of the probabilities is 1, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \tag{4.1}$$

Example 4.2

Suppose Nancy has classes **3 days** a week. She attends classes 3 days a week **80%** of the time, **2 days 15%** of the time, **1 day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Let *X* = the number of days Nancy ____________________ .

Let *X* = the number of days Nancy **attends class per week**.

X takes on what values?

0, 1, 2, and 3

Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in the previous example. The table should have two columns labeled *x* and P(x). What does the P(x) column sum to?

x	P(x)
0	0.01
1	0.04
2	0.15
3	0.80

4.3 Mean or Expected Value and Standard Deviation

The **expected value** is often referred to as the **"long-term"average or mean** . This means that over the long term of doing an experiment over and over, you would **expect** this average.

The **mean** of a random variable *X* is μ. If we do an experiment many times (for instance, flip a fair coin, as Karl Pearson did, 24,000 times and let *X* = the number of heads) and record the value of *X* each time, the average is likely to get closer and closer to μ as we keep repeating the experiment. This is known as the **Law of Large Numbers**.

To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

A Step-by-Step Example

A men's soccer team plays soccer 0, 1, or 2 days a week. The probability that they play 0 days is 0.2, the probability that they play 1 day is 0.5, and the probability that they play 2 days is 0.3. Find the long-term average, μ, or expected value of the days per week the men's soccer team plays soccer.

To do the problem, first let the random variable *X* = the number of days the men's soccer team plays soccer per week. *X* takes on the values 0, 1, 2. Construct a PDF table, adding a column xP(x). In this column, you will multiply each *x* value by its probability.

Table 4.3 Expected Value Table This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	P(x)	*x*P(x)
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Add the last column to find the long term average or expected value: (0)(0.2)+(1)(0.5)+(2)(0.3)= 0 + 0.5 + 0.6 = 1.1.

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long term average or expected value if the men's soccer team plays soccer week after week after week. We say μ=1.1

Example 4.3

Find the expected value for the example about the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times a newborn wakes its mother after midnight.

Table 4.4 You expect a newborn to wake its mother after midnight 2.1 times, on the average.

x	P(X)	*x*P(X)
0	$P(x=0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$
1	$P(x=1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$
2	$P(x=2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$
3	$P(x=3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$
4	$P(x=4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$
5	$P(x=5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$

Add the last column to find the expected value. $\mu = \text{Expected Value} = \frac{105}{50} = 2.1$

Go back and calculate the expected value for the number of days Nancy attends classes a week. Construct the third column to do so.

2.74 days a week.

Example 4.4

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from 0 to 9 with replacement. You pay $2 to play and could profit $100,000 if you match all 5 numbers in order (you get your $2 back plus $100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let *X* = the amount of money you profit. The values of *x* are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of *x* are 100,000 dollars and -2 dollars.

To win, you must get all 5 numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are 10 numbers. You may choose a number more than once. The probability of choosing all 5 numbers correctly and in order is:

$$\frac{1}{10}*\frac{1}{10}*\frac{1}{10}*\frac{1}{10}*\frac{1}{10}* = 1*10^{-5} = 0.00001 \tag{4.2}$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

$$1 - 0.00001 = 0.99999 \tag{4.3}$$

The expected value table is as follows.

Table 4.5 Add the last column. -1.99998 + 1 = -0.99998

	x	P(x)	*x*P(x)
Loss	-2	0.99999	(-2)(0.99999)=-1.99998
Profit	100,000	0.00001	(100000)(0.00001)=1

Since -0.99998 is about -1, you would, on the average, expect to lose approximately one dollar for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Example 4.5

Suppose you play a game with a biased coin. You play each game by tossing the coin once. P(heads) = $\frac{2}{3}$ and P(tails) = $\frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

Define a random variable *X*.

X = amount of profit

Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

	x	**P(x)**	***x*P(x)**
WIN	10	$\frac{1}{3}$	$\mathbf{\frac{10}{3}}$
LOSE	**-6**	$\mathbf{\frac{2}{3}}$	$\frac{-12}{3}$

What is the expected value, μ? Do you come out ahead?

Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root . To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x - \mu)^2 \cdot P\left(x\right)$ and take the square root.

Table 4.6

x	P(x)	xP(x)	$(x - \mu)^2 P(x)$
0	0.2	(0)(0.2) = 0	$(0 - 1.1)^2\left(.2\right) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1 - 1.1)^2\left(.5\right) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2 - 1.1)^2\left(.3\right) = 0.243$

Add the last column in the table. $0.242 + 0.005 + 0.243 = 0.490$. The standard deviation is the square root of 0.49. $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas that calculate μ and σ.

4.4 Common Discrete Probability Distribution Functions

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

4.5 Binomial

The characteristics of a binomial experiment are:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter *n* denotes the number of trials.
2. There are only 2 possible outcomes, called "success" and, "failure" for each trial. The letter *p* denotes the probability of a success on one trial and *q* denotes the probability of a failure on one trial. $p + q = 1$.
3. The *n* trials are independent and are repeated using identical conditions. Because the *n* trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, *p*, of a success and probability, *q*, of a failure remain the same. For example, randomly guessing at a true - false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true - false question with probability $p = 0.6$. Then, $q = 0.4$ .This means that for every true - false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the *n* independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution is $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then $\sigma = \sqrt{npq}$.

Any experiment that has characteristics 2 and 3 and where n = 1 is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example 4.6

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable is $X =$ the number of students who withdraw from the randomly selected elementary physics class.

Example 4.7

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55% and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, what is the probability that you win 15 of the 20 games? Here, if you define X = the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Example 4.8

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than 10 heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. The probability question can be stated mathematically as $P(x > 10)$.

Example 4.9

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

This is a binomial problem because there is only a success or a __________, there are a definite number of trials, and the probability of a success is 0.70 for each trial.

failure

If we are interested in the number of students who do their homework, then how do we define X?

X = the number of statistics students who do their homework on time

What values does x take on?

0, 1, 2, …, 50

What is a "failure", in words?

Failure is a student who does not do his or her homework on time.

The probability of a success is $p = 0.70$. The number of trial is $n = 50$.

If $p + q = 1$, then what is q?

$q = 0.30$

The words "at least" translate as what kind of inequality for the probability question $P(x$____$40)$.

greater than or equal to (≥)

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

Notation for the Binomial: B = Binomial Probability Distribution Function

$X \sim B(n, p)$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p. n = number of trials p = probability of a success on each trial

Example 4.10

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$ and $p = 0.41$. $q = 1 - 0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (calculator or computer)

Using the TI-83+ or the TI-84 calculators, the calculations are as follows. Go into 2nd DISTR. The syntax for the instructions are

To calculate (x = value): binompdf(n, p, number) If "number" is left out, the result is the binomial probability table.

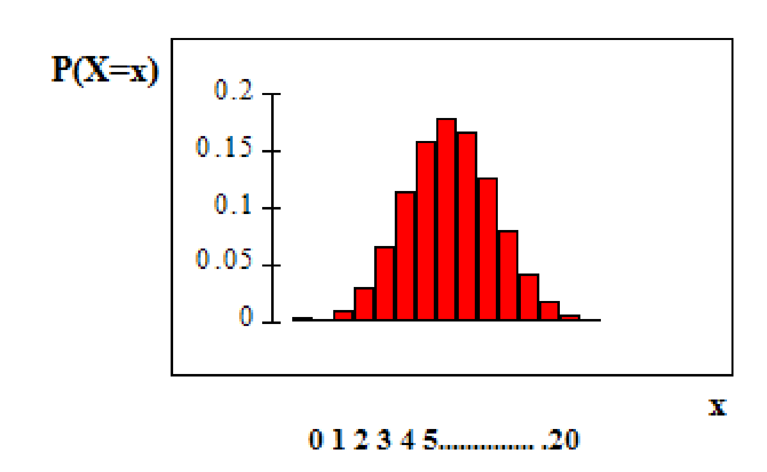
To calculate $P(x \leq$ value): binomcdf(n, p, number) If "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to A:binomcdf. Press ENTER. Enter 20,.41,12). The result is $P(x \leq 12) = 0.9738$.

If you want to find $P(x = 12)$, use the pdf (0:binompdf). If you want to find illegal children in mi, use 1 - binomcdf(20,.41,12).

The probability at most 12 workers have a high school diploma but do not pursue any further education is 0.9738

The graph of $x \sim B(20, 0.41)$ is:



The y-axis contains the probability of x, where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$. $\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20$.

Example 4.11

The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of 10 staff members and 6 students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? All names of the committee are put into a box and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw produces a student. The probability is $\frac{6}{15}$ when the first draw produces a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

4.6 Geometric (optional)

The characteristics of a geometric experiment are:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bull's eye until you hit the bull's eye. The first time you hit the bull's eye is a "success" so you stop throwing the dart. It might take you 6 tries until you hit the bull's eye. You can think of the trials as failure, failure, failure, failure, failure, success. STOP.
2. In theory, the number of trials could go on forever. There must be at least one trial.

3. The probability,*p*, of a success and the probability, *q*, of a failure is the same for each trial. $p + q = 1$ and $q = 1 - p$. For example, the probability of rolling a 3 when you throw one fair die is $\frac{1}{6}$. This is true no matter how many times you roll the die. Suppose you want to know the probability of getting the first 3 on the fifth roll. On rolls 1, 2, 3, and 4, you do not get a face with a 3. The probability for each of rolls 1, 2, 3, and 4 is $q = \frac{5}{6}$, the probability of a failure. The probability of getting a 3 on the fifth roll is $\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = 0.0804$

X = the number of independent trials until the first success. The mean and variance are in the summary in this chapter.

Example 4.12

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

Example 4.13

A safety engineer feels that 35% of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) **until** she finds one that shows an accident caused by failure of employees to follow instructions. On the average, how many reports would the safety engineer **expect** to look at until she finds a report showing an accident caused by employee failure to follow instructions? What is the probability that the safety engineer will have to examine at least 3 reports until she finds a report showing an accident caused by employee failure to follow instructions?

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

Example 4.14

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

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

Let *X* = the number of ____________ you must ask ____________ one says yes.

Let *X* = the number of **students** you must ask **until** one says yes.

What values does *X* take on?

1, 2, 3, …, (total number of students)

What are *p* and *q*?

- *p* = 0.55
- *q* = 0.45

The probability question is P(_______).

P(*x* = 4)

Notation for the Geometric: G = Geometric Probability Distribution Function

X ~ G(p)

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

Example 4.15

Assume that the probability of a defective computer component is 0.02. Components are randomly selected. Find the probability that the first defect is caused by the 7th component tested. How many components do you expect to test until one is found to be defective?

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

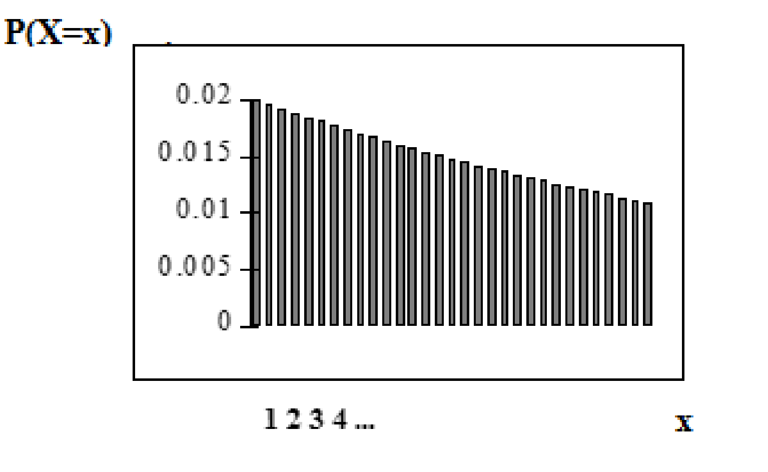
X takes on the values 1, 2, 3, ... where p = 0.02. X ~ G(0.02)

Find $P(x = 7)$. $P(x = 7) = 0.0177$. (calculator or computer)

TI-83+ and TI-84: **For a general discussion, see this example (http://cnx.org/content/m16820/1.16/##element-678) (binomial)**. The syntax is similar. The geometric parameter list is (p, number) If "number" is left out, the result is the geometric probability table. For this problem: **After you are in 2nd DISTR, arrow down to D:geometpdf. Press ENTER. Enter .02,7). The result is *P*(*x* = 7) = 0.0177.**

The probability that the 7th component is the first defect is 0.0177.

The graph of X ~ G(0.02) is:



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

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

The formula for the mean is $\mu = \frac{1}{p} = \frac{1}{0.02} = 50$

The formula for the variance is $\sigma^2 = \frac{1}{p} \cdot \left(\frac{1}{p} - 1\right) = \frac{1}{0.02} \cdot \left(\frac{1}{0.02} - 1\right) = 2450$

The standard deviation is $\sigma = \sqrt{\frac{1}{p} \cdot \left(\frac{1}{p} - 1\right)} = \sqrt{\frac{1}{0.02} \cdot \left(\frac{1}{0.02} - 1\right)} = 49.5$

4.7 Hypergeometric (optional)

The characteristics of a hypergeometric experiment are:

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

The outcomes of a hypergeometric experiment fit a **hypergeometric probability** distribution. The random variable X = the number of items from the group of interest. The mean and variance are given in the summary.

Example 4.16

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

Example 4.17

Suppose a shipment of 100 VCRs is known to have 10 defective VCRs. An inspector randomly chooses 12 for inspection. He is interested in determining the probability that, among the 12, at most 2 are defective. The two groups are the 90 non-defective VCRs and the 10 defective VCRs. The group of interest (first group) is the defective group because the probability question asks for the probability of at most 2 defective VCRs. The size of the sample is 12 VCRs. (They may be non-defective or defective.) Let X = the number of defective VCRs in the sample of 12. X takes on the values 0, 1, 2, ..., 10. X may not take on the values 11 or 12. The sample size is 12, but there are only 10 defective VCRs. The inspector wants to know $P(x \leq 2)$ ("At most" means "less than or equal to").

Example 4.18

You are president of an on-campus special events organization. You need a committee of 7 to plan a special birthday party for the president of the college. Your organization consists of 18 women and 15 men. You are interested in the number of men on your committee. If the members of the committee are randomly selected, what is the probability that your committee has more than 4 men?

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

Are you choosing with or without replacement?

Without

What is the group of interest?

The men

How many are in the group of interest?

15 men

How many are in the other group?

18 women

Let X = __________ on the committee. What values does X take on?

Let X = **the number of men** on the committee. x = 0, 1, 2, …, 7.

The probability question is P(_______).

P(x>4)

Notation for the Hypergeometric: H = Hypergeometric Probability Distribution Function

X~H(r, b, n)

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

Example 4.19

A school site committee is to be chosen randomly from 6 men and 5 women. If the committee consists of 4 members chosen randomly, what is the probability that 2 of them are men? How many men do you expect to be on the committee?

Let X = the number of men on the committee of 4. The men are the group of interest (first group).

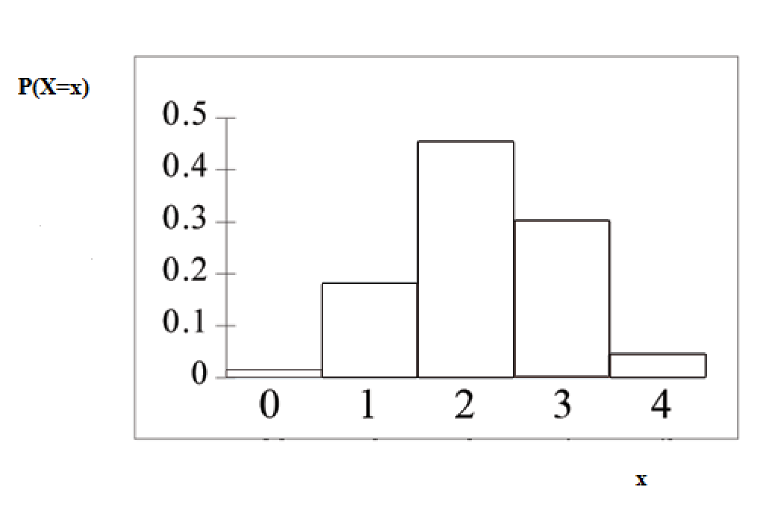
X takes on the values 0, 1, 2, 3, 4, where r = 6, b = 5 , and n = 4. X~H(6, 5, 4)

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

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

The probability that there are 2 men on the committee is about 0.45.

The graph of *X*~H(6, 5, 4) is:



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

You would expect *m* = 2.18(about 2) men on the committee.

The formula for the mean is $\mu = \frac{n \cdot r}{r + b} = \frac{4 \cdot 6}{6 + 5} = 2.18$

The formula for the variance is fairly complex. You will find it in the **Summary of the Discrete Probability Functions Chapter**.

4.8 Poisson

Characteristics of a Poisson experiment:

1. The Poisson gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are 5 words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1000). You will verify the relationship in the homework exercises. ***n*** is the number of trials and ***p*** is the probability of a "success."

Poisson probability distribution. The random variable *X* = the number of occurrences in the interval of interest. The mean and variance are given in the summary.

Example 4.20

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in 5 minutes. The time interval of interest is 5 minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in 5 minutes is 3?

Let *X* = the number of loaves of bread put on the shelf in 5 minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, **then the average number of loaves put on the shelf in 5 minutes is**

$\left(\frac{5}{30}\right) \cdot 12 = 2$ loaves of bread

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

Example 4.21

A certain bank expects to receive 6 bad checks per day, on average. What is the probability of the bank getting fewer than 5 bad checks on any given day? Of interest is the number of checks the bank receives in 1 day, so the time interval of interest is 1 day. Let *X* = the number of bad checks the bank receives in one day. If the bank expects to receive 6 bad checks per day then the average is 6 checks per day. The probability question asks for *P*(*x*<5).

Example 4.22

You notice that a news reporter says "uh", on average, 2 times per broadcast. What is the probability that the news reporter says "uh" more than 2 times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

What is the interval of interest?

One broadcast

What is the average number of times the news reporter says "uh" during one broadcast?

2

Let *X* = ____________. What values does *X* take on?

Let *X* = **the number of times the news reporter says "uh" during one broadcast**.
x = 0, 1, 2, 3, ...

The probability question is P(______).

P(x > 2)

Notation for the Poisson: P = Poisson Probability Distribution Function

X ~ P(μ)

Read this as "*X* is a random variable with a Poisson distribution." The parameter is μ (or λ). μ (or λ) = the mean for the interval of interest.

Example 4.23

Leah's answering machine receives about 6 telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than 1 call **in the next 15 minutes?**

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

x = 0, 1, 2, 3, ...

If Leah receives, on the average, 6 telephone calls in 2 hours, and there are eight 15 minutes intervals in 2 hours, then Leah receives

$$\frac{1}{8} \cdot 6 = 0.75$$

calls in 15 minutes, on the average. So, μ = 0.75 for this problem.

X ~ P(0.75)

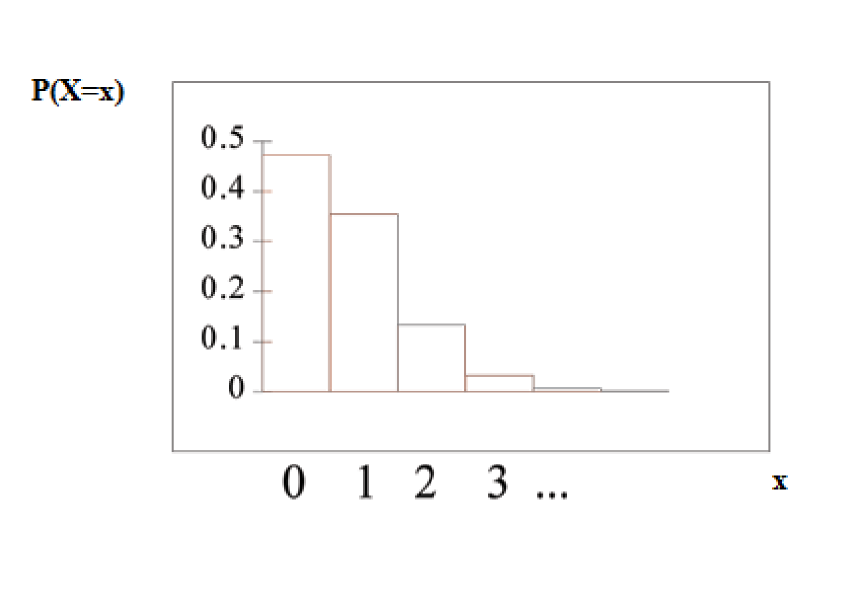
Find *P*(*x* > 1). *P*(*x* > 1) = 0.1734 (calculator or computer)

TI-83+ and TI-84: For a general discussion, see **this example (Binomial) (http://cnx.org/content/element-678/latest/#element-678)** . The syntax is similar. The Poisson parameter list is (μ for the interval of interest, number). **For this problem:**

Press 1- and then press 2nd DISTR. Arrow down to C:poissoncdf. Press ENTER. Enter .75,1). The result is *P*(*x* > 1) = 0.1734. NOTE: The TI calculators use λ (lambda) for the mean.

The probability that Leah receives more than 1 telephone call in the next fifteen minutes is about 0.1734.

The graph of *X* ~ P(0.75) is:



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

4.9 Summary of Functions

Formula

$X \sim B(n, p)$

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x =$ 0,1, 2, 3, ...,n

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1 \qquad q = 1 - p$

The mean is μ = np. The standard deviation is σ = $\sqrt{npq}$.

Formula

$X \sim G(p)$

X = the number of independent trials until the first success (count the failures and the first success)

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

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1$

$q = 1 - p$

The mean is $\mu = \frac{1}{p}$

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

Formula

$X \sim H(r, b, n)$

X = the number of items from the group of interest that are in the chosen sample.

X may take on the values x= 0, 1, ..., up to the size of the group of interest. (The minimum value for X may be larger than 0 in some instances.)

r = the size of the group of interest (first group)

b= the size of the second group

n= the size of the chosen sample.

$n \leq r + b$

The mean is: $\mu = \frac{nr}{r+b}$

The standard deviation is: σ = $\sqrt{\frac{rbn(r+b+n)}{(r+b)^2(r+b-1)}}$

Formula

X ~ P(μ)

X = the number of occurrences in the interval of interest

X takes on the values x = 0, 1, 2, 3, ...

The mean μ is typically given. (λ is often used as the mean instead of μ.) When the Poisson is used to approximate the binomial, we use the binomial mean μ = *np*. *n* is the binomial number of trials. *p* = the probability of a success for each trial. This formula is valid when n is "large" and *p* "small" (a general rule is that *n* should be greater than or equal to 20 and *p* should be less than or equal to 0.05). If *n* is large enough and *p* is small enough then the Poisson approximates the binomial very well. The variance is σ^2 = μ and the standard deviation is σ = $\sqrt{\mu}$

4.10 Practice 1: Discrete Distribution

Student Learning Outcomes

- The student will analyze the properties of a discrete distribution.

Given:

A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let P(x) = the probability that a student will study ballet x years.

Organize the Data

Complete the table below using the data provided.

Table 4.7

x	P(x)	x*P(x)
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

Discussion Question

4.11 Practice 2: Binomial Distribution

Student Learning Outcomes

- The student will construct the Binomial Distribution.

Given

The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. (*Source: http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf).*)

Suppose that you randomly pick 8 first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status

Interpret the Data

4.12 Practice 3: Poisson Distribution

Student Learning Outcomes

- The student will analyze the properties of a Poisson distribution.

Given

On average, eight teens in the U.S. die from motor vehicle injuries per day. As a result, states across the country are debating raising the driving age. (*Source: http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html)*)

Interpret the Data

4.13 Practice 4: Geometric Distribution

Student Learning Outcomes

- The student will analyze the properties of a geometric distribution.

Given:

Use the information from the **Binomial Distribution Practice** shown below.

The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. (*Source: http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf*)

Suppose that you randomly select freshman from the study until you find one who replies “yes.” You are interested in the number of freshmen you must ask.

Interpret the Data

4.14 Practice 5: Hypergeometric Distribution

Student Learning Outcomes

- The student will analyze the properties of a hypergeometric distribution.

Given

Suppose that a group of statistics students is divided into two groups: business majors and non-business majors. There are 16 business majors in the group and 7 non-business majors in the group. A random sample of 9 students is taken. We are interested in the number of business majors in the group.

Interpret the Data

4.15 Homework

For each problem:

a. In words, define the Random Variable X.
b. List the values that X may take on.
c. Give the distribution of X. X~

Then, answer the questions specific to each individual problem.

The next 2 questions refer to the following: In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages.

Try these multiple choice problems.

For the next three problems: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13 year win history of 382 wins out of 1034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

Let X = the number of games won in that upcoming month.

For the next two questions: The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is 10. We are interested in the number of times her cats wake her up each week.

**Exercises 38 - 43 contributed by Roberta Bloom

4.16 Review

The next two questions refer to the following:

A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2000 per month. Of those respondents who charge more than $2000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

Table 4.8

			Ethnic Group		
Gender	White	Asian	Hispanic	Black	American Indian
Male	400	168	115	35	16
Female	440	132	140	40	14

The next four questions refer to the following: Recently, a nurse commented that when a patient calls the medical advice line claiming to have **the flu**, the chance that he/she truly has **the flu** (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have **the flu**, we are interested in how many actually have **the flu**.

The next two questions refer to the following: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

4.17 Lab 1: Discrete Distribution (Playing Card Experiment)

Class Time:

Names:

Student Learning Outcomes:

- The student will compare empirical data and a theoretical distribution to determine if everyday experiment fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies:

- One full deck of playing cards

Procedure

The experiment procedure is to pick one card from a deck of shuffled cards.

1. The theorectical probability of picking a diamond from a deck is: _________
2. Shuffle a deck of cards.

3. Pick one card from it.
4. Record whether it was a diamond or not a diamond.
5. Put the card back and reshuffle.
6. Do this a total of 10 times
7. Record the number of diamonds picked.
8. Let *X* = number of diamonds. Theoretically, *X* ~ *B*(_____,_____)

Organize the Data

1. Record the number of diamonds picked for your class in the chart below. Then calculate the relative frequency.

Table 4.9

x	Frequency	Relative Frequency
0	__________	__________
1	__________	__________
2	__________	__________
3	__________	__________
4	__________	__________
5	__________	__________
6	__________	__________
7	__________	__________
8	__________	__________
9	__________	__________
10	__________	__________

2. Calculate the following:

   a. $\bar{x}$ =

   b. *s* =

3. Construct a histogram of the empirical data.

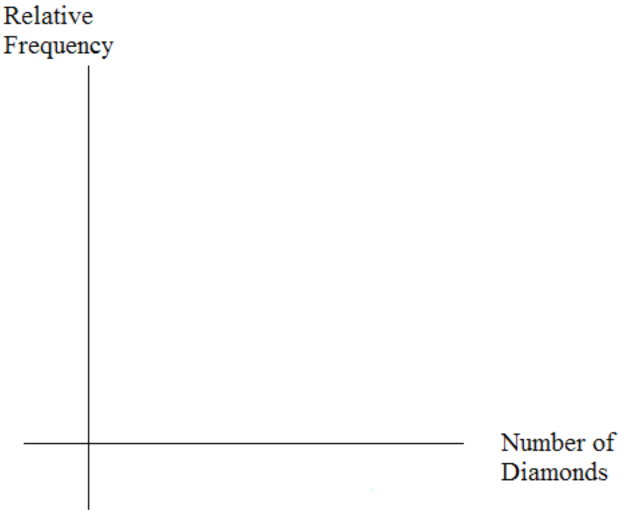


Figure 4.1

Theoretical Distribution

1. Build the theoretical PDF chart based on the distribution in the Procedure section above.

Table 4.10

x	*P*(*x*)
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

2. Calculate the following:

   a. μ = ___________

   b. σ = ___________

3. Construct a histogram of the theoretical distribution.

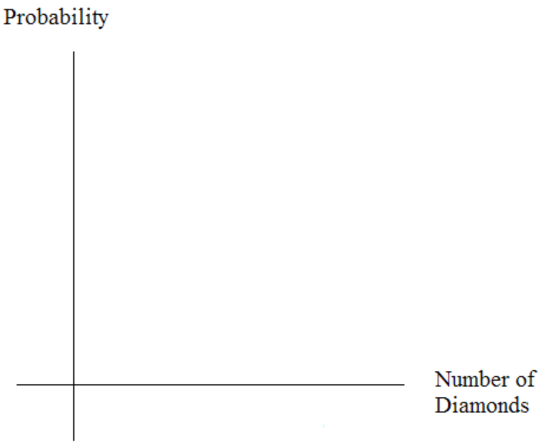


Figure 4.2

Using the Data

Calculate the following, rounding to 4 decimal places:

RF = relative frequency

Use the table from the section titled "Theoretical Distribution" here:

- $P(x = 3) =$
- $P(1 < x < 4) =$
- $P(x \geq 8) =$

Use the data from the section titled "Organize the Data" here:

- RF($x = 3$) =
- RF($1 < x < 4$) =
- RF($x \geq 8$) =

Discussion Questions

For questions 1. and 2., think about the shapes of the two graphs, the probabilities and the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences. (Note: These answers may vary and still be correct.)
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions. (Note: These answers may vary and still be correct.)
3. Using your answers from the two previous questions, does it appear that the data fit the theoretical distribution? In 1 - 3 complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Which table (from "Organize the data" and "Theoretical Distributions") would you expect to change (and how would it change)? Why? Why wouldn't the other table change?

4.18 Lab 2: Discrete Distribution (Lucky Dice Experiment)

Class Time:

Names:

Student Learning Outcomes:

- The student will compare empirical data and a theoretical distribution to determine if a Tet gambling game fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies:

- 1 game "Lucky Dice" or 3 regular dice

For a detailed game description, refer **here (http://cnx.org/content/m16823/1.20/##element-650)** . (The link goes to the beginning of Discrete Random Variables Homework. Please refer to Problem #14.)

Round relative frequencies and probabilities to four decimal places.

The Procedure

1. The experiment procedure is to bet on one object. Then, roll 3 Lucky Dice and count the number of matches. The number of matches will decide your profit.
2. What is the theoretical probability of 1 die matching the object? _________
3. Choose one object to place a bet on. Roll the 3 Lucky Dice. Count the number of matches.
4. Let X = number of matches. Theoretically, $X \sim B$(______,______)
5. Let Y = profit per game.

Organize the Data

In the chart below, fill in the Y value that corresponds to each X value. Next, record the number of matches picked for your class. Then, calculate the relative frequency.

1. Complete the table.

Table 4.11

x	y	Frequency	Relative Frequency
0			
1			
2			
3			

2. Calculate the Following:

   a. $\overline{x} =$

   b. $s_x =$

   c. $\overline{y} =$

   d. $s_y =$

3. Explain what $\overline{x}$ represents.
4. Explain what $\overline{y}$ represents.
5. Based upon the experiment:

   a. What was the average profit per game?

   b. Did this represent an average win or loss per game?

c. How do you know? Answer in complete sentences.

6. Construct a histogram of the empirical data

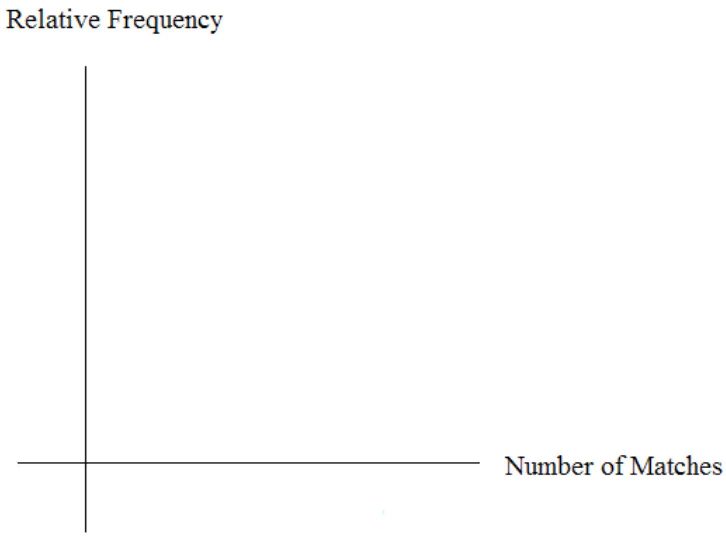


Figure 4.3

Theoretical Distribution

Build the theoretical PDF chart for X and Y based on the distribution from the section titled "The Procedure".

1.

Table 4.12

x	y	$P(x) = P(y)$
0		
1		
2		
3		

2. Calculate the following

a. μ_x =

b. σ_x =

c. μ_y =

3. Explain what μ_x represents.
4. Explain what μ_y represents.
5. Based upon theory:

a. What was the expected profit per game?

b. Did the expected profit represent an average win or loss per game?

c. How do you know? Answer in complete sentences.

6. Construct a histogram of the theoretical distribution.

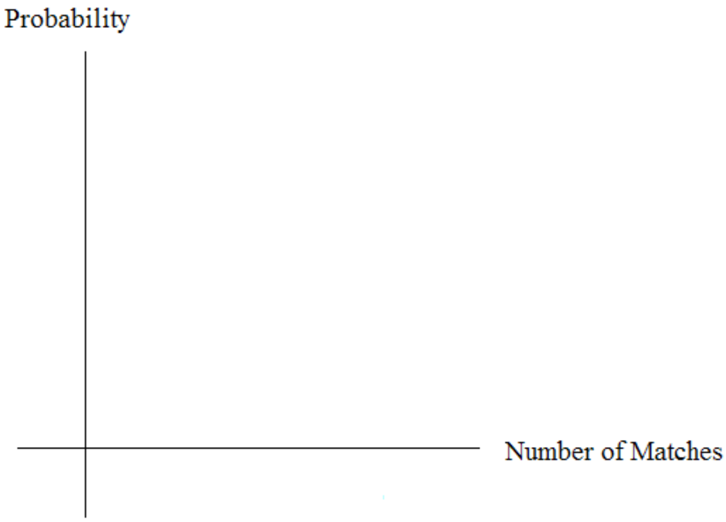


Figure 4.4

Use the Data

Calculate the following (rounded to 4 decimal places):

RF = relative frequency

Use the data from the section titled "Theoretical Distribution" here:

1. $P(x = 3)$ = __________
2. $P(0 < x < 3)$ = __________
3. $P(x \geq 2)$ = __________

Use the data from the section titled "Organize the Data" here:

1. $RF(x = 3)$ = __________
2. $RF(0 < x < 3)$ = __________
3. $RF(x \geq 2)$ = __________

Discussion Question

For questions 1. and 2., consider the graphs, the probabilities and relative frequencies, the means and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences. (Note: these answers may vary and still be correct.)
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions. (Note: these answers may vary and still be correct.)
3. Thinking about your answers to 1. and 2.,does it appear that the data fit the theoretical distribution? In 1 - 3 complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Which table (from "Organize the Data" or "Theoretical Distribution") would you expect to change? Why? How might the table change?

Glossary

Bernoulli Trials: An experiment with the following characteristics:

- There are only 2 possible outcomes called “success” and “failure” for each trial.
- The probability p of a success is the same for any trial (so the probability $q = 1 − p$ of a failure is the same for any trial).

Binomial Distribution: A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: ***X*~*B*(*n*, *p*)**. The mean is μ = np and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P\left(X = x\right) = \binom{n}{x}p^x q^{n-x}$.

Expected Value: Expected arithmetic average when an experiment is repeated many times. (Also called the mean). Notations: $E(x)$, μ. For a discrete random variable (RV) with probability distribution function $P(x)$, the definition can also be written in the form $E(x) = \mu = \sum xP(x)$.

Geometric Distribution: A discrete random variable (RV) which arises from the Bernoulli trials. The trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. Notation: $\boldsymbol{X \sim G(p)}$. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p} \cdot \left(\frac{1}{p} - 1\right)}$ The probability of exactly x failures before the first success is given by the formula: $P(X = x) = p(1 - p)^{x - 1}$.

Hypergeometric Distribution: A discrete random variable (RV) that is characterized by

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

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

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

Poisson Distribution: A discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval. Characteristics of the variable:

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

The distribution is defined by the mean μ of the event in the interval. Notation: $X \sim P(\mu)$. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(X = x) = e^{-\mu}\frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution when n is “large” and p is “small” (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to .05).

Probability Distribution Function (PDF): A mathematical description of a discrete random variable (RV), given either in the form of an equation (formula) , or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Example .

A biased coin with probability 0.7 for a head (in one toss of the coin) is tossed 5 times. We are interested in the number of heads (the RV X = the number of heads). X is Binomial, so $X \sim B(5, 0.7)$ and $P(X = x) = \binom{5}{x}.7^x.3^{5-x}$or in the form of the table:

x	*P*(*X* = *x*)
0	0.0024
1	0.0284
2	0.1323
3	0.3087
4	0.3602
5	0.1681

Random Variable (RV): see **Variable**

Variable (Random Variable): A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

5 CONTINUOUS RANDOM VARIABLES

5.1 Continuous Random Variables

Student Learning Outcomes

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

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Introduction

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

This chapter gives an introduction to continuous random variables and the many continuous distributions. We will be studying these continuous distributions for several chapters.

NOTE

The values of discrete and continuous random variables can be ambiguous. For example, if *X* is equal to the number of miles (to the nearest mile) you drive to work, then *X* is a discrete random variable. You count the miles. If *X* is the distance you drive to work, then you measure values of *X* and *X* is a continuous random variable. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated: **pdf**). We use the symbol *f*(*x*) to represent the curve. *f*(*x*) is the function that corresponds to the graph; we use the density function *f*(*x*) to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated: **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to 1.
- Probability is found for intervals of x values rather than for individual x values.
- *P*(c<*x*<d) is the probability that the random variable X is in the interval between the values c and d. *P*(c<*x*<d) is the area under the curve, above the x-axis, to the right of c and the left of d.
- *P*(*x* = *c*) = 0 The probability that x takes on any single individual value is 0. The area below the curve, above the x-axis, and between x=c and x=c has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also 0.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to best model and fit the particular situation.

In this chapter and the next chapter, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

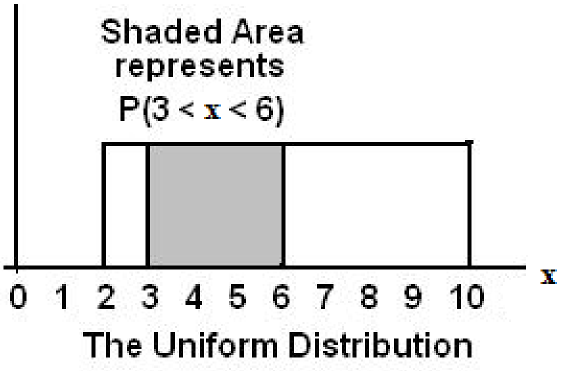


Figure 5.1 The graph shows a Uniform Distribution with the area between x=3 and x=6 shaded to represent the probability that the value of the random variable X is in the interval between 3 and 6.

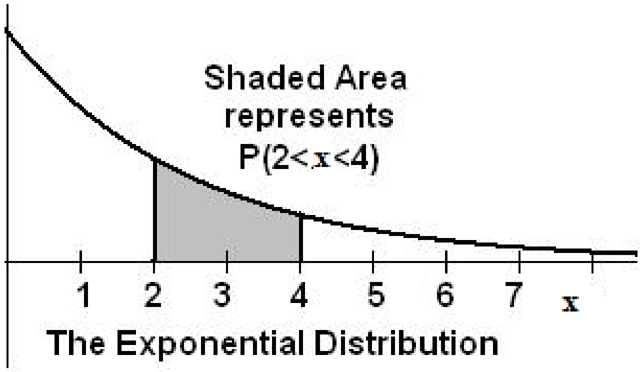


Figure 5.2 The graph shows an Exponential Distribution with the area between x=2 and x=4 shaded to represent the probability that the value of the random variable X is in the interval between 2 and 4.

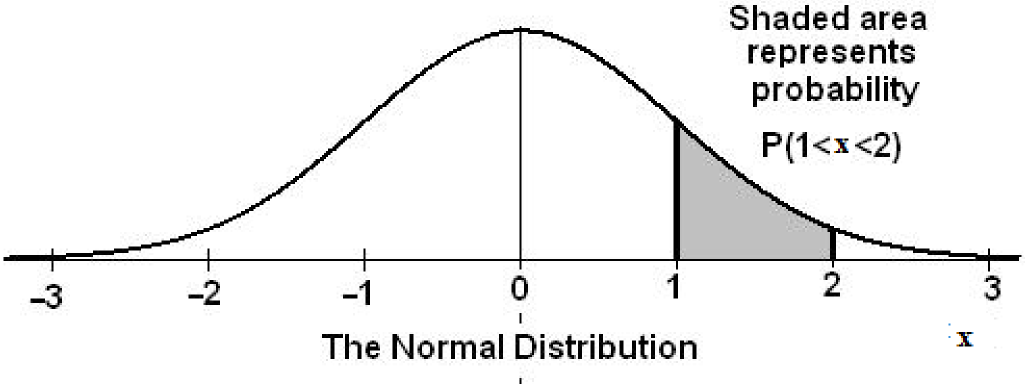


Figure 5.3 The graph shows the Standard Normal Distribution with the area between x=1 and x=2 shaded to represent the probability that the value of the random variable X is in the interval between 1 and 2.

**With contributions from Roberta Bloom

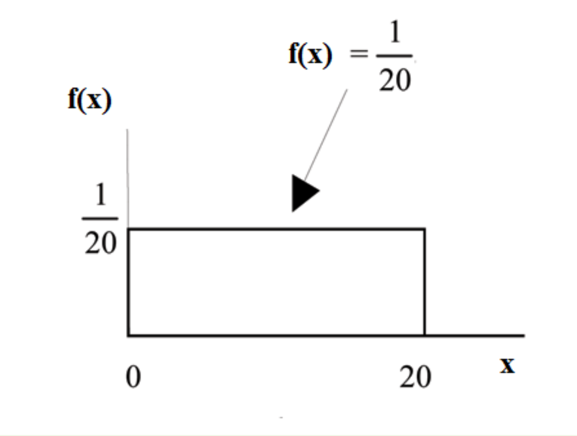
5.2 Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation *f*(*x*). Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function *f*(*x*) so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one.

For continuous probability distributions, PROBABILITY = AREA.

Example 5.1

Consider the function $f(x) = \frac{1}{20}$ for $0 \le x \le 20$. x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \le x \le 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive .



$f(x) = \frac{1}{20}$ **for** $0 \le x \le 20$.

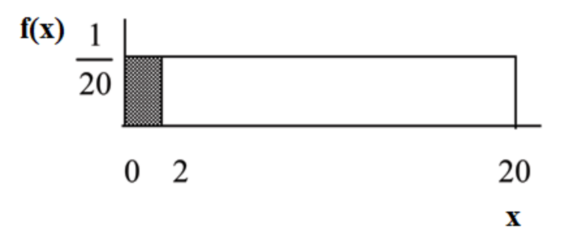
The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \le x \le 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \le x \le 20$ and the x-axis is the area of a rectangle with base = 20 and height $= \frac{1}{20}$.

AREA $= 20 \cdot \frac{1}{20} = 1$

This particular function, where we have restricted *x* so that the area between the function and the x-axis is 1, is an example of a continuous probability density function. It is used as a tool to calculate probabilities.

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



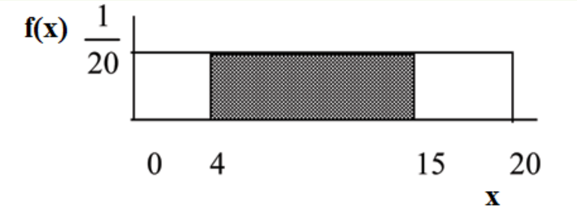
AREA $= \left(2 - 0\right) \cdot \frac{1}{20} = 0.1$

$(2 - 0) = 2$ = base of a rectangle

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

The area corresponds to a probability. The probability that *x* is between 0 and 2 is 0.1, which can be written mathematically as $P(0<x<2) = P(x<2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $4<x<15$.



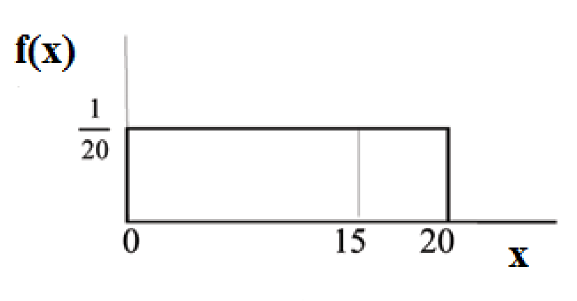
AREA $= \left(15 - 4\right) \cdot \frac{1}{20} = 0.55$

$(15 - 4) = 11$ = the base of a rectangle

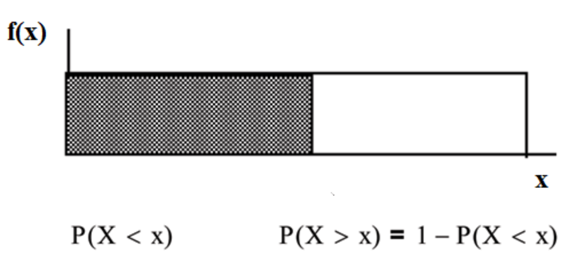
$\frac{1}{20}$ = the height.

The area corresponds to the probability $P(4<x<15) = 0.55$.

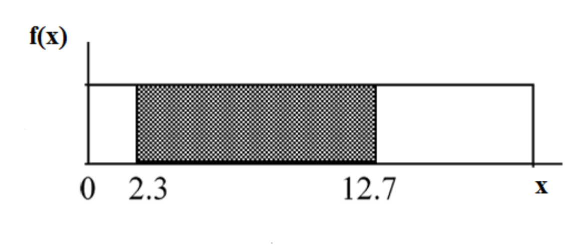
Suppose we want to find $P(x=15)$. On an x-y graph, $x=15$ is a vertical line. A vertical line has no width (or 0 width). Therefore, $P(x=15) = \text{(base)(height)} = \left(0\right)\left(\frac{1}{20}\right) = 0$.



$P(X \le x)$ (can be written as $P(X<x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X<x)$.



Label the graph with f(x) and *x*. Scale the x and y axes with the maximum *x* and *y* values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.



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

5.3 The Uniform Distribution

Example 5.2

The previous problem is an example of the **uniform probability distribution**.

Illustrate the uniform distribution. The data that follows are 55 smiling times, in seconds, of an eight-week old baby.

Table 5.1

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

sample mean = 11.49 and sample standard deviation = 6.23

We will assume that the smiling times, in seconds, follow a uniform distribution between 0 and 23 seconds, inclusive. This means that any smiling time from 0 to and including 23 seconds is **equally likely**. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let *X* = length, in seconds, of an eight-week old baby's smile.

The notation for the uniform distribution is

$X \sim U(\mathrm{a},b)$ where a = the lowest value of x and b = the highest value of x.

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

For this example, $x \sim U(0,23)$ and $f(x) = \frac{1}{23-0}$ for $0 \leq x \leq 23$.

Formulas for the theoretical mean and standard deviation are

$\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

For this problem, the theoretical mean and standard deviation are

$\mu = \frac{0+23}{2}$ = 11.50 seconds and $\sigma = \sqrt{\frac{(23-0)^2}{12}}$ = 6.64 seconds

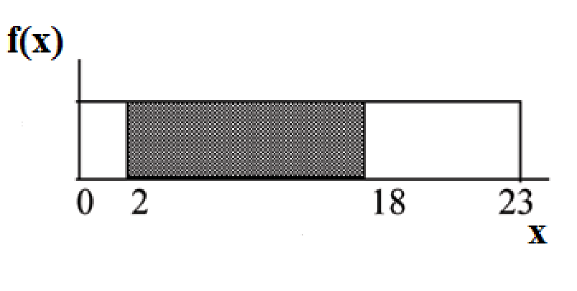
Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation.

Example 5.3

What is the probability that a randomly chosen eight-week old baby smiles between 2 and 18 seconds?

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

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



Find the 90th percentile for an eight week old baby's smiling time.

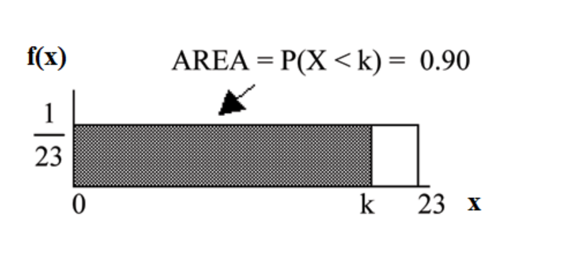
Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x<k)=0.90$

$P(x<k)=0.90$

(base)(height) = 0.90

$\left(k-0\right)\cdot\frac{1}{23}=0.90$

$k=23\cdot0.90=20.7$



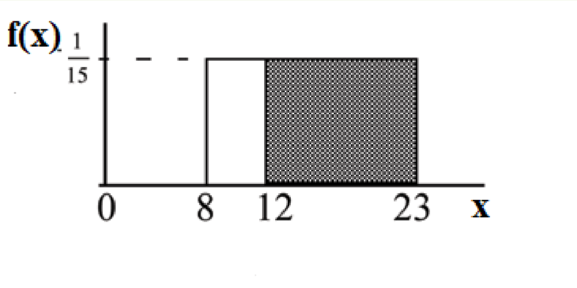
Find the probability that a random eight week old baby smiles more than 12 seconds **KNOWING** that the baby smiles **MORE THAN 8 SECONDS**.

Find $P(x>12\mid x>8)$ There are two ways to do the problem. **For the first way**, use the fact that this is a **conditional** and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than 8 seconds.

Write a new *f*(*x*): $f(x)=\frac{1}{23-8}=\frac{1}{15}$

for $8<x<23$

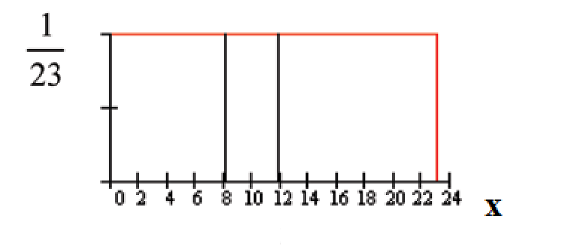
$P(x>12\mid x>8)=\left(23-12\right)\cdot\frac{1}{15}=\frac{11}{15}$



For the second way, use the conditional formula from **Probability Topics** with the original distribution $X \sim U(0, 23)$:

$P\left(A \mid B\right) = \frac{P(A \text{ AND } B)}{P(B)}$ For this problem, A is ($x > 12$) and B is ($x > 8$).

So, $P(x > 12 \mid x > 8) = \frac{(x > 12 \text{ AND } x > 8)}{P(x > 8)} = \frac{P(x > 12)}{P(x > 8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = 0.733$



Example 5.4

Uniform: The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between 0 and 15 minutes, inclusive.

What is the probability that a person waits fewer than 12.5 minutes?

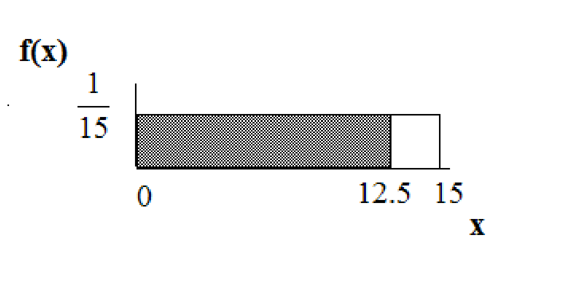
Let X = the number of minutes a person must wait for a bus. $a = 0$ and $b = 15$. $x \sim U(0, 15)$. Write the probability density function.

$f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \le x \le 15$.

Find $P(x<12.5)$. Draw a graph.

$P(x<k) = (\text{base})(\text{height}) = \left(12.5 - 0\right) \cdot \frac{1}{15} = 0.8333$

The probability a person waits less than 12.5 minutes is 0.8333.



On the average, how long must a person wait?

Find the mean, μ, and the standard deviation, σ.

$\mu = \frac{a+b}{2} = \frac{15+0}{2}$ = 7.5. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}}$ = 4.3. The Standard deviation is 4.3 minutes.

Ninety percent of the time, the time a person must wait falls below what value?

Note

This asks for the 90th percentile.

Find the 90th percentile. Draw a graph. Let *k* = the 90th percentile.

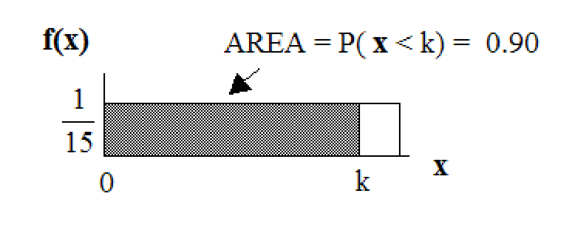
$P(x<k)$ = (base)(height) = $\left(k-0\right) \cdot \left(\frac{1}{15}\right)$

$0.90 = k \cdot \frac{1}{15}$

k = (0.90)(15) = 13.5

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.



Example 5.5

Uniform: Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let *X* = the time, in minutes, it takes a nine-year old child to eat a donut. Then *X* ~ *U*(0.5, 4).

The probability that a randomly selected nine-year old child eats a donut in at least two minutes is _______.

0.5714

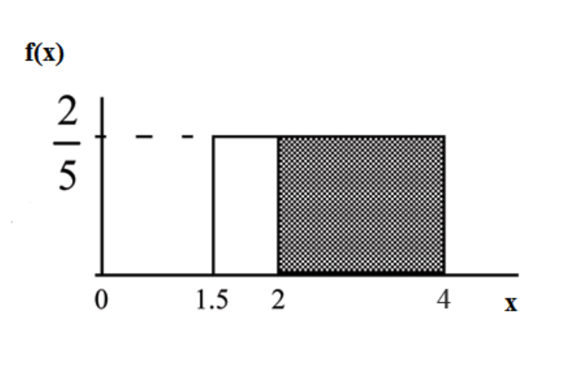
Find the probability that a different nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes.

The second probability question has a conditional (refer to "Probability Topics"). You are asked to find the probability that a nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways (see the first example). You must reduce the sample space. **First way**: Since you already know the child has already been eating the donut for more than 1.5 minutes, you are no longer starting at a = 0.5 minutes. Your starting point is 1.5 minutes.

Write a new f(x):

$f(x) = \frac{1}{4-1.5} = \frac{2}{5}$ for $1.5 \le x \le 4$.

Find $P(x > 2 \mid x > 1.5)$. Draw a graph.



$P(x > 2 \mid x > 1.5)$ = (base)(new height) = $(4 - 2)(2 / 5)$ = ?

$\frac{4}{5}$

The probability that a nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $x \sim U(0.5, 4)$. Use the conditional formula

$$P(x > 2 \mid x > 1.5) = \frac{P(x > 2 \text{ AND } x > 1.5)}{P(x > 1.5)} = \frac{P(x > 2)}{P(x > 1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$$

See "**Summary of the Uniform and Exponential Probability Distributions**" for a full summary.

Example 5.6

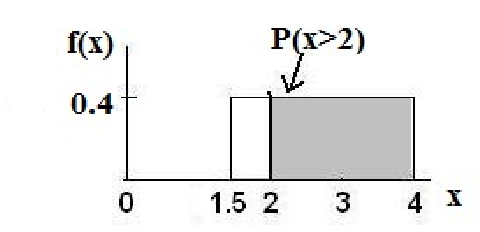
Uniform: Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and 4 hours. Let x = the time needed to fix a furnace. Then $x \sim U(1.5, 4)$.

1. Find the problem that a randomly selected furnace repair requires more than 2 hours.
2. Find the probability that a randomly selected furnace repair requires less than 3 hours.
3. Find the 30th percentile of furnace repair times.
4. The longest 25% of repair furnace repairs take at least how long? (In other words: Find the minimum time for the longest 25% of repair times.) What percentile does this represent?
5. Find the mean and standard deviation

Find the probability that a randomly selected furnace repair requires longer than 2 hours.

To find $f(x)$: $f(x) = \frac{1}{4 - 1.5} = \frac{1}{2.5}$ so $f(x)$ = =0.4

P(x>2) = (base)(height) = (4 − 2)(0.4) = 0.8

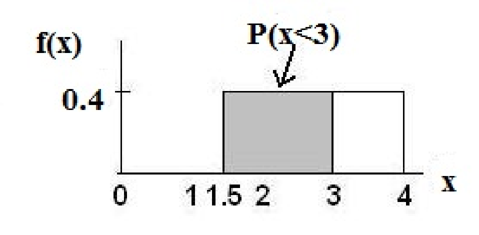


Example 4 Figure 1 Uniform Distribution between 1.5 and 4 with shaded area between 2 and 4 representing the probability that the repair time x is greater than 2

Find the probability that a randomly selected furnace repair requires less than 3 hours. Describe how the graph differs from the graph in the first part of this example.

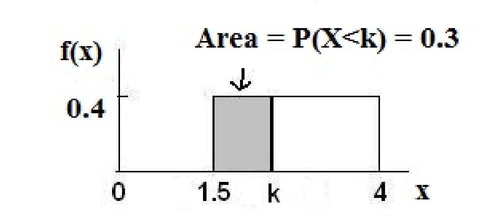
$P(x<3)$ = (base)(height) = (3 − 1.5)(0.4) = 0.6

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between x=1.5 and x=3. Note that the shaded area starts at x=1.5 rather than at x=0; since X~U(1.5,4), x can not be less than 1.5.



Example 4 Figure 2 Uniform Distribution between 1.5 and 4 with shaded area between 1.5 and 3 representing the probability that the repair time x is less than 3

Find the 30th percentile of furnace repair times.



Example 4 Figure 3 Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

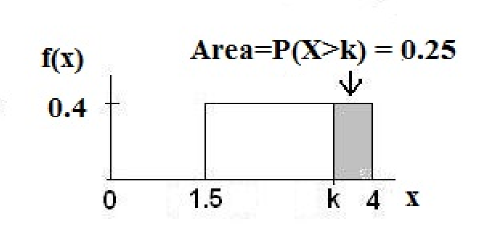
$P(x<k)$ = 0.30

$P(x<k)$ = (base)(height) = $(k - 1.5) \cdot (0.4)$

0.3 = (k − 1.5) (0.4) ; Solve to find k:
0.75 = k − 1.5 , obtained by dividing both sides by 0.4
k = 2.25 , obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.5 hours or less.

The **longest 25%** of furnace repair times take **at least** how long? (Find the minimum time for the longest 25% of repairs.)



Example 4 Figure 4 Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.

$P(x > k)$ = 0.25

$P(x > k)$ = (base)(height) = $(4 - k) \cdot (0.4)$

0.25 = (4 − k)(0.4) ; Solve for k:
0.625 = 4 − k , obtained by dividing both sides by 0.4
−3.375 = −k , obtained by subtracting 4 from both sides

k=3.375

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the **75th percentile** of furnace repair times.

Find the mean and standard deviation

$\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$\mu = \frac{1.5+4}{2} = 2.75$ hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

See "**Summary of the Uniform and Exponential Probability Distributions**" for a full summary.

**Example 5 contributed by Roberta Bloom

5.4 The Exponential Distribution

The **exponential** distribution is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people that spend less money and fewer people that spend large amounts of money.

The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

Example 5.7

Illustrates the exponential distribution: Let *X* = amount of time (in minutes) a postal clerk spends with his/her customer. The time is known to have an exponential distribution with the average amount of time equal to 4 minutes.

X is a **continuous random variable** since time is measured. It is given that μ = 4 minutes. To do any calculations, you must know *m*, the decay parameter.

$m = \frac{1}{\mu}$. Therefore, $m = \frac{1}{4} = 0.25$

The standard deviation, σ, is the same as the mean. μ = σ

The distribution notation is $X \sim Exp(m)$. Therefore, $X \sim Exp(0.25)$.

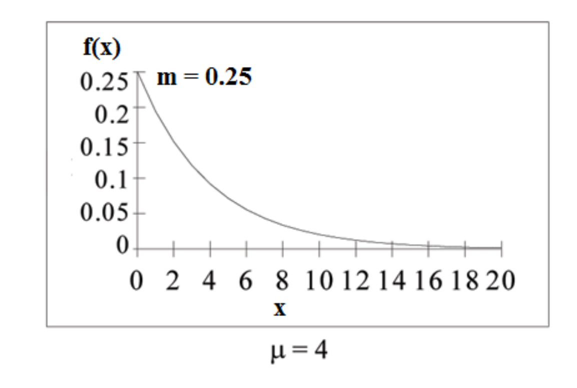
The probability density function is $f(x) = m \cdot e^{-m \cdot x}$ The number *e* = 2.71828182846... It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter 1 for *x*, the calculator will display the value *e*.

The curve is:

$f(x) = 0.25 \cdot e^{-0.25 \cdot x}$ where *x* is at least 0 and *m* = 0.25.

For example, $f(5) = 0.25 \cdot e^{-0.25 \cdot 5} = 0.072$

The graph is as follows:



Notice the graph is a declining curve. When $x = 0$,

$f(x) = 0.25 \cdot e^{-0.25 \cdot 0} = 0.25 \cdot 1 = 0.25 = m$

Example 5.8

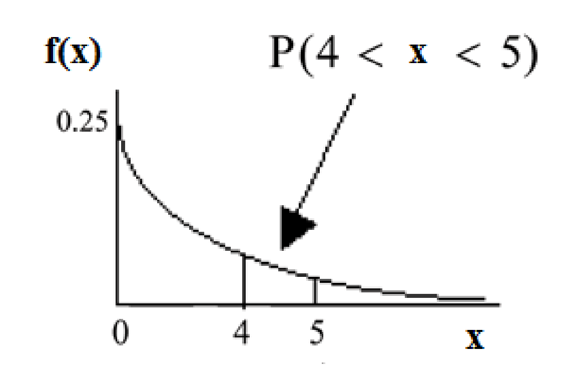
Find the probability that a clerk spends four to five minutes with a randomly selected customer.

Find $P(4<x<5)$.

The **cumulative distribution function (CDF)** gives the area to the left.

$P(x<x) = 1 - e^{-m \cdot x}$

$P(x<5) = 1 - e^{-0.25 \cdot 5} = 0.7135$ and $P(x<4) = 1 - e^{-0.25 \cdot 4} = 0.6321$



You can do these calculations easily on a calculator.

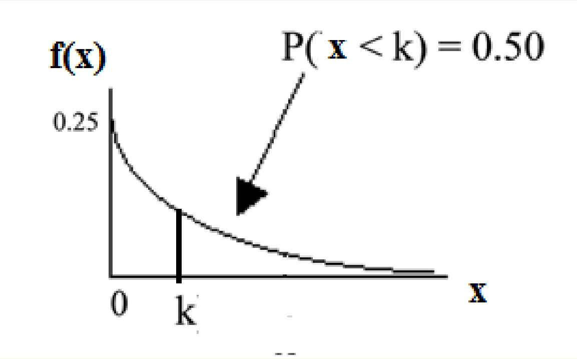
The probability that a postal clerk spends four to five minutes with a randomly selected customer is

$P(4<x<5) = P(x<5) - P(x<4) = 0.7135 - 0.6321 = 0.0814$

TI-83+ and TI-84: On the home screen, enter (1-e^(-.25*5))-(1-e^(-.25*4)) or enter e^(-.25*4)-e^(-.25*5).

Half of all customers are finished within how long? (Find the 50th percentile)

Find the 50th percentile.



$P(x<k) = 0.50$, $k = 2.8$ minutes (calculator or computer)

Half of all customers are finished within 2.8 minutes.

You can also do the calculation as follows:

$P(x<k) = 0.50$ and $P(x<k) = 1 - e^{-0.25 \cdot k}$

Therefore, $0.50 = 1 - e^{-0.25 \cdot k}$ and $e^{-0.25 \cdot k} = 1 - 0.50 = 0.5$

Take natural logs: $\ln\left(e^{-0.25 \cdot k}\right) = \ln\left(0.50\right)$. So, $-0.25 \cdot k = \ln(0.50)$

Solve for k: $k = \frac{\ln(.50)}{-0.25} = 2.8$ minutes

A formula for the percentile k is $k = \frac{\text{LN}(1-\text{AreaToTheLeft})}{-m}$ where LN is the natural log.

TI-83+ and TI-84: On the home screen, enter LN(1-.50)/-.25. Press the (-) for the negative.

Which is larger, the mean or the median?

Is the mean or median larger?

From part b, the median or 50th percentile is 2.8 minutes. The theoretical mean is 4 minutes. The mean is larger.

Optional Collaborative Classroom Activity

Have each class member count the change he/she has in his/her pocket or purse. Your instructor will record the amounts in dollars and cents. Construct a histogram of the data taken by the class. Use 5 intervals. Draw a smooth curve through the bars. The graph should look approximately exponential. Then calculate the mean.

Let X = the amount of money a student in your class has in his/her pocket or purse.

The distribution for X is approximately exponential with mean, μ = _______ and m = _______. The standard deviation, σ = ________.

Draw the appropriate exponential graph. You should label the x and y axes, the decay rate, and the mean. Shade the area that represents the probability that one student has less than $.40 in his/her pocket or purse. (Shade $P(x<0.40)$).

Example 5.9

On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.

What is the probability that a computer part lasts more than 7 years?

Let x = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

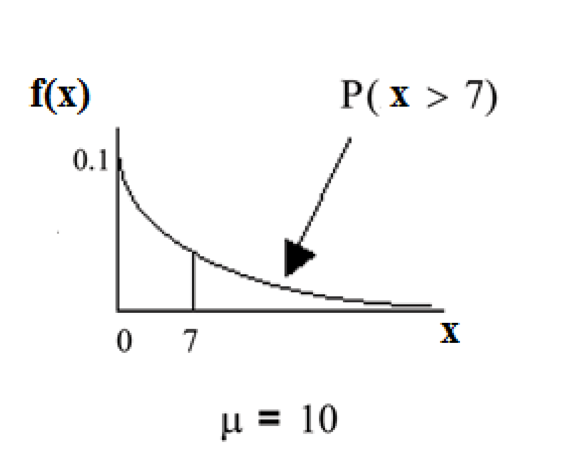
Find $P(x > 7)$. Draw a graph.

$P(x > 7) = 1 - P(x < 7)$.

Since $P(X<x) = 1 - e^{-mx}$ then $P(X > x) = 1 - \left(1 - e^{-m \cdot x}\right) = e^{-m \cdot x}$

$P(x > 7) = e^{-0.1 \cdot 7} = 0.4966$. The probability that a computer part lasts more than 7 years is 0.4966.

TI-83+ and TI-84: On the home screen, enter e^(-.1*7).



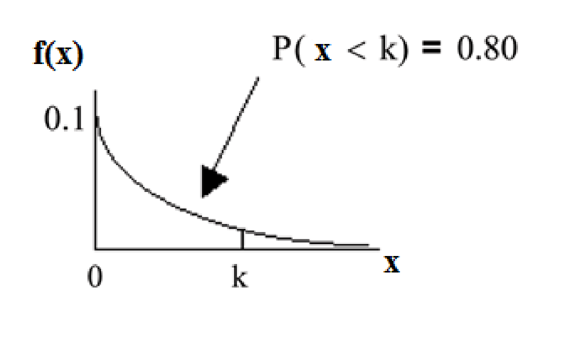
On the average, how long would 5 computer parts last if they are used one after another?

On the average, 1 computer part lasts 10 years. Therefore, 5 computer parts, if they are used one right after the other would last, on the average,

(5)(10) = 50 years.

Eighty percent of computer parts last at most how long?

Find the 80th percentile. Draw a graph. Let k = the 80th percentile.



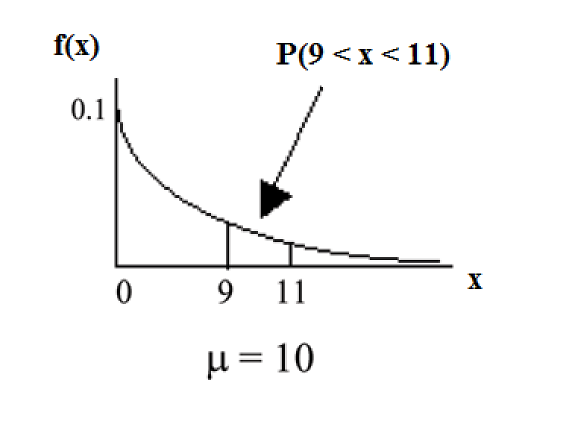
Solve for k: $k = \frac{\ln(1-.80)}{-0.1} = 16.1$ years

Eighty percent of the computer parts last at most 16.1 years.

TI-83+ and TI-84: On the home screen, enter LN(1 - .80)/-.1

What is the probability that a computer part lasts between 9 and 11 years?

Find $P(9<x<11)$. Draw a graph.



$P(9<x<11) = P(x<11) - P(x<9) = \left(1 - e^{-0.1\cdot 11}\right) - \left(1 - e^{-0.1\cdot 9}\right) = 0.6671 - 0.5934 = 0.0737$. (calculator or computer)

The probability that a computer part lasts between 9 and 11 years is 0.0737.

TI-83+ and TI-84: On the home screen, enter e^(-.1*9) - e^(-.1*11).

Example 5.10

Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter = $\frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than 5 minutes. Let X = the length of a phone call, in minutes.

What is m, μ, and σ? The probability that you must wait more than 5 minutes is _______ .

- $m = \frac{1}{12}$
- $\mu = 12$
- $\sigma = 12$

$P(x > 5) = 0.6592$

A summary for exponential distribution is available in "**Summary of The Uniform and Exponential Probability Distributions**".

5.5 Summary of the Uniform and Exponential Probability Distributions

Formula

X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X ; b = largest X

***X* ~ *U*(a,*b*)**

The mean is μ = $\frac{a+b}{2}$

The standard deviation is σ = $\sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f(X) = \frac{1}{b-a}$ for $a \leq X \leq b$

Area to the Left of x: $P(X<x)$ = (base)(height)

Area to the Right of x: $P(X > x)$ = (base)(height)

Area Between c and d: $P(c<X<d)$ = (base)(height) = $(d - c)$(height).

Formula

***X* ~ Exp(*m*)**

X = a real number, 0 or larger. *m* = the parameter that controls the rate of decay or decline

The mean and standard deviation **are the same.**

$\mu = \sigma = \frac{1}{m}$ and $m = \frac{1}{\mu} = \frac{1}{\sigma}$

The probability density function: $f\left(X\right) = m \cdot e^{-m \cdot X}$, $X \geq\ 0$

Area to the Left of x: $P(X<x) = 1 - e^{-m \cdot x}$

Area to the Right of x: $P(X > x) = e^{-m \cdot x}$

Area Between c and d: $P(c<X<d) = P(X<d) - P(X<c) = \left(1 - e^{- m \cdot d}\right) - \left(1 - e^{- m \cdot c}\right) = e^{- m \cdot c} - e^{- m \cdot d}$

Percentile, k: $k = \frac{\text{LN(1-AreaToTheLeft)}}{\text{-m}}$

5.6 Practice 1: Uniform Distribution

Student Learning Outcomes

- The student will analyze data following a uniform distribution.

Given

The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

Describe the Data

Probability Distribution

Random Probability

Quartiles

5.7 Practice 2: Exponential Distribution

Student Learning Outcomes

- The student will analyze data following the exponential distribution.

Given

Carbon-14 is a radioactive element with a half-life of about 5730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121 . We start with 1 gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

Describe the Data

Probability

5.8 Homework

For each probability and percentile problem, DRAW THE PICTURE!

Try these multiple choice problems

The next three questions refer to the following information. The average lifetime of a certain new cell phone is 3 years. The manufacturer will replace any cell phone failing within 2 years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.

The next three questions refer to the following information. The Sky Train from the terminal to the rental car and long term parking center is supposed to arrive every 8 minutes. The waiting times for the train are known to follow a uniform distribution.

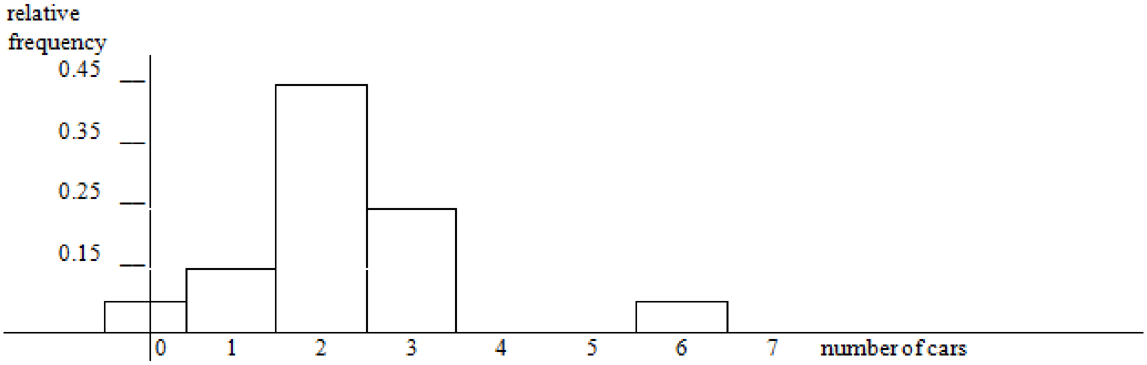
5.9 Review

Exercise 0.0 **–** Exercise 0.0 **refer to the following study:** A recent study of mothers of junior high school children in Santa Clara County reported that 76% of the mothers are employed in paid positions. Of those mothers who are employed, 64% work full-time (over 35 hours per week), and 36% work part-time. However, out of all of the mothers in the population, 49% work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County.

Let $E =$ employed, Let $F =$ full-time employment

Exercise 0.0 **-** Exercise 0.0 **refer to the following:** We randomly pick 10 mothers from the above population. We are interested in the number of the mothers that are employed. Let $X =$ number of mothers that are employed.

Exercise 0.0 **–** Exercise 0.0 **refer to the following:** 64 faculty members were asked the number of cars they owned (including spouse and children's cars). The results are given in the following graph:



Exercise 0.0 **–** Exercise 0.0 **refer to the following study done of the Girls soccer team "Snow Leopards":**

Table 5.2

Hair Style		Hair Color	
	blond	brown	black
ponytail	3	2	5
plain	2	2	1

Suppose that one girl from the Snow Leopards is randomly selected.

5.10 Lab: Continuous Distribution

Class Time:

Names:

Student Learning Outcomes:

- The student will compare and contrast empirical data from a random number generator with the Uniform Distribution.

Collect the Data

Use a random number generator to generate 50 values between 0 and 1 (inclusive). List them below. Round the numbers to 4 decimal places or set the calculator MODE to 4 places.

1. Complete the table:

Table 5.3

2. Calculate the following:

a. $\overline{X}$ =

b. S =

c. 1st quartile =

d. 3rd quartile =

e. Median =

Organize the Data

1. Construct a histogram of the empirical data. Make 8 bars.

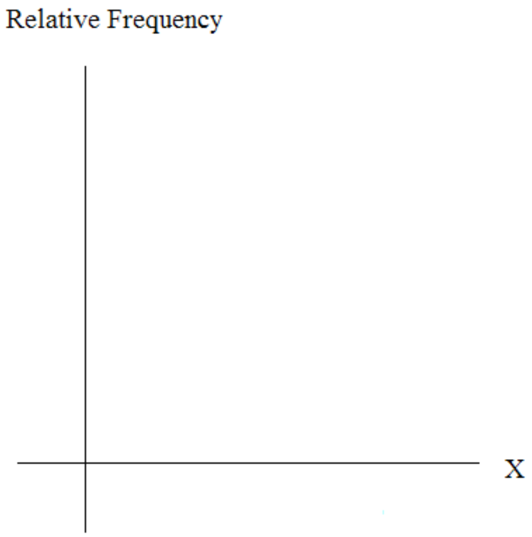


Figure 5.4

2. Construct a histogram of the empirical data. Make 5 bars.

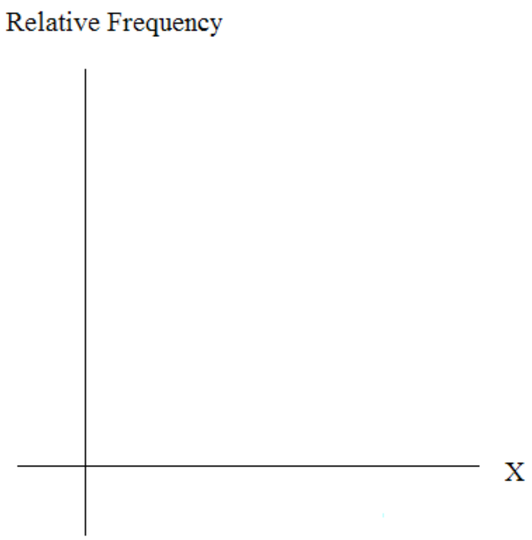


Figure 5.5

Describe the Data

1. Describe the shape of each graph. Use 2 – 3 complete sentences. (Keep it simple. Does the graph go straight across, does it have a V shape, does it have a hump in the middle or at either end, etc.? One way to help you determine a shape, is to roughly draw a smooth curve through the top of the bars.)
2. Describe how changing the number of bars might change the shape.

Theoretical Distribution

1. In words, *X* =
2. The theoretical distribution of *X* is *X* ~ *U*(0, 1). Use it for this part.
3. In theory, based upon the distribution *X* ~ *U*(0, 1), complete the following.
   a. μ=
   b. σ =
   c. 1st quartile =
   d. 3rd quartile =
   e. median = __________
4. Are the empirical values (the data) in the section titled "Collect the Data" close to the corresponding theoretical values above? Why or why not?

Plot the Data

1. Construct a box plot of the data. Be sure to use a ruler to scale accurately and draw straight edges.
2. Do you notice any potential outliers? If so, which values are they? Either way, numerically justify your answer. (Recall that any DATA are less than Q1 – 1.5*IQR or more than Q3 + 1.5*IQR are potential outliers. IQR means interquartile range.)

Compare the Data

1. For each part below, use a complete sentence to comment on how the value obtained from the data compares to the theoretical value you expected from the distribution in the section titled "Theoretical Distribution."
   a. minimum value:
   b. 1st quartile:
   c. median:
   d. third quartile:
   e. maximum value:
   f. width of IQR:
   g. overall shape:
2. Based on your comments in the section titled "Collect the Data", how does the box plot fit or not fit what you would expect of the distribution in the section titled "Theoretical Distribution?"

Discussion Question

1. Suppose that the number of values generated was 500, not 50. How would that affect what you would expect the empirical data to be and the shape of its graph to look like?

Glossary

Conditional Probability: The likelihood that an event will occur given that another event has already occurred.

Exponential Distribution: A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Exponential Distribution: A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Uniform Distribution: A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a,b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(X) = \frac{1}{b-a}$ for $a<x<b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Uniform Distribution: A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a,b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(x) = \frac{1}{b-a}$ for $a<x<b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

6 THE NORMAL DISTRIBUTION

6.1 The Normal Distribution

Student Learning Outcomes

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

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

Introduction

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real estate prices fit a normal distribution. The normal distribution is extremely important but it cannot be applied to everything in the real world.

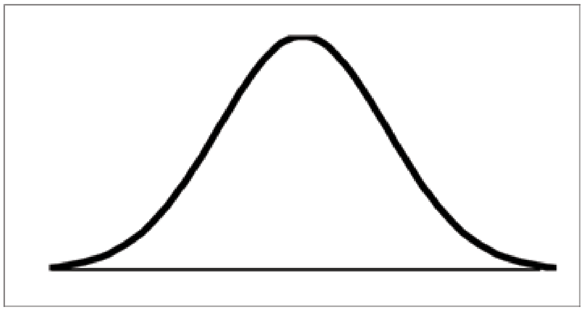
In this chapter, you will study the normal distribution, the standard normal, and applications associated with them.

Optional Collaborative Classroom Activity

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the x-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If *X* is a quantity to be measured that has a normal distribution with mean (μ) and the standard deviation (σ), we designate this by writing

NORMAL:*X*~N(μ, σ)



The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

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

The cumulative distribution function is ***P*(*X*<*x*)** . It is calculated either by a calculator or a computer or it is looked up in a table. Technology has made the tables basically obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions in this chapter. See the NOTE in this chapter in **Calculation of Probabilities**.

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median since the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

6.2 The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z-scores**. **A z-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is 5 and the standard deviation is 2, the value 11 is 3 standard deviations above (or to the right of) the mean. The calculation is:

$$x = \mu + (z)\sigma = 5 + (3)(2) = 11 \tag{6.1}$$

The z-score is 3.

The mean for the standard normal distribution is 0 and the standard deviation is 1. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution ***Z~ N*(0, 1)** . The value *x* comes from a normal distribution with mean μ and standard deviation σ.

6.3 Z-scores

If *X* is a normally distributed random variable and *X*~N(μ, σ), then the z-score is:

$$z = \frac{x - \mu}{\sigma} \tag{6.2}$$

The z-score tells you how many standard deviations that the value *x* is above (to the right of) or below (to the left of) the mean, μ. Values of *x* that are larger than the mean have positive z-scores and values of *x* that are smaller than the mean have negative z-scores. If *x* equals the mean, then *x* has a z-score of 0.

Example 6.1

Suppose *X* ~ N(5, 6). This says that *X* is a normally distributed random variable with mean μ = 5 and standard deviation σ = 6. Suppose x = 17. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2 \tag{6.3}$$

This means that x = 17 is **2 standard deviations** (2σ) above or to the right of the mean μ = 5. The standard deviation is σ = 6.

Notice that:

$$5 + 2 \cdot 6 = 17 \qquad \text{(The pattern is } \mu + z\sigma = x\text{.)} \tag{6.4}$$

Now suppose x=1. Then:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67 \qquad \text{(rounded to two decimal places)} \tag{6.5}$$

This means that x = 1 is 0.67 standard deviations (- 0.67σ) below or to the left of the mean μ = 5. Notice that:

5 + (-0.67)(6) is approximately equal to 1 (This has the pattern μ + (-0.67)σ = 1)

Summarizing, when *z* is positive, *x* is above or to the right of μ and when *z* is negative, *x* is to the left of or below μ.

Example 6.2

Some doctors believe that a person can lose 5 pounds, on the average, in a month by reducing his/her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let *X* = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of 2 pounds. *X*~N(5, 2). Fill in the blanks.

Suppose a person **lost** 10 pounds in a month. The z-score when x = 10 pounds is z = 2.5 (verify). This z-score tells you that x = 10 is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).

This z-score tells you that x = 10 is **2.5** standard deviations to the **right** of the mean **5**.

Suppose a person **gained** 3 pounds (a negative weight loss). Then *z* = __________. This z-score tells you that x = -3 is ________ standard deviations to the __________ (right or left) of the mean.

z = **-4**. This z-score tells you that x = -3 is **4** standard deviations to the **left** of the mean.

Suppose the random variables *X* and *Y* have the following normal distributions: *X* ~N(5, 6) and Y ~ N(2, 1). If x = 17, then *z* = 2. (This was previously shown.) If y = 4, what is *z*?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2 \qquad \text{where } \mu\text{=2 and } \sigma\text{=1.} \tag{6.6}$$

The z-score for y = 4 is z = 2. This means that 4 is z = 2 standard deviations to the right of the mean. Therefore, x = 17 and y = 4 are both 2 (of **their**) standard deviations to the right of **their** respective means.

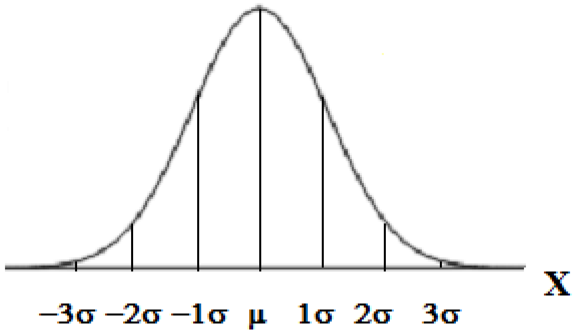
The z-score allows us to compare data that are scaled differently. To understand the concept, suppose *X* ~N(5, 6) represents weight gains for one group of people who are trying to gain weight in a 6 week period and *Y* ~N(2, 1) measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since x = 17 and y = 4 are each 2 standard deviations to the right of their means, they represent the same weight gain **relative to their means**.

The Empirical Rule

If *X* is a random variable and has a normal distribution with mean µ and standard deviation σ then the **Empirical Rule** says (See the figure below)

- About 68.27% of the *x* values lie between -1σ and +1σ of the mean µ (within 1 standard deviation of the mean).

- About 95.45% of the *x* values lie between -2σ and +2σ of the mean μ (within 2 standard deviations of the mean).
- About 99.73% of the *x* values lie between -3σ and +3σ of the mean μ (within 3 standard deviations of the mean). Notice that almost all the *x* values lie within 3 standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and -1, respectively.
- The z-scores for +2σ and –2σ are +2 and -2, respectively.
- The z-scores for +3σ and –3σ are +3 and -3 respectively.



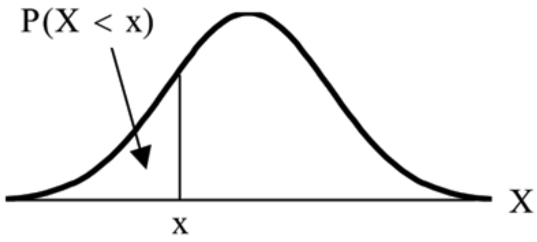
Example 6.3

Suppose *X* has a normal distribution with mean 50 and standard deviation 6.

- About 68.27% of the *x* values lie between -1σ = (-1)(6) = -6 and 1σ = (1)(6) = 6. The values -6 and 6 are within 1 standard deviation of the mean 50. The z-scores are -1 and +1 for -6 and 6, respectively.
- About 95.45% of the *x* values lie between -2σ = (-2)(6) = -12 and 2σ = (2)(6) = 12. The values -12 and 12 are within 2 standard deviations of the mean 50. The z-scores are -2 and +2 for -12 and 12, respectively.
- About 99.73% of the *x* values lie between -3σ = (-3)(6) = -18 and 3σ = (3)(6) = 18. The values -18 and 18 are within 3 standard deviations of the mean 50. The z-scores are -3 and +3 for -18 and 18, respectively.

6.4 Areas to the Left and Right of x

The arrow in the graph below points to the area to the left of *x*. This area is represented by the probability ***P*(*X*<*x*).** Normal tables, computers, and calculators provide or calculate the probability *P*(*X*<*x*).



The area to the right is then *P*(*X* > *x*) = 1 − *P*(*X*<*x*).

Remember, *P*(*X*<*x*) = **Area to the left** of the vertical line through *x*.

P(*X* > *x*) = 1 − *P*(*X*<*x*) = . **Area to the right** of the vertical line through *x*

P(*X*<*x*) is the same as *P*(*X* ≤ *x*) and *P*(*X* > *x*) is the same as *P*(*X* ≥ *x*) for continuous distributions.

6.5 Calculations of Probabilities

Probabilities are calculated by using technology. There are instructions in the chapter for the TI-83+ and TI-84 calculators.

In the Table of Contents for **Collaborative Statistics**, entry **15. Tables** has a link to a table of normal probabilities. Use the probability tables if so desired, instead of a calculator. The tables include instructions for how to use then.

Example 6.4

If the area to the left is 0.0228, then the area to the right is 1 − 0.0228 = 0.9772.

Example 6.5

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

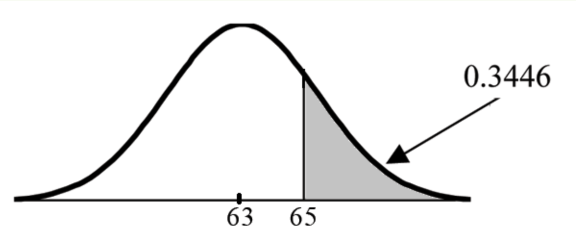
Find the probability that a randomly selected student scored more than 65 on the exam.

Let X = a score on the final exam. $X \sim N(63, 5)$, where μ = 63 and σ = 5

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$ (calculator or computer)



The probability that one student scores more than 65 is 0.3446.

Using the TI-83+ or the TI-84 calculators, the calculation is as follows. Go into `2nd DISTR`.

After pressing `2nd DISTR`, press `2:normalcdf`.

The syntax for the instructions are shown below.

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1`, the `EE` key (a 2nd key) and then `99`. Or, you can enter `10^99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be -1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Historical Note

The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$z = \frac{65-63}{5} = 0.4$. Area to the left is 0.6554. $P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$

Find the probability that a randomly selected student scored less than 85.

Draw a graph.

Then find $P(x<85)$. Shade the graph. $P(x<85) = 1$ (calculator or computer)

The probability that one student scores less than 85 is approximately 1 (or 100%).

The TI-instructions and answer are as follows:

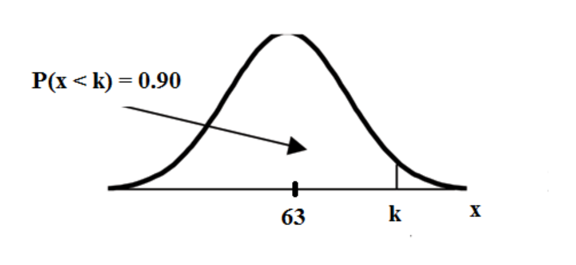
normalcdf(0,85,63,5) = 1 (rounds to 1)

Find the 90th percentile (that is, find the score k that has 90 % of the scores below k and 10% of the scores above k).

Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let *k* = the 90th percentile. k is located on the x-axis. $P(x<k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k and 10% are the same or higher. ***k*** is often called a **critical value**.

$k = 69.4$ (calculator or computer)



The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. For the TI-83+ or TI-84 calculators, use `invNorm` in `2nd DISTR`. invNorm(area to the left, mean, standard deviation) For this problem, invNorm(0.90,63,5) = 69.4

Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Find the 70th percentile.

Draw a new graph and label it appropriately. *k* = 65.6

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

invNorm(0.70,63,5) = 65.6

Example 6.6

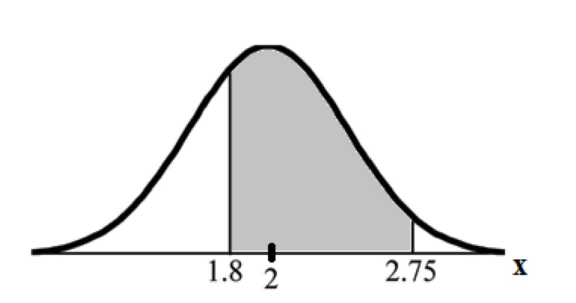
A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Find the probability that a household personal computer is used between 1.8 and 2.75 hours per day.

Let *X* = the amount of time (in hours) a household personal computer is used for entertainment. $x \sim N(2, 0.5)$ where μ = 2 and σ = 0.5.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** *x* = 1.8 and *x* = 2.75. $P(1.8<x<2.75) = 0.5886$

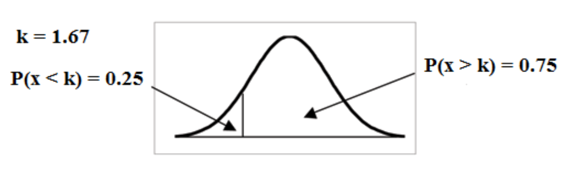


normalcdf(1.8,2.75,2,0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Find the maximum number of hours per day that the bottom quartile of households use a personal computer for entertainment.

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile, *k***, where $P(x<k) = 0.25$.



invNorm(0.25,2,.5) = 1.66

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

6.6 Summary of Formulas

Formula

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Formula

$Z \sim N(0, 1)$

z = a standardized value (z-score)

mean = 0 standard deviation = 1

Formula

To find the **kth** percentile when the z-score is known: $\boldsymbol{k = \mu + (z)\sigma}$

Formula

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

Formula

The area to the left: $P(X<x)$

Formula

The area to the right: $P(X > x) = 1 - P(X<x)$

6.7 Practice: The Normal Distribution

Student Learning Outcomes

- The student will analyze data following a normal distribution.

Given

The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for 3 years. We are interested in the length of time a CD player lasts.

Normal Distribution

6.8 Homework

Try These Multiple Choice Questions

The questions below refer to the following: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

The questions below refer to the following: The length of time to find a parking space at 9 A.M. follows a normal distribution with a mean of 5 minutes and a standard deviation of 2 minutes.

6.9 Review

The next two questions refer to: $X \sim U(3, 13)$

6.10 Lab 1: Normal Distribution (Lap Times)

Class Time:

Names:

Student Learning Outcome:

- The student will compare and contrast empirical data and a theoretical distribution to determine if Terry Vogel's lap times fit a continuous distribution.

Directions:

Round the relative frequencies and probabilities to 4 decimal places. Carry all other decimal answers to 2 places.

Collect the Data

1. Use the data from **Terri Vogel's Log Book**. Use a Stratified Sampling Method by Lap (Races 1 – 20) and a random number generator to pick 6 lap times from each stratum. Record the lap times below for Laps 2 – 7.

Table 6.1

________	________	________	________	________	________
________	________	________	________	________	________
________	________	________	________	________	________
________	________	________	________	________	________
________	________	________	________	________	________
________	________	________	________	________	________

2. Construct a histogram. Make 5 - 6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

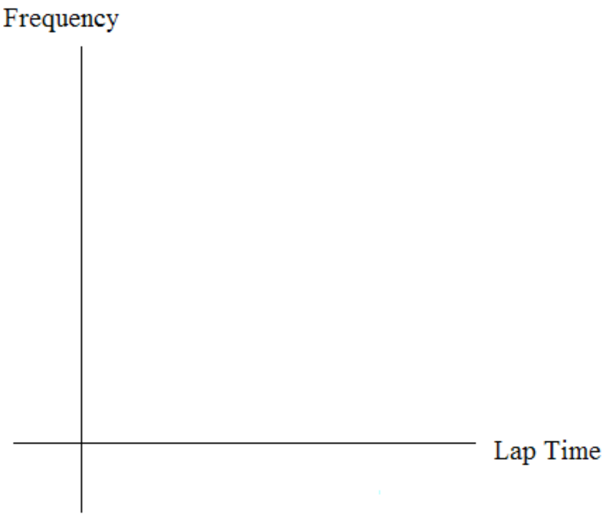


Figure 6.1

3. Calculate the following.

a. $\bar{x}$ =

b. *s* =

4. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a V-shape, does it have a hump in the middle or at either end, etc.?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram to help, what was the approximate theoretical distribution of the data?

- *X* ~
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Use the Data from the section titled "Collect the Data" to complete the following statements.

- The IQR goes from __________ to __________.
- IQR = __________. (IQR=Q3-Q1)
- The 15th percentile is:
- The 85th percentile is:
- The median is:
- The empirical probability that a randomly chosen lap time is more than 130 seconds =
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution from the section titled "Analyse the Distribution" complete the following statements:

- The IQR goes from __________ to __________.
- IQR =
- The 15th percentile is:
- The 85th percentile is:
- The median is:
- The probability that a randomly chosen lap time is more than 130 seconds =
- Explain the meaning of the 85th percentile of this distribution.

Discussion Questions

- Do the data from the section titled "Collect the Data" give a close approximation to the theoretical distibution in the section titled "Analyze the Distribution"? In complete sentences and comparing the result in the sections titled "Describe the Data" and "Theoretical Distribution", explain why or why not.

6.11 Lab 2: Normal Distribution (Pinkie Length)

Class Time:

Names:

Student Learning Outcomes:

- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

Collect the Data

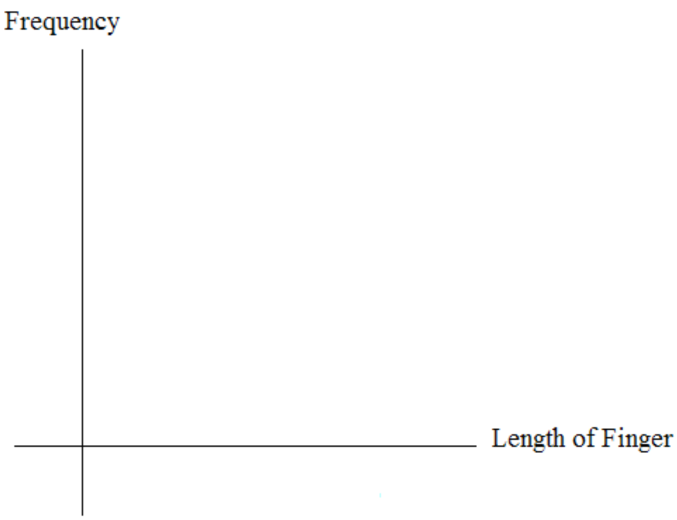
Measure the length of your pinkie finger (in cm.)

1. Randomly survey 30 adults. Round to the nearest 0.5 cm.

Table 6.2

________	________	________	________	________
________	________	________	________	________
________	________	________	________	________
________	________	________	________	________
________	________	________	________	________
________	________	________	________	________

2. Construct a histogram. Make 5-6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the Following
   - **a.** $\bar{x}$ =
   - **b.** s =
4. Draw a smooth curve through the top of the bars of the histogram. Use 1-2 complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a V-shape, does it have a hump in the middle or at either end, etc.?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram to help, what was the approximate theoretical distribution of the data from the section titled "Collect the Data"?

- X ~
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Using the data in the section titled "Collect the Data" complete the following statements. (Hint: order the data)

> Remember
>
> (IQR = $Q3 - Q1$)

- IQR =
- 15th percentile is:
- 85th percentile is:
- Median is:
- What is the empirical probability that a randomly chosen pinkie length is more than 6.5 cm?
- Explain the meaning the 85th percentile of this data.

Theoretical Distribution

Using the Theoretical Distribution in the section titled "Analyze the Distribution"

- IQR =
- 15th percentile is:
- 85th percentile is:
- Median is:
- What is the theoretical probability that a randomly chosen pinkie length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Discussion Questions

- Do the data from the section entitled "Collect the Data" give a close approximation to the theoretical distribution in "Analyze the Distribution." In complete sentences and comparing the results in the sections titled "Describe the Data" and "Theoretical Distribution", explain why or why not.

Glossary

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Standard Normal Distribution: A continuous random variable (RV) $X \sim N(0,1)$.. When X follows the standard normal distribution, it is often noted as $Z \sim N(0,1)$.

z-score: The linear transformation of the form $z = \frac{x - \mu}{\sigma}$. If this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$, the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ , the result is called the z-score of x. Z-scores allow us to compare data that are normally distributed but scaled differently.

7 THE CENTRAL LIMIT THEOREM

7.1 The Central Limit Theorem

Student Learning Outcomes

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

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Averages.
- Apply and interpret the Central Limit Theorem for Sums.

Introduction

What does it mean to be average? Why are we so concerned with averages? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study averages and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size *n* from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size *n* and *n* is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means (averages) and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, *n*, that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more observations are needed for the sample averages or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times. (The 8, 7, 9, and 11 were randomly chosen.)

Each time a person rolls more than one die, he/she calculates the **average** of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The average is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one average when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more averages for a total of 10 averages.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the average. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "average" when you roll one die, is just the face on the die, what distribution do these "averages" appear to be representing?

Draw the graph for the averages using 2 dice. Do the averages show any kind of pattern?

Draw the graph for the averages using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the averages using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The average of the averages remains approximately the same.
2. The spread of the averages (the standard deviation of the averages) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means (averages) tend toward a normal distribution (the sampling distribution).**

7.2 The Central Limit Theorem for Sample Means (Averages)

Suppose *X* is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. μ_X = the mean of *X*

b. σ_X = the standard deviation of *X*

If you draw random samples of size *n*, then as *n* increases, the random variable $\overline{X}$ which consists of sample means, tends to be **normally distributed** and

$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$

The Central Limit Theorem for Sample Means (Averages) says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means (averages) form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by *n*, the sample size. *n* is the number of values that are averaged together not the number of times the experiment is done.

The random variable $\overline{X}$ has a different z-score associated with it than the random variable *X*. $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \tag{7.1}$$

μ_X is both the average of *X* and of $\overline{X}$.

$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} =$ standard deviation of $\overline{X}$ and is called the **standard error of the mean.**

Example 7.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size *n* = 25 are drawn randomly from the population.

Find the probability that the **sample mean** is between 85 and 92.

Let *X* = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean (or average)**.

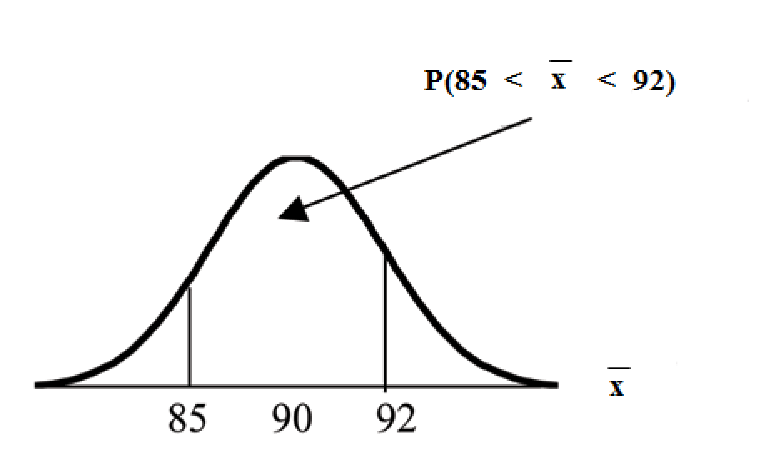
Let $\overline{X}$ = the mean or average of a sample of size 25. Since μ_X = 90, σ_X = 15, and *n* = 25;

then $\overline{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$

Find $P(85<\overline{x}<92)$ Draw a graph.

$P(85<\overline{x}<92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.



TI-83 or 84: `normalcdf`(lower value, upper value, mean for averages, `stdev` for averages)

`stdev` = standard deviation

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

`normalcdf`(85,92,90,$\frac{15}{\sqrt{25}}$) = 0.6997

Find the average value that is 2 standard deviations above the the mean of the averages.

To find the average value that is 2 standard deviations above the mean of the averages, use the formula

$$\text{value} = \mu_X + (\#\text{ofSTDEVs})\left(\frac{\sigma_X}{\sqrt{n}}\right)$$

$$\text{value} = 90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$$

So, the average value that is 2 standard deviations above the mean of the averages is 96.

Example 7.2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size *n* = 50** is drawn randomly from the population.

Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean or average time, in hours**, it takes to play one soccer match.

Let $\overline{X}$ = the **average** time, in hours, it takes to play one soccer match.

If μ_X = __________, σ_X = ___________, and n = ____________, then $\overline{X} \sim N$(______, ______) by the Central Limit Theorem for Averages of Sample Means.

$\mu_X =$ **2**, $\sigma_X =$ **0.5**, $n =$ **50**, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8<\overline{x}<2.3)$. Draw a graph.

$P(1.8<\overline{x}<2.3) = 0.9977$

`normalcdf`$\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the sample mean is between 1.8 hours and 2.3 hours is ______.

7.3 The Central Limit Theorem for Sums

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

a. $\mu_X =$ the mean of X

b. $\sigma_X =$ the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

$$\Sigma X \sim N\left(n \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right)$$

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution). **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.

b. $z = \dfrac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$

a. $n \cdot \mu_X =$ the mean of ΣX

b. $\sqrt{n} \cdot \sigma_X =$ standard deviation of ΣX

Example 7.3

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

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
b. Find the sum that is 1.5 standard deviations below the mean of the sums.

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

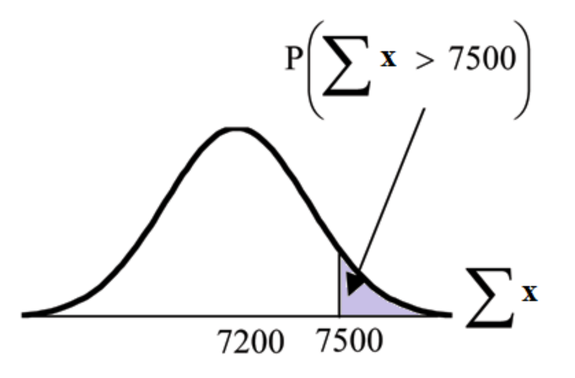
ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then

$\Sigma X \sim N(80 \cdot 90, \sqrt{80} \cdot 15)$

a. mean of the sums = $n \cdot \mu_X = (80)(90) = 7200$
b. standard deviation of the sums = $\sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
c. sum of 80 values = Σx = 7500

Find $P(\Sigma x > 7500)$ Draw a graph.

$P(\Sigma x > 7500) = 0.0127$



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

The parameter list is abbreviated (lower, upper, $n \cdot \mu_X$, $\sqrt{n} \cdot \sigma_X$)

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

Reminder: 1E99 = 10^{99}. Press the EE key for E.

7.4 Using the Central Limit Theorem

It is important for you to understand when to use the **CLT**. If you are being asked to find the probability of an average or mean, use the CLT for means or averages. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for averages and sums.

If you are being asked to find the probability of an **individual** value, do **not** use the CLT. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample gets closer and closer to μ. From the Central Limit Theorem, we know that as *n* gets larger and larger, the sample averages follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ must be close to the population mean μ. We can say that μ is the value that the sample averages approach as *n* gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean (Average) and Sum Examples

Example 7.4

A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the **average stress score** for the 75 students is less than 2.

2. The 90th percentile for the **average stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.

Problems 1. and 2. ask you to find a probability or a percentile for an **average** or **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$ (See **Continuous Random Variables** for the uniform).

$$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$$

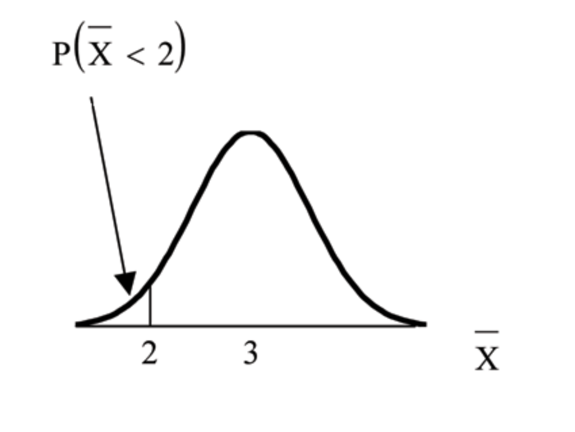
For problems 1. and 2., let $\overline{X}$ = the average stress score for the 75 students. Then,

$\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where n = 75.

Find $P\left(\overline{X}<2\right)$. Draw the graph.

$P\left(\overline{X}<2\right) = 0$

The probability that the average stress score is less than 2 is about 0.



$\texttt{normalcdf}\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

Reminder

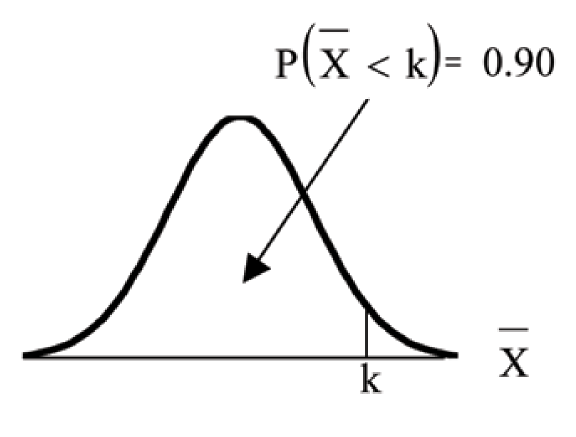
The smallest stress score is 1. Therefore, the smallest average for 75 stress scores is 1.

Find the 90th percentile for the average of 75 stress scores. Draw a graph.

Let k = the 90th precentile.

Find k where $P\left(\overline{X}<k\right) = 0.90$.

$k = 3.2$



The 90th percentile for the average of 75 scores is about 3.2. This means that 90% of all the averages of 75 stress scores are at most 3.2 and 10% are at least 3.2.

`invNorm` $\left(.90, 3, \frac{1.15}{\sqrt{75}}\right) = 3.2$

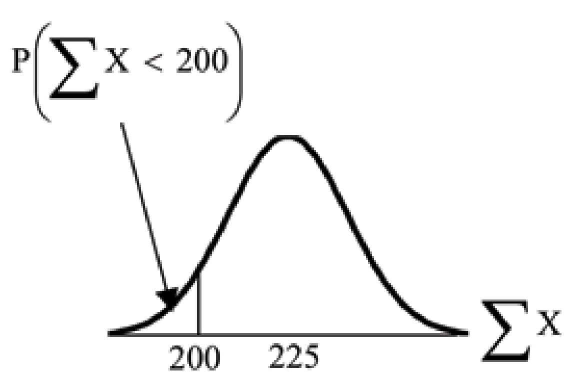
For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N[(75) \cdot (3), \sqrt{75} \cdot 1.15]$

Find $P(\Sigma X<200)$. Draw the graph.

The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75} \cdot 1.15 = 9.96$

$P(\Sigma X<200) = 0$



The probability that the total of 75 scores is less than 200 is about 0.

`normalcdf` $(75, 200, 75 \cdot 3, \sqrt{75} \cdot 1.15) = 0.$

Reminder

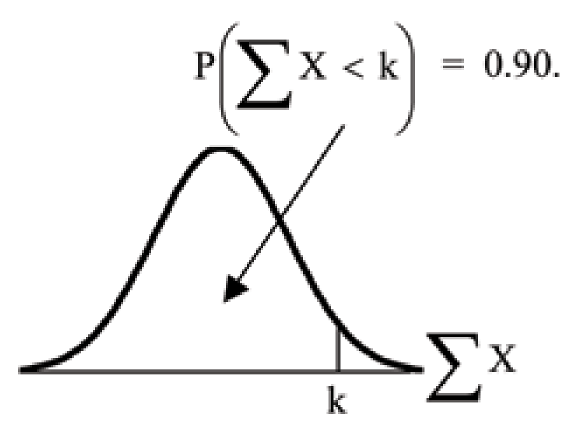
The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Let k = the 90th percentile.

Find k where $P(\Sigma X<k) = 0.90$.

$k = 237.8$



The 90th percentile for the sum of 75 scores is about 237.8. This means that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

`invNorm` (.90, 75 · 3, $\sqrt{75}$ · 1.15) = 237.8

Example 7.5

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim \text{Exp}\left(\frac{1}{22}\right)$ From Chapter 5, we know that μ = 22 and σ = 22.

Let $\bar{X}$ = the AVERAGE excess time used by a sample of n = 80 customers who exceed their contracted time allowance.

$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means or Averages

a. Find the probability that the average excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P\left(\bar{X} > 20\right)$ Draw the graph.

b. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(X > 20)$

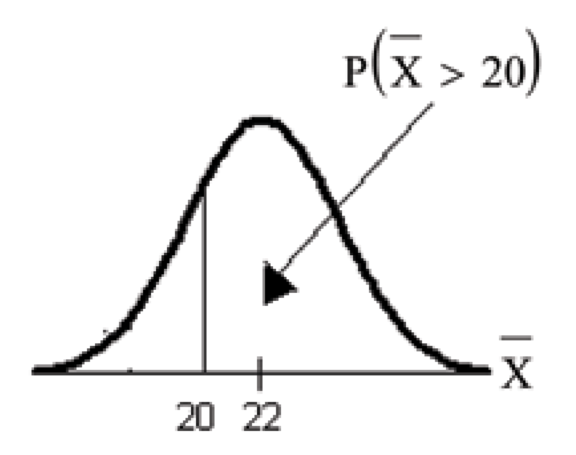
c. Explain why the probabilities in (a) and (b) are different.

Part a.

Find: $P\left(\bar{X} > 20\right)$

$P\left(\bar{X} > 20\right) = 0.7919$ using `normalcdf` $\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the average excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



Reminder

1E99 = $10^{99}$and-1E99 = -10^{99}. Press the `EE` key for E. Or just use 10^99 instead of 1E99.

Part b.

Find P(X>20) . Remember to use the exponential distribution for an **individual: X~Exp(1/22)**.

P(X>20) = e^(–(1/22)*20) or e^(–.04545*20) = 0.4029

$P(X > 20) = 0.4029$ but $P\left(\overline{X} > 20\right) = 0.7919$

The probabilities are not equal because we use different distributions to calculate the probability for individuals and for averages.

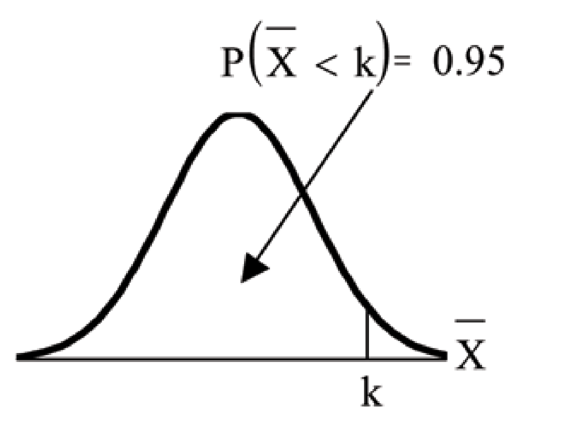
When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the CLT. Use the CLT with the normal distribution when you are being asked to find the probability for an average.

Using the CLT to find Percentiles:

Find the 95th percentile for the **sample average excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Let k = the 95th percentile. Find k where $P\left(\overline{X}<k\right) = 0.95$

k = 26.0 using `invNorm`$\left(.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$



The 95th percentile for the **sample average excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

95% of such samples would have averages under 26 minutes; only 5% of such samples would have averages above 26 minutes.

(HISTORICAL): Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for *n* (say, 20). To calculate the probabilities with large values of *n*, you had to use the binomial formula which could be very complicated. Using the **Normal Approximation to the Binomial** simplified the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number *n* of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success *p*

Recall that if X is the binomial random variable, then $\boldsymbol{X \sim B(n, p)}$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to X or subtract 0.5 from X (use $X + 0.5$ or $X - 0.5$). The number 0.5 is called the **continuity correction factor**.

Example 7.6

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K - 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let X= the number that favor a charter school for grades K - 5. $\boldsymbol{X \sim B(n, p)}$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N(159, 8.6447)$. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5)=0.8641$.

`normalcdf` (149.5, 10^99, 159, 8.6447) = 0.8641.

For Problem 2., you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5)=0.5689$.

`normalcdf` (0, 160.5, 159, 8.6447) = 0.5689

For Problem 3., you **exclude 155** so $P(X > 155)$ has normal approximation $P(Y > 155.5)=0.6572$.

`normalcdf` (155.5, 10^99, 159, 8.6447) = 0.6572

For Problem 4., you **exclude 147** so $P(X<147)$ has normal approximation $P(Y<146.5)=0.0741$.

`normalcdf` (0, 146.5, 159, 8.6447) = 0.0741

For Problem 5., $P(X=175)$ has normal approximation $P(174.5 < Y < 175.5)=0.0083$.

`normalcdf` (174.5, 175.5, 159, 8.6447) = 0.0083

Because of calculators and computer software that easily let you calculate binomial probabilities for large values of *n*, it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial. In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.

For **Example 3**, the probabilities are calculated using the binomial (n=300 and p=0.53) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.

$P(X \geq 150)$: `1 - binomialcdf` (300, 0.53, 149)=0.8641

$P(X \leq 160)$: `binomialcdf` (300, 0.53, 160)=0.5684

$P(X > 155)$: `1 - binomialcdf` (300, 0.53, 155)=0.6576

$P(X<147)$: `binomialcdf` (300, 0.53, 146)=0.0742

$P(X=175)$: (You use the binomial pdf.) `binomialpdf` (175, 0.53, 146)=0.0083

**Contributions made to Example 2 by Roberta Bloom

7.5 Summary of Formulas

Formula

$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ **The Mean $\left(\overline{X}\right)$:** μ_X

Formula

$z = \dfrac{\bar{x} - \mu_X}{\left(\dfrac{\sigma_X}{\sqrt{n}}\right)}$ **Standard Error of the Mean (Standard Deviation $\left(\bar{X}\right)$):** $\dfrac{\sigma_X}{\sqrt{n}}$

Formula

$\Sigma X \sim N\left[(n) \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right]$ **Mean for Sums (ΣX):** $n \cdot \mu_X$

Formula

$z = \dfrac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$ **Standard Deviation for Sums (ΣX):** $\sqrt{n} \cdot \sigma_X$

7.6 Practice: The Central Limit Theorem

Student Learning Outcomes

- The student will calculate probabilities using the Central Limit Theorem.

Given

Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately 4 hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\bar{X}$ be the random variable representing the average time to complete the 16 reviews. Let ΣX be the total time it takes Yoonie to complete all of the month's reviews.

Distribution

Complete the distributions.

1. $X \sim$
2. $\bar{X} \sim$
3. $\Sigma X \sim$

Graphing Probability

For each problem below:

a. Sketch the graph. Label and scale the horizontal axis. Shade the region corresponding to the probability.
b. Calculate the value.

Discussion Question

7.7 Homework

Try these multiple choice questions (Exercises19 - 23).

The next two questions refer to the following information: The time to wait for a particular rural bus is distributed uniformly from 0 to 75 minutes. 100 riders are randomly sampled to learn how long they waited.

The next three questions refer to the following information: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

**Exercise 24 contributed by Roberta Bloom

7.8 Review

The next three questions refer to the following information: Richard’s Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

**Exercise 9 contributed by Roberta Bloom

7.9 Lab 1: Central Limit Theorem (Pocket Change)

Class Time:

Names:

Student Learning Outcomes:

- The student will demonstrate and compare properties of the Central Limit Theorem.

This lab works best when sampling from several classes and combining data.

Collect the Data

1. Count the change in your pocket. (Do not include bills.)
2. Randomly survey 30 classmates. Record the values of the change.

Table 7.1

__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________

3. Construct a histogram. Make 5 - 6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

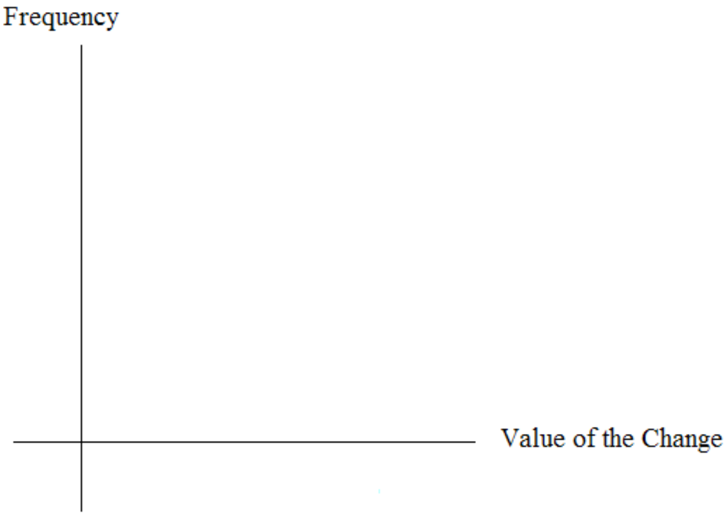


Figure 7.1

4. Calculate the following (*n* = 1; surveying one person at a time):

   a. $\overline{x}$ =

   b. *s* =

5. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Collecting Averages of Pairs

Repeat steps 1 - 5 (of the section above titled "Collect the Data") with one exception. Instead of recording the change of 30 classmates, record the average change of 30 pairs.

1. Randomly survey 30 **pairs** of classmates. Record the values of the average of their change.

Table 7.2

____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____

2. Construct a histogram. Scale the axes using the same scaling you did for the section titled "Collecting the Data". Sketch the graph using a ruler and a pencil.

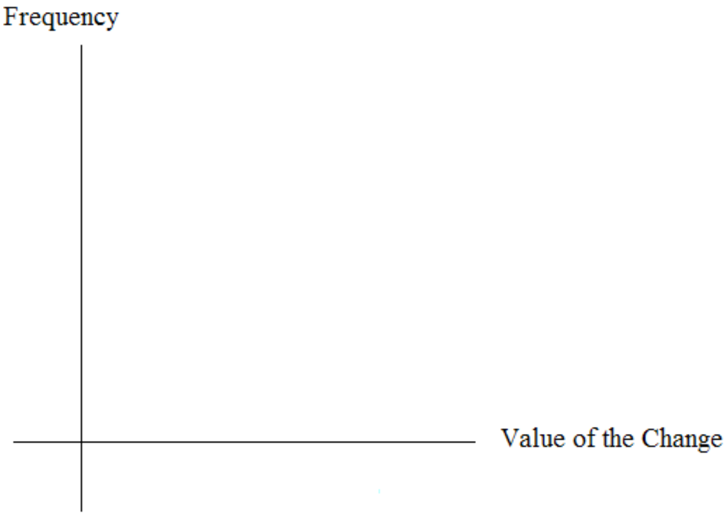


Figure 7.2

3. Calculate the following (*n* = 2; surveying two people at a time):

   a. $\bar{x}$ =

   b. *s* =

4. Draw a smooth curve through tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Collecting Averages of Groups of Five

Repeat steps 1 – 5 (of the section titled "Collect the Data") with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of 5.

1. Randomly survey 30 **groups of 5** classmates. Record the values of the average of their change.

Table 7.3

2. Construct a histogram. Scale the axes using the same scaling you did for the section titled "Collect the Data". Sketch the graph using a ruler and a pencil.

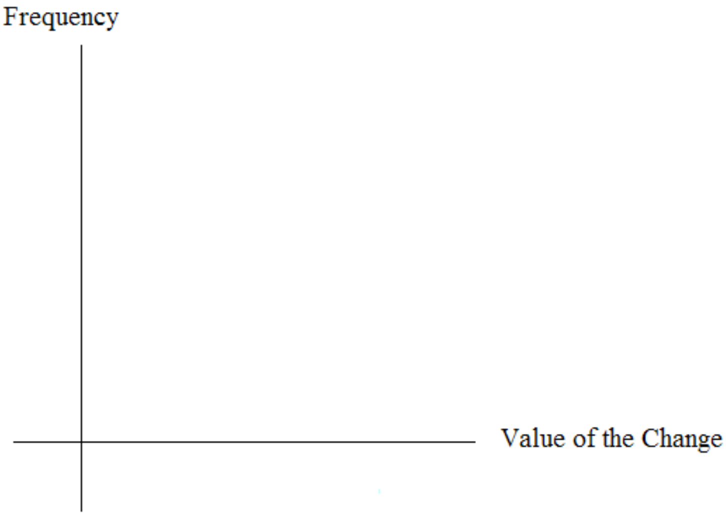


Figure 7.3

3. Calculate the following (n = 5; surveying five people at a time):

   a. $\bar{x}$ =

   b. s =

4. Draw a smooth curve through tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Discussion Questions

1. As n changed, why did the shape of the distribution of the data change? Use 1 – 2 complete sentences to explain what happened.
2. In the section titled "Collect the Data", what was the approximate distribution of the data? X ~
3. In the section titled "Collecting Averages of Groups of Five", what was the approximate distribution of the averages? $\bar{X}$ ~
4. In 1 – 2 complete sentences, explain any differences in your answers to the previous two questions.

7.10 Lab 2: Central Limit Theorem (Cookie Recipes)

Class Time:

Names:

Student Learning Outcomes:

- The student will demonstrate and compare properties of the Central Limit Theorem.

Given:

X = length of time (in days) that a cookie recipe lasted at the Olmstead Homestead. (Assume that each of the different recipes makes the same quantity of cookies.)

Table 7.4

Recipe #	X		Recipe #	X		Recipe #	X		Recipe #	X
1	1		16	2		31	3		46	2
2	5		17	2		32	4		47	2
3	2		18	4		33	5		48	11
4	5		19	6		34	6		49	5
5	6		20	1		35	6		50	5
6	1		21	6		36	1		51	4
7	2		22	5		37	1		52	6
8	6		23	2		38	2		53	5
9	5		24	5		39	1		54	1
10	2		25	1		40	6		55	1
11	5		26	6		41	1		56	2
12	1		27	4		42	6		57	4
13	1		28	1		43	2		58	3
14	3		29	6		44	6		59	6
15	2		30	2		45	2		60	5

Calculate the following:

a. μ_X =

b. σ_X =

Collect the Data

Use a random number generator to randomly select 4 samples of size *n* = 5 from the given population. Record your samples below. Then, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

1. Complete the table:

Table 7.5

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means:	$\bar{x}$ =	$\bar{x}$ =	$\bar{x}$ =	$\bar{x}$ =	

2. Calculate the following:

   a. $\bar{X}$ =

   b. $s_{\bar{x}}$ =

3. Again, use a random number generator to randomly select 4 samples from the population. This time, make the samples of size $n = 10$. Record the samples below. As before, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

Table 7.6

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means:	$\overline{x}$ =	$\overline{x}$ =	$\overline{x}$ =	$\overline{x}$ =	

4. Calculate the following:

   a. $\overline{x}$ =

   b. $s_{\overline{x}}$ =

5. For the original population, construct a histogram. Make intervals with bar width = 1 day. Sketch the graph using a ruler and pencil. Scale the axes.

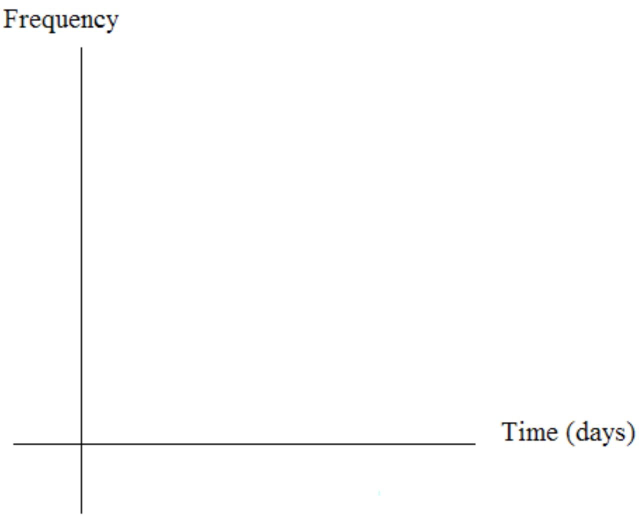


Figure 7.4

6. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Repeat the Procedure for n=5

1. For the sample of *n* = 5 days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths = $\frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.

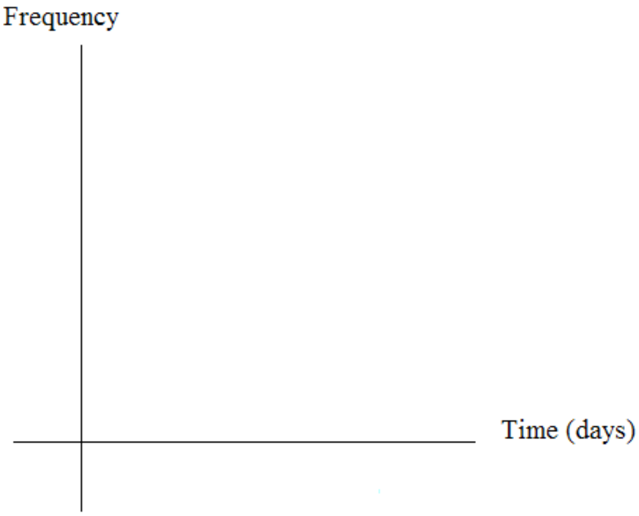


Figure 7.5

2. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Repeat the Procedure for n=10

1. For the sample of *n* = 10 days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths = $\frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.

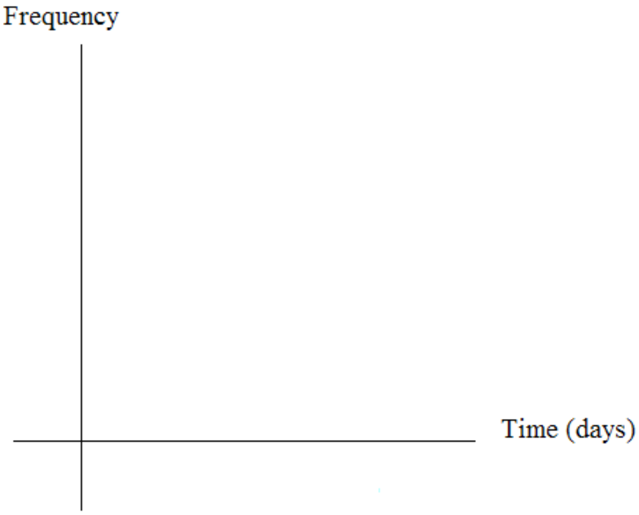


Figure 7.6

2. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Discussion Questions

1. Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.
2. State the theoretical (according to the CLT) distributions for the sample means.

   a. *n* = 5: $\overline{X}$ ~

   b. *n* = 10: $\overline{X}$ ~
3. Are the sample means for n = 5 and n = 10 “close” to the theoretical mean, μ_X? Explain why or why not.
4. Which of the two distributions of sample means has the smaller standard deviation? Why?
5. As n changed, why did the shape of the distribution of the data change? Use 1 – 2 complete sentences to explain what happened.

This lab was designed and contributed by Carol Olmstead.

Glossary

Average: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Average: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Average: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Central Limit Theorem: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Central Limit Theorem: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Central Limit Theorem: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Exponential Distribution: A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: $X \sim Exp(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

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

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Standard Error of the Mean: The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$.

Uniform Distribution: A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a,b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(X) = \frac{1}{b-a}$ for $a<X<b$ or $a \le X \le b$. The cumulative distribution is $P(X \le x) = \frac{x-a}{b-a}$.

8 CONFIDENCE INTERVALS

8.1 Confidence Intervals

Student Learning Objectives

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

- Calculate and interpret confidence intervals for one population average and one population proportion.
- Interpret the student-t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the student-t distributions.

Introduction

Suppose you are trying to determine the average rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percent of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population** **parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct confidence intervals in which we believe the parameter lies.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student-t, and how it is used with these intervals.

If you worked in the marketing department of an entertainment company, you might be interested in the average number of compact discs (CD's) a consumer buys per month. If so, you could conduct a survey and calculate the sample average, $\overline{x}$, and the sample standard deviation, *s*. You would use $\overline{x}$ to estimate the population mean and *s* to estimate the population standard deviation. The sample mean, $\overline{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, *s*, is the point estimate for the population standard deviation, σ.

Each of $\overline{x}$ and *s* is also called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is an estimated range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the CD example we do not know the population mean μ but we do know that the population standard deviation is σ = 1 and our sample size is 100. Then by the Central Limit Theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, will be within two standard deviations of the population mean μ. For our CD example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\overline{x}$ is within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is within 0.2 units of $\overline{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations ((2)(0.1)) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\overline{x} - 0.2$ and $\overline{x} + 0.2$ in 95% of all the samples.

For the CD example, suppose that a sample produced a sample mean $\overline{x} = 2$. Then the unknown population mean μ is between

$$\overline{x} - 0.2 = 2 - 0.2 = 1.8 \text{ and } \overline{x} + 0.2 = 2 + 0.2 = 2.2$$

We say that we are **95% confident** that the unknown population mean number of CDs is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (100% - 95%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. A confidence interval has the form

(point estimate - margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence.

Optional Collaborative Classroom Activity

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be 3 meals. Construct an approximate 95% confidence interval for the true average number of meals students eat out each week.

1. Calculate the sample mean.
2. σ = 3 and *n* = the number of students surveyed.
3. Construct the interval $\left(\overline{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}\right)$

We say we are approximately 95% confident that the true average number of meals that students eat out in a week is between __________ and __________.

8.2 Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ **, where the population standard deviation is known,** we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error is called the **error bound for a population mean** (abbreviated **EBM**). The sample mean $\bar{x}$ is the **point estimate** of the unknown population mean μ

(point estimate - error bound, point estimate + error bound) or, in symbols, $\left(\bar{x} - \text{EBM}, \bar{x} + \text{EBM}\right)$

The margin of error depends on the **confidence level** (abbreviated **CL**). The confidence level is the probability that the confidence interval estimate that we will calculate will contain the true population parameter. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because he wants to be reasonably certain of his conclusions.

There is another probability called alpha (α). α is related to the confidence level CL. α is the probability that the sample produced a point estimate that is not within the appropriate margin of error of the unknown population parameter.

Example 8.1

Suppose we have collected data from a sample. We know the sample average but we do not know the average for the entire population. The sample mean is 7 and the error bound for the mean is 2.5.

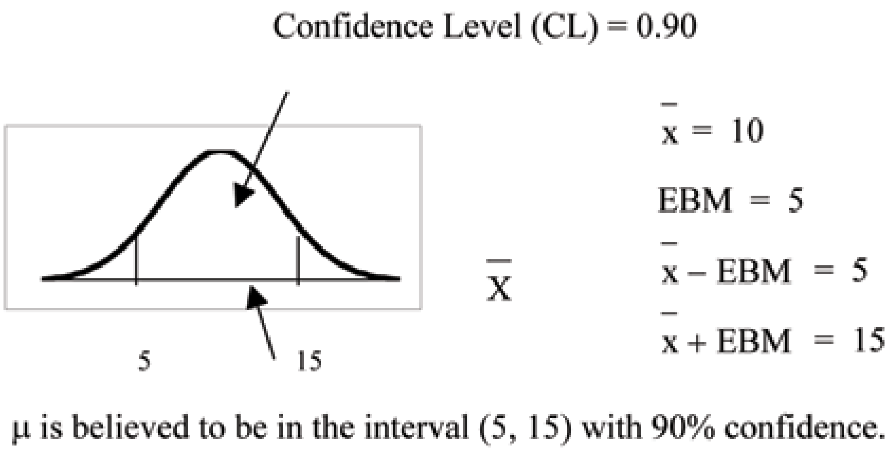
$\bar{x} = 7$ and EBM = 2.5.

The confidence interval is (7 − 2.5, 7 + 2.5); calculating the values gives (4.5, 9.5).

If the confidence level (CL) is 95%, then we say that "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where EBM = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of 10% in both tails, or 5% in each tail, of the normal distribution.


Confidence Level (CL) = 0.90
5
15
$\bar{X}$
$\bar{x}$ = 10
EBM = 5
$\bar{x}$ − EBM = 5
$\bar{x}$ + EBM = 15
μ is believed to be in the interval (5, 15) with 90% confidence.

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. 1.645 is the z-score from a Standard Normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating. So in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. $\frac{\sigma}{\sqrt{n}}$ is commonly called the "standard error of the mean" in order to clearly distinguish the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the Central Limit Theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a Normal distribution to calculate the error bound.**

Calculating the Confidence Interval:

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\overline{x}$ from the sample data. Remember, in this section, we already know the population standard deviation σ.
- Find the Z-score that corresponds to the confidence level.
- Calculate the error bound EBM
- Construct the confidence interval
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding z for the stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution Z~N(0,1).

The confidence level, CL, is he area in the middle of the standard normal distribution. CL = 1 − α. So α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$

For example, when CL = 0.95 then α = 0.05 and $\frac{\alpha}{2}$ = 0.025 ; we write $z_{\frac{\alpha}{2}} = z_{.025}$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 1-0.025 = 0.975

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a Standard Normal probability table.

Using the TI83, TI83+ or TI84+ calculator: `invNorm`(.975, 0, 1) = 1.96

CALCULATOR NOTE: Remember to use area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invnorm command are 0,1 because you are using a Standard Normal Distribution Z~N(0,1)

EBM: Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

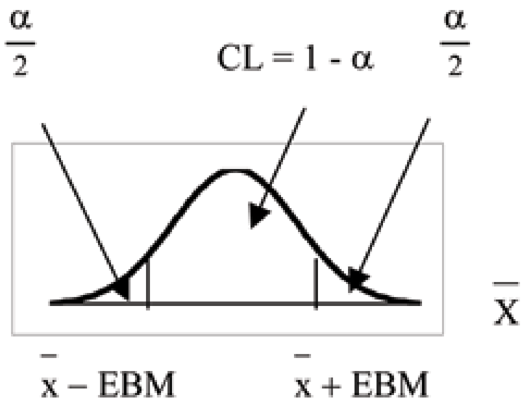
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Constructing the Confidence Interval

- The confidence interval estimate has the format $\left(\overline{x} - \text{EBM}, \overline{x} + \text{EBM}\right)$.

The graph gives a picture of the entire situation.

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



Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean or average**), and should state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population average (include context of the problem) is between ___ and ___ (include appopriate units)."

Example 8.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean (sample average score) of 68. Find a confidence interval estimate for the population mean exam score (the average score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

- You can use technology to directly calculate the confidence interval
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+ and 84+ calculators (Solution B).

Solution A

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

$\bar{x} = 68$

$$\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3$; $n = 36$; The confidence level is 90% (CL=0.90)

CL = 0.90 so $\alpha = 1 - \text{CL} = 1 - 0.90 = 0.10$

$\frac{\alpha}{2} = 0.05 \qquad z_{\frac{\alpha}{2}} = z_{.05}$

The area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 1−0.05=0.95

$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

using invnorm(.95,0,1) on the TI-83,83+,84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.

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

$\bar{x} - \text{EBM} = 68 - 0.8225 = 67.1775$

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

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

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `7:ZInterval`.
Press `ENTER`.
Arrow to `Stats` and press `ENTER`.
Arrow down and enter 3 for σ, 68 for $\bar{x}$, 36 for *n*, and .90 for `C-level`.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (to 3 decimal places) (67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true average statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Changing the Confidence Level or Sample Size

Example 8.3 Changing the Confidence Level

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

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

$\bar{x} = 68$

$$\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

$\sigma = 3$; $n = 36$; The confidence level is 95% (CL=0.95)

CL = 0.95 so $\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \qquad z_{\frac{\alpha}{2}} = z_{.025}$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 1−0.025=0.975

$z_{\frac{\alpha}{2}} = z_{.025} = 1.96$

using invnorm(.975,0,1) on the TI-83,83+,84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.)

$$\text{EBM} = 1.96 \cdot \left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$\bar{x} - \text{EBM} = 68 - 0.98 = 67.02$

$\bar{x} + \text{EBM} = 68 + 0.98 = 68.98$

Interpretation

We estimate with 95 % confidence that the true population average for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

95% of all confidence intervals constructed in this way contain the true value of the population average statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider.

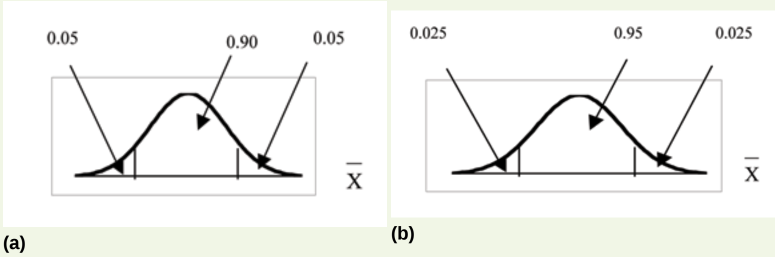


Figure 8.1

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Example 8.4 Changing the Sample Size:

Suppose we change the original problem to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use n=100 instead of n=36? What happens if we decrease the sample size to n=25 instead of n=36?

- $\overline{x} = 68$
- $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90) ; $z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100$: $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) = 1.645 \cdot \left(\frac{3}{\sqrt{100}}\right) = 0.4935$

B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right) = 1.645 \cdot \left(\frac{3}{\sqrt{25}}\right) = 0.987$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Working Backwards to Find the Error Bound or Sample Mean

Working Bacwards to find the Error Bound or the Sample Mean

When we calculate a confidence interval, we find the sample mean and calculate the error bound and use them to calculate the confidence interval. But sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean
- OR, From the upper value for the interval, subtract the lower value. Then divide the difference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval
- OR, Average the upper and lower endpoints of the confidence interval

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 8.5

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68. Or perhaps our source only gave the confidence interval and did not tell us the value of the the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $\text{EBM} = 68.82 - 68 = 0.82$
- If we don't know the sample mean: $\text{EBM} = \frac{(68.82 - 67.18)}{2} = 0.82$

Calculate the Sample Mean:

- If we know the error bound: $\overline{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\overline{x} = \frac{(67.18 + 68.82)}{2} = 68$

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$

The formula for sample size is $n = \frac{z^2\sigma^2}{\text{EBM}^2}$, found by solving the error bound formula for n

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 8.6

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within 2 years of the true population mean age of Foothill College students , how many randomly selected Foothill College students must be surveyed?

From the problem, we know that σ = 15 and EBM=2

$z = z_{.025} = 1.96$, because the confidence level is 95%.

$n = \frac{z^2\sigma^2}{\text{EBM}^2} = \frac{1.96^2 15^2}{2^2}$ =216.09 using the sample size equation.

Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within 2 years of the true population age of Foothill College students.

**With contributions from Roberta Bloom

8.3 Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gossett (1876-1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student-t distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid 1990s, statisticians used the **normal distribution** approximation for large sample sizes and only used the Student-t distribution for sample sizes of at most 30. With the common use of graphing calculators and computers, the practice is to use the Student-t distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student-t distribution with $n - 1$ degrees of freedom**. The t-score has the same interpretation as the **z-score**. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student-t distribution.

The **degrees of freedom**, $\boldsymbol{n - 1}$, come from the calculation of the sample standard deviation $\boldsymbol{s}$. In Chapter 2, we used n deviations $\left(\boldsymbol{x - \bar{x}} \textbf{ values}\right)$ to calculate $\boldsymbol{s}$. Because the sum of the deviations is 0, we can find the last deviation once we know the other $\boldsymbol{n - 1}$ deviations. The other $\boldsymbol{n - 1}$ deviations can change or vary freely. **We call the number $\boldsymbol{n - 1}$ the degrees of freedom (df).**

Properties of the Student-t Distribution

- The graph for the Student-t distribution is similar to the Standard Normal curve.
- The mean for the Student-t distribution is 0 and the distribution is symmetric about 0.
- The Student-t distribution has more probability in its tails than the Standard Normal distribution because the spread of the t distribution is greater than the spread of the Standard Normal. So the graph of the Student-t distribution will be thicker in the tails and shorter in the center than the graph of the Standard Normal distribution.
- The exact shape of the Student-t distribution depends on the "degrees of freedom". As the degrees of freedom increases, the graph Student-t distribution becomes more like the graph of the Standard Normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean **μ** and unknown population standard deviation **σ**. In the real world, however, as long as the underlying population is large and bell-shaped, and the data are a simple random sample, practitioners often consider the assumptions met.

Calculators and computers can easily calculate any Student-t probabilities. The TI-83,83+,84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student-t distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student-t distribution.) When using t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student-t table (See the Table of Contents **15. Tables**) gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student-t probabilities.**

The notation for the Student-t distribution is (using T as the random variable) is

- $T \sim t_{df}$ where df = $n - 1$.
- For example, if we have a sample of size n=20 items, then we calculate the degrees of freedom as df=n−1=20−1=19 and we write the distribution as $T \sim t_{19}$

If the population standard deviation is not known, the **error bound for a population mean** is:

- $\text{EBM} = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$
- use df = $n - 1$ degrees of freedom
- s = sample standard deviation

The format for the confidence interval is:

$\left(\bar{x} - \text{EBM}, \bar{x} + \text{EBM}\right)$.

The TI-83, 83+ and 84 calculators have a function that calculates the confidence interval directly. To get to it,
Press `STAT`
Arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or just press 8).

Example 8.7

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+ and 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses the Ti-83+ and Ti-84 calculators (Solution B).

Solution A

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

$\bar{x} = 8.2267$ $s = 1.6722$ $n = 15$

df = 15 − 1 = 14

CL = 0.95 so α = 1 − CL = 1 − 0.95 = 0.05

$\frac{\alpha}{2} = 0.025$ $t_{\frac{\alpha}{2}} = t_{.025}$

The area to the right of $t_{.025}$ is 0.025 and the area to the left of $t_{.025}$ is 1−0.025=0.975

$t_{\frac{\alpha}{2}} = t_{.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$\text{EBM} = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

$$\text{EBM} = 2.14 \cdot \left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$\bar{x} - \text{EBM} = 8.2267 - 0.9240 = 7.3$

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

The 95% confidence interval is **(7.30, 9.15)**.

We estimate with 95% confidence that the true population average sensory rate is between 7.30 and 9.15.

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to `Data` and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter .95
Arrow down to `Calculate` and press `ENTER`.
The 95% confidence interval is (7.3006, 9.1527)

When calculating the error bound, a probability table for the Student-t distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

**With contributions from Roberta Bloom

8.4 Confidence Interval for a Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within 3 percentage points. Often, election polls are calculated with 95% confidence. So, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 : (0.40 – 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound,** and the **confidence level** for a proportion is similar to that for the population mean. The formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying **distribution is binomial**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n = the number of trials and p = the probability of a success. To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large, we can use the **normal distribution** to approximate the binomial.

$$X \sim N(n \cdot p, \sqrt{n \cdot p \cdot q})$$

If we divide the random variable by n, the mean by n, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion = the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right)$$

Using algebra to simplify : $\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$

P' follows a normal distribution for proportions: $P' \sim N\left(p, \sqrt{\left(\frac{p \cdot q}{n}\right)}\right)$

The confidence interval has the form **(p' – EBP, p' + EBP)**.

$$p' = \frac{x}{n}$$

p' = the **estimated proportion** of successes (p' is a **point estimate** for p, the true proportion)

x = the **number** of successes.

n = the size of the sample

The error bound for a proportion is

$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\left(\frac{p' \cdot q'}{n}\right)}$ $\qquad q' = 1 - p'$

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{p \cdot q}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p' \cdot q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{p \cdot q}{n}}$

However, in the error bound formula, the standard deviation is $\sqrt{\left(\frac{p' \cdot q'}{n}\right)}$.

In the error bound formula, the **sample proportions *p'* and *q'* are estimates of the unknown population proportions *p* and *q***. The estimated proportions *p'* and *q'* are used because *p* and *q* are not known. *p'* and *q'* are calculated from the data. *p'* is the estimated proportion of successes. *q'* is the estimated proportion of failures.

For the normal distribution of proportions, the z-score formula is as follows.

If $P' \sim N\left(p, \sqrt{\left(\frac{p \cdot q}{n}\right)}\right)$ then the z-score formula is $z = \frac{p' - p}{\sqrt{\left(\frac{p \cdot q}{n}\right)}}$

Example 8.8

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adults residents of this city who have cell phones.

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

A

Let *X* = the number of people in the sample who have cell phones. *X* is binomial. $X \sim B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find *p'*, *q'*, and EBP.

$n = 500$ $\qquad x$ = the number of successes = 421

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

$p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$q' = 1 - p' = 1 - 0.842 = 0.158$

Since CL = 0.95, then α = 1 − CL = 1 − 0.95 = 0.05 $\qquad \frac{\alpha}{2} = 0.025$.

Then $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$

Use the TI-83, 83+ or 84+ calculator command invnorm(.975,0,1) to find $z_{.025}$. Remember that the area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.96 \cdot \sqrt{\left[\frac{(.842) \cdot (.158)}{500}\right]} = 0.032$

$p' - \text{EBP} = 0.842 - 0.032 = 0.81$

$p' + \text{EBP} = 0.842 + 0.032 = 0.874$

The confidence interval for the true binomial population proportion is (p' − EBP, p' + EBP) = **(0.810, 0.874)**.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

95% of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

B

Using a function of the TI-83, 83+ or 84 calculators:

Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:PropZint`. Press `ENTER`.
Arrow down to *x* and enter 421.
Arrow down to *n* and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.81003, 0.87397).

Example 8.9

For a class project, a political science student at a large university wants to determine the percent of students that are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students that are registered voters and interpret the confidence interval.

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution A

$x = 300$ and $n = 500$.

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since CL = 0.90, then $\alpha = 1 - CL = 1 - 0.90 = 0.10$ $\qquad \frac{\alpha}{2} = 0.05$.

$$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$$

Use the TI-83, 83+ or 84+ calculator command invnorm(.95,0,1) to find $z_{.05}$. Remember that the area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.645 \cdot \sqrt{\left[\frac{(.60) \cdot (.40)}{500}\right]} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP)$ = **(0.564, 0.636)**.

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Using a function of the TI-83, 83+ or 84 calculators:

Press `STAT` and arrow over to `TESTS`.

Arrow down to `A:PropZint`. Press `ENTER`.
Arrow down to *x* and enter 300.
Arrow down to *n* and enter 500.
Arrow down to `C-Level` and enter .90.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.564, 0.636).

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

- $\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\left(\frac{p'q'}{n}\right)}$
- Solving for *n* gives you an equation for the sample size.
- $n = \dfrac{z_{\frac{\alpha}{2}}^{2} \cdot p'q'}{\text{EBP}^2}$

Example 8.10

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ that use text messaging on their cell phone. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ that use text messaging on their cell phone.

Solution

From the problem, we know that **EBP=0.03** (3%=0.03) and

$z_{\frac{\alpha}{2}} = z_{.05}$ = 1.645 because the confidence level is 90%

However, in order to find n , we need to know the estimated (sample) proportion p'. Remember that q'=1-p'. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because p'q'= (.5)(.5)=.25 results in the largest possible product. (Try other products: (.6)(.4)=.24; (.3)(.7)=.21; (.2)(.8)=.16 and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$n = \dfrac{z^2 p'q'}{\text{EBP}^2}$ gives $n = \dfrac{1.645^2 (.5)(.5)}{.03^2}$=751.7

Round the answer to the next higher value. The sample size should be 758 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50+ that use text messaging on their cell phone.

**With contributions from Roberta Bloom.

8.5 Summary of Formulas

Formula

(lower value, upper value) = (point estimate − error bound, point estimate + error bound)

Formula

error bound = upper value − point estimate OR error bound = $\dfrac{\text{upper value} - \text{lower value}}{2}$

Formula

Use the Normal Distribution for Means $\text{EBM} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $\left(\overline{x} - \text{EBM}, \overline{x} + \text{EBM}\right)$.

Formula

Use the Student-t Distribution with degrees of freedom df = $n - 1$. $\text{EBM} = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

Formula

Use the Normal Distribution for a single population proportion $p' = \frac{x}{n}$

$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}}$ $\qquad p' + q' = 1$

The confidence interval has the format (p' − EBP, p' + EBP).

Formula

$\bar{x}$ is a point estimate for μ

p' is a point estimate for ρ

s is a point estimate for σ

8.6 Practice 1: Confidence Intervals for Averages, Known Population Standard Deviation

Student Learning Outcomes

- The student will explore the properties of Confidence Intervals for Averages when the population standard deviation is known.

Given

The average age for all Foothill College students for Fall 2005 was 32.7. The population standard deviation has been pretty consistent at 15. Twenty-five Winter 2006 students were randomly selected. The average age for the sample was 30.4. We are interested in the true average age for Winter 2006 Foothill College students. (**http://research.fhda.edu/factbook/FHdemofs/Fact_sheet_fh_2005f.pdf (http://research.fhda.edu/factbook/FHdemofs/Fact_sheet_fh_2005f.pdf)**)

Let $X =$ the age of a Winter 2006 Foothill College student

Calculating the Confidence Interval

Explaining the Confidence Interval

Construct a 95% Confidence Interval for the true average age of Winter 2006 Foothill College students.

Discussion Questions

8.7 Practice 2: Confidence Intervals for Averages, Unknown Population Standard Deviation

Student Learning Outcomes

- The student will explore the properties of confidence intervals for averages when the population standard deviation is unknown.

Given

The following real data are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true average number of colors on a national flag. Let $X =$ the number of colors on a national flag.

Table 8.1

X	Freq.
1	1
2	7
3	18
4	7
5	6

Calculating the Confidence Interval

Confidence Interval for the True Average Number

Construct a 95% Confidence Interval for the true average number of colors on national flags.

Discussion Questions

8.8 Practice 3: Confidence Intervals for Proportions

Student Learning Outcomes

- The student will explore the properties of the confidence intervals for proportions.

Given

The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 - 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 - 12, in all beginning ice-skating classes at the Ice Chalet.

Estimated Distribution

Explaining the Confidence Interval

Construct a 92% Confidence Interval for the true proportion of girls in the age 8 - 12 beginning ice-skating classes at the Ice Chalet.

Discussion Questions

8.9 Homework

If you are using a student-t distribution for a homework problem below, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Try these multiple choice questions.

The next three problems refer to the following: According a Field Poll conducted February 8 – 17, 2005, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

The next two problems refer to the following:

A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample average is 13.30 with a sample standard deviation is 1.55. Assume the underlying population is normally distributed.

The next two problems refer to the following:

Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness and 338 did not.

8.10 Review

The next three problems refer to the following situation: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to 3 pounds.

The next three questions refer to the following situation: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

- So, if an accident occurs at 4 p.m., the amount of time, in hours, it took for the accident to occur is _______.
- μ = ______
- σ^2 = ______

The next two questions refer to the following situation: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from 1 to 15 days.

The next five problems refer to the following study: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

The next two problems refer to the following information: $P(A) = 0.2$, $P(B) = 0.3$, A and B are independent events.

8.11 Lab 1: Confidence Interval (Home Costs)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate the 90% confidence interval for the average cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will examine the effects that changing conditions has on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. (Note: many papers only list them one day per week. Also, we will assume that homes come up for sale randomly.) Record the sales prices for 35 randomly selected homes recently listed in the county.

1. Complete the table:

Table 8.2

Describe the Data

1. Compute the following:

   a. $\overline{x}$ =

   b. s_x =

   c. n =

2. Define the Random Variable $\overline{X}$, in words. $\overline{X}$ =
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.

   a. Confidence Interval:

   b. Error Bound:

2. How much area is in both tails (combined)? α =
3. How much area is in each tail? $\frac{\alpha}{2}$ =
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.

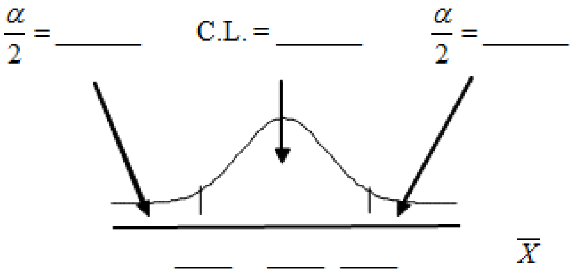


Figure 8.2

5. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a Confidence Interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this Confidence Interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the above information, construct a confidence interval for each confidence level given.

Table 8.3

Confidence level	EBM / Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

2. What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.12 Lab 2: Confidence Interval (Place of Birth)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate the 90% confidence interval for proportion of students in this school that were born in this state.
- The student will interpret confidence intervals.
- The student will examine the effects that changing conditions have on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them if they were born in this state. Let X = the number that were born in this state.

   a. n =____________

   b. X =____________

2. Define the Random Variable P' in words.
3. State the estimated distribution to use.

Find the Confidence Interval and Error Bound

1. Calculate the confidence interval and the error bound.

   a. Confidence Interval:

   b. Error Bound:

2. How much area is in both tails (combined)? α=
3. How much area is in each tail? $\frac{\alpha}{2}$ =
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.

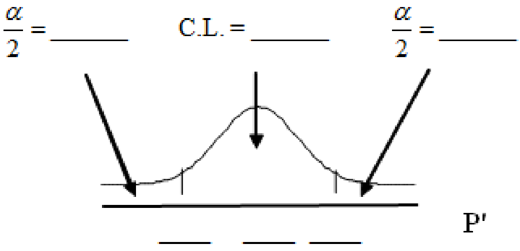


Figure 8.3

Describe the Confidence Interval

1. In two to three complete sentences, explain what a Confidence Interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this Confidence Interval means for this particular study.
3. Using the above information, construct a confidence interval for each given confidence level given.

Table 8.4

Confidence level	EBP / Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

4. What happens to the EBP as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.13 Lab 3: Confidence Interval (Womens' Heights)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate a 90% confidence interval using the given data.
- The student will examine the relationship between the confidence level and the percent of constructed intervals that contain the population average.

Given:

1.

Table 8.5 Heights of 100 Women (in Inches)

59.4	71.6	69.3	65.0	62.9
66.5	61.7	55.2	67.5	67.2
63.8	62.9	63.0	63.9	68.7
65.5	61.9	69.6	58.7	63.4
61.8	60.6	69.8	60.0	64.9
66.1	66.8	60.6	65.6	63.8
61.3	59.2	64.1	59.3	64.9
62.4	63.5	60.9	63.3	66.3
61.5	64.3	62.9	60.6	63.8
58.8	64.9	65.7	62.5	70.9
62.9	63.1	62.2	58.7	64.7
66.0	60.5	64.7	65.4	60.2
65.0	64.1	61.1	65.3	64.6
59.2	61.4	62.0	63.5	61.4
65.5	62.3	65.5	64.7	58.8
66.1	64.9	66.9	57.9	69.8
58.5	63.4	69.2	65.9	62.2
60.0	58.1	62.5	62.4	59.1
66.4	61.2	60.4	58.7	66.7
67.5	63.2	56.6	67.7	62.5

Listed above are the heights of 100 women. Use a random number generator to randomly select 10 data values.

2. Calculate the sample mean and sample standard deviation. Assume that the population standard deviation is known to be 3.3 inches. With these values, construct a 90% confidence interval for your sample of 10 values. Write the confidence interval you obtained in the first space of the table below.
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into the table below:

Table 8.6 90% Confidence Intervals

__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________
__________	__________	__________	__________	__________

Discussion Questions

1. The actual population mean for the 100 heights given above is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percent of confidence intervals that contains the mean μ. Write this percent below.
3. Is the percent of confidence intervals that contain the population mean μ close to 90%?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percent of confidence intervals that contained the population mean?
5. When we construct a 90% confidence interval, we say that we are **90% confident that the true population mean lies within the confidence interval.** Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated on that page. How many of the 100 data values lie within your confidence interval? What percent is this? Is this percent close to 90%?

7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of 10 women and calculated a confidence interval from this information. Without knowing the population mean μ, would you have any way of knowing **for certain** if your interval actually contained the value of μ? Explain.

This lab was designed and contributed by Diane Mathios.

Glossary

Binomial Distribution: A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, *n*, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: ***X*~*B*(*n*, *p*)**. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly *x* successes in *n* trials is $P\left(X = x\right) = \binom{n}{x}p^{x}q^{n-x}$.

Confidence Interval (CI): An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Interval (CI): An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Interval (CI): An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Interval (CI): An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL): The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Confidence Level (CL): The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Confidence Level (CL): The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Degrees of Freedom (df): The number of objects in a sample that are free to vary.

Error Bound for a Population Mean (EBM): The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Mean (EBM): The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion(EBP): The margin of error. Depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

Inferential Statistics : Also called statistical inference or inductive statistics. This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if 4 out of the 100 calculators sampled are defective we might infer that 4 percent of the production is defective.

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^{2}/2\sigma^{2}}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: *X* ~ *N*(μ, σ). If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^{2}/2\sigma^{2}}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: *X* ~ *N*(μ, σ). If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Parameter: A numerical characteristic of the population.

Point Estimate: A single number computed from a sample and used to estimate a population parameter.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student-t Distribution: Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

9 HYPOTHESIS TESTING: SINGLE MEAN AND SINGLE PROPORTION

9.1 Hypothesis Testing: Single Mean and Single Proportion

Student Learning Objectives

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

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called **"hypothesis testing."** A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not the data supports the claim that is made about the population.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will support one of the hypotheses.
5. Make a decision and write a meaningful conclusion.

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See the Table of Contents topic "Solution Sheets".

9.2 Null and Alternate Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternate hypothesis**. These hypotheses contain opposing viewpoints.

H_o: **The null hypothesis:** It is a statement about the population that will be assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt.

H_a: **The alternate hypothesis:** It is a claim about the population that is contradictory to H_o and what we conclude when we reject H_o.

Example 9.1

$\boldsymbol{H_o}$: No more than 30% of the registered voters in Santa Clara County voted in the primary election.

$\boldsymbol{H_a}$: More than 30% of the registered voters in Santa Clara County voted in the primary election.

Example 9.2

We want to test whether the average grade point average in American colleges is 2.0 (out of 4.0) or not.

H_o: $\mu = 2.0$ H_a: $\mu \neq 2.0$

Example 9.3

We want to test if college students take less than five years to graduate from college, on the average.

H_o: $\mu \geq 5$ H_a: $\mu < 5$

Example 9.4

In an issue of **U. S. News and World Report**, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U. S. students take advanced placement exams and 4.4 % pass. Test if the percentage of U. S. students who take advanced placement exams is more than 6.6%.

H_o: $p = 0.066$ H_a: $p > 0.066$

Since the null and alternate hypotheses are contradictory, you must examine evidence to decide which hypothesis the evidence supports. The evidence is in the form of sample data. The sample might support either the null hypothesis or the alternate hypothesis but not both.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "reject H_o" if the sample information favors the alternate hypothesis or "do not reject H_o" if the sample information favors the null hypothesis, meaning that there is not enough information to reject the null.

Mathematical Symbols Used in H_o and H_a:

Table 9.1

H_o	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

H_o always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the Null Hypothesis, even with > or < as the symbol in the Alternate Hypothesis. This practice is acceptable because we only make the decision to reject or not reject the Null Hypothesis.

Optional Collaborative Classroom Activity

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write a null and alternate hypotheses. Discuss your hypotheses with the rest of the class.

9.3 Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four outcomes depending on the actual truth (or falseness) of the null hypothesis H_o and the decision to reject or not. The outcomes are summarized in the following table:

Table 9.2

ACTION	H_o IS ACTUALLY	...
	True	False
Do not reject H_o	Correct Outcome	Type II error
Reject H_o	Type I Error	Correct Outcome

The four outcomes in the table are:

- The decision is to **not reject H_o** when, in fact, **H_o is true (correct decision).**
- The decision is to **reject H_o** when, in fact, **H_o is true** (incorrect decision known as a **Type I error**).
- The decision is to **not reject H_o** when, in fact, **H_o is false** (incorrect decision known as a **Type II error**).
- The decision is to **reject H_o** when, in fact, **H_o is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters **α** and **β** represent the probabilities.

α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely 0.

The Power of the Test is 1 − β. Ideally, we want a high power that is as close to 1 as possible.

The following are examples of Type I and Type II errors.

Example 9.5

Suppose the null hypothesis, H_O, is: Frank's rock climbing equipment is safe.

Type I error: Frank concludes that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank concludes that his rock climbing equipment is safe when, in fact, it is not safe.

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

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Example 9.6

Suppose the null hypothesis, H_O, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew concludes that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew concludes that the victim is alive when, in fact, the victim is dead.

α = probability that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). **β = probability** that the emergency crew thinks the victim is alive when, in fact, he is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

9.4 Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a **normal distribution** or a **student-t distribution.** (Remember, use a student-t distribution when the population **standard deviation** is unknown and the population from which the sample is taken is normal.) In this chapter we perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a **single population mean**, the distribution for the test is for **averages**:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \qquad \text{or} \qquad t_{df}$$

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

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

$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

9.5 Assumption

When you perform a **hypothesis test of a single population mean μ** using a **Student-t distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is larger than 30, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean μ** using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is larger than 30 or both. You know the value of the population standard deviation.

When you perform a **hypothesis test of a single population proportion *p***, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\left(\frac{p \cdot q}{n}\right)}$. Remember that $q = 1 - p$.

9.6 Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption** - it is not a fact and it may or may not be true. But your sample data is real and it is showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200}$ = 0.005. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

9.7 Using the Sample to Support One of the Hypotheses

Use the sample (data) to calculate the actual probability of getting the test result, called the **p-value**. The p-value is the **probability that an outcome of the data (for example, the sample mean) will happen purely by chance when the null hypothesis is true**.

A large p-value calculated from the data indicates that the sample result is likely happening purely by chance. The data support the **null hypothesis** so we do not reject it. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against the null hypothesis.

The p-value is sometimes called the **computed α** because it is calculated from the data. You can think of it as the probability of (incorrectly) rejecting the null hypothesis when the null hypothesis is actually true.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 9.7 (to illustrate the p-value)

Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The average height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm.

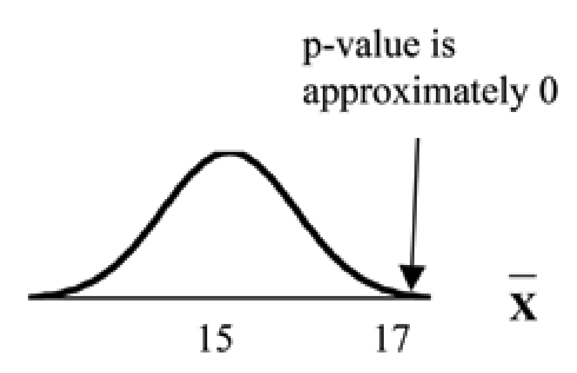
The null hypothesis could be H_o: μ ≤ 15 The alternate hypothesis is H_a: μ > 15

The words **"is more than"** translates as a " > " so "μ > 15" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since **σ is known** (σ = 0.5 cm.), the distribution for the test is normal with mean μ = 15 and standard deviation $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the average height of the loaves is no more than 15 cm). Then is the average height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample average would be if the null hypothesis were true. The graph shows how far out the sample average is on the normal curve. How far out the sample average is on the normal curve is measured by the p-value. The p-value is the probability that, if we were to take other samples, any other sample average would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample average is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for averages from Chapter 7.



p-value = $P(\overline{X} > 17)$ which is approximately 0.

A p-value of approximately 0 tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on the average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE**. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the average height is at most 15 cm.). There is sufficient evidence that the true average height for the population of the baker's loaves of bread is greater than 15 cm.

9.8 Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the **p-value** and a **preset or preconceived α (also called a "significance level")**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_o, do as follows:

- If α > p-value, reject H_o. The results of the sample data are significant. There is sufficient evidence to conclude that H_o is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If α ≤ p-value, do not reject H_o. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.
- When you "do not reject H_o", it does not mean that you should believe that H_o is true. It simply means that the sample data has **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

9.9 Additional Information

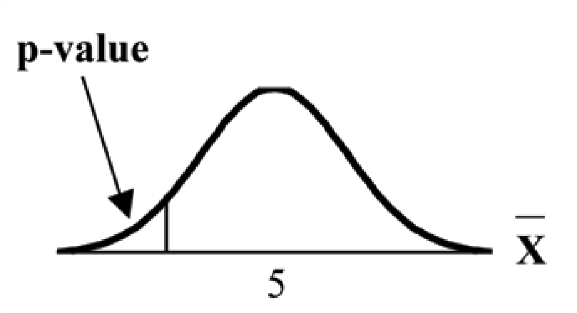
- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, we generally can use α = 0.05.**
- When you calculate the **p-value** and draw the picture, the p-value is in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternate hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.

The following examples illustrate a left, right, and two-tailed test.

Example 9.8

H_o: μ = 5 H_a: μ < 5

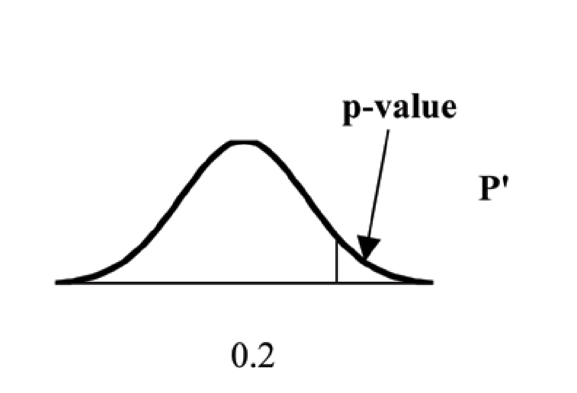
Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:



Example 9.9

H_o: p ≤ 0.2 H_a: p > 0.2

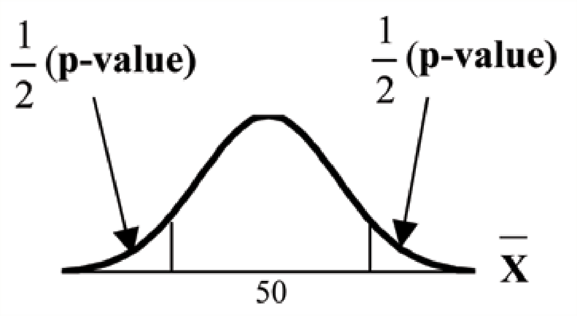
This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:



Example 9.10

H_o: μ = 50 H_a: μ ≠ 50

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.



9.10 Summary of the Hypothesis Test

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_o and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the **p-value**. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or cannot reject H_o), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity 1 – β is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

9.11 Examples

Example 9.11

Jeffrey, as an eight-year old, **established an average time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster by using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's average time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset α = 0.05. Assume that the swim times for the 25-yard freestyle are normal.

Set up the Hypothesis Test:

Since the problem is about a mean (average), this is a **test of a single population mean**.

H_o: μ = 16.43 H_a: μ < 16.43

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the average time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population **standard deviation** is known: σ = 0.8)

$\bar{X} \sim N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right)$ Therefore, $\bar{X} \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

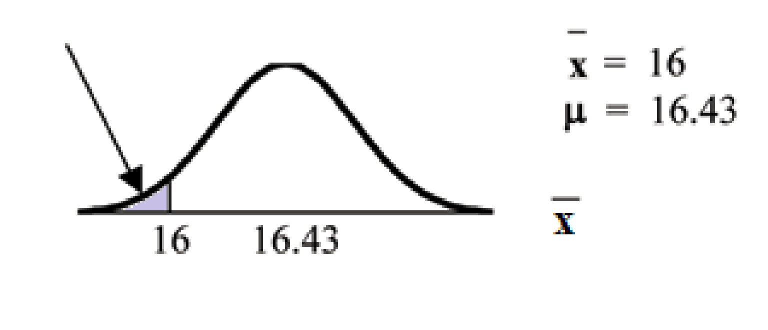
μ = 16.43 comes from H_0 and not the data. σ = 0.8, and n = 15.

Calculate the p-value using the normal distribution for a mean:

p-value = $P\left(\bar{x} < 16\right)$ = 0.0187 where the sample mean in the problem is given s 16.

p-value = 0.0187 (This is called the **actual level of significance**.) The p-value is the area to the left of the sample mean is given as 16.

Graph:



μ = 16.43 comes from H_0. Our assumption is μ = 16.43.

Interpretation of the p-value: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean (or average) time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is not happening randomly. It is a rare event.

Compare α and the p-value:

α = 0.05 p-value = 0.0187 α > p-value

Make a decision: Since α > p-value, reject H_0.

This means that you reject μ = 16.43. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for *n*. Arrow down to μ: (alternate hypothesis) and arrow over to <μ_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (*p* = 0.0187) but it also calculates the test statistic (z-score) for the sample mean. μ<16.43 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with *z* = -2.08 (test statistic) and *p* = 0.0187 (p-value). Make sure when you use `Draw` that no other equations are highlighted in *Y* = and the plots are turned off.

When the calculator does a Z-Test, the `Z-Test` function finds the p-value by doing a normal probability calculation using the **Central Limit Theorem**:

$P(\bar{x} < 16)$ = `2nd DISTR normcdf` (-10^99, 16, 16.43, 0.8 / $\sqrt{15}$).

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Historical Note: The traditional way to compare the two probabilities, α and the p-value, is to compare their test statistics (z-scores). The calculated test statistic for the p-value is -2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$. For this problem, $\bar{x}$ = 16, μ_X = 16.43 from the null hypothesis, σ_X = 0.8, and *n* = 15.) You can find the test statistic for α = 0.05 in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between -1.65 and -1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is -1.645. Since -1.645 > -2.08 (which demonstrates that α > p-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common and advantageous. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. It is difficult to know that the p-value is traditionally smaller than α by just examining the test statistics. The graph shows α, the p-value, and the two test statistics (z scores).

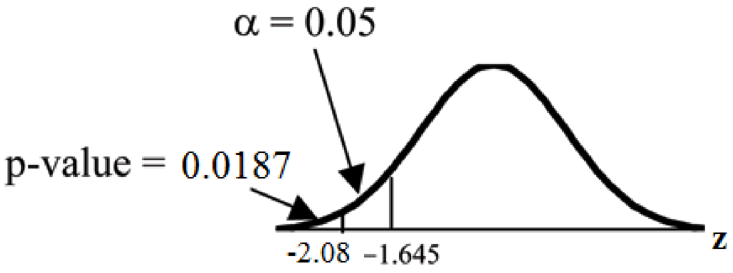


Figure 9.1

Example 9.12

A college football coach thought that his players could bench press an **average of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the average was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1). (Source: data from Reuben Davis, Kraig Evans, and Scott Gunderson.)

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press average is **more than 275 pounds**.

Set up the Hypothesis Test:

Since the problem is about a mean (average), this is a **test of a single population mean**.

H_o: $\mu = 275$ H_a: $\mu > 275$ This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X}$ = the average weight lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$\bar{X} \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

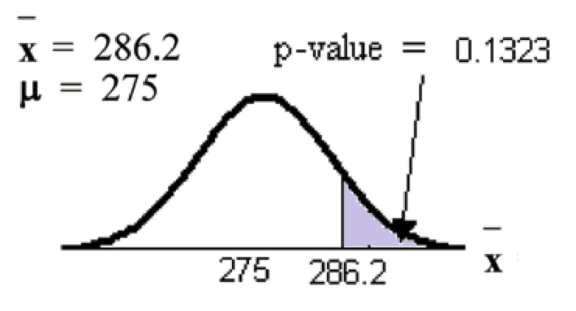
$\bar{x}$ = 286.2 pounds (from the data).

σ = 55 pounds **(Always use σ if you know it.)** We assume μ = 275 pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean:

p-value = $P(\bar{x} > 286.2) = 0.1323$ where the sample mean is calculated as 286.2 pounds from the data.

Interpretation of the p-value: If H_o is true, then there is a 0.1323 probability (13.23%) that the football players can lift a mean (or average) weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is happening randomly and is not a rare event.



Compare α and the p-value:

α = 0.025 p-value = 0.1323

Make a decision: Since α<p-value, do not reject H_o.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ : and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (p = 0.1331, a little different from the above calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. μ > 275 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with z = 1.112 (test statistic) and p = 0.1331 (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

Example 9.13

Statistics students believe that the average score on the first statistics test is 65. A statistics instructor thinks the average score is higher than 65. He samples ten statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5% level of significance. The data are from a normal distribution.

Set up the Hypothesis Test:

A 5% level of significance means that α = 0.05. This is a test of a **single population mean**.

H_0: μ = 65 H_a: μ > 65

Since the instructor thinks the average score is higher, use a " > ". The " > " means the test is right-tailed.

Determine the distribution needed:

Random variable: $\overline{X}$ = average score on the first statistics test.

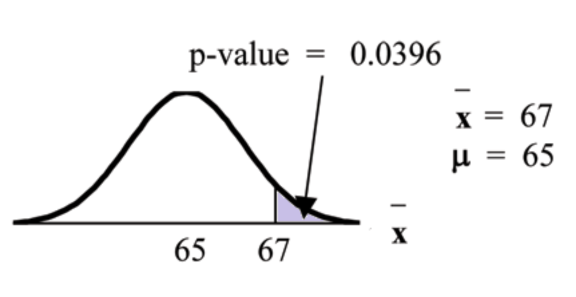
Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given n = 10 sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student-t.

Use t_{df}. Therefore, the distribution for the test is t_9 where n = 10 and df = 10 − 1 = 9.

Calculate the p-value using the Student-t distribution:

p-value = $P(\overline{x} > 67)$ = 0.0396 where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 67 or more.



Compare α and the p-value:

Since α = .05 and p-value = 0.0396. Therefore, α > p-value.

Make a decision: Since α > p-value, reject H_0.

This means you reject μ = 65. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ : and arrow over to $> \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. μ > 65 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press ENTER. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 9.14

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Set up the Hypothesis Test:

The 1% level of significance means that α = 0.01. This is a **test of a single population proportion**.

H_0: $p = 0.50$ H_a: $p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of an average. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

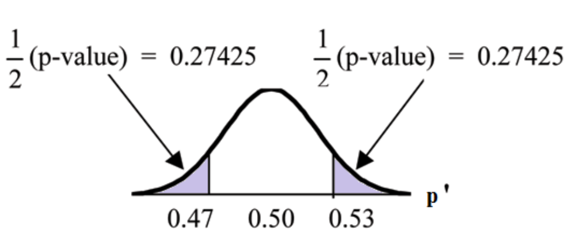
$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ Therefore, $P' \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$ where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$.

Calculate the p-value using the normal distribution for proportions:

p-value = $P(P' < 0.47 \text{ or } P' > 0.53) = 0.5485$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph below).



$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since α = 0.01 and p-value = 0.5485. Therefore, α< p-value.

Make a decision: Since α<p-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press STAT and arrow over to TESTS. Press `5:1-PropZTest`. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press ENTER. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides that are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is to conclude that the proportion of first-time brides that are younger than their grooms is equal to 50% when, in fact, the proportion is different from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Example 9.15

Suppose a consumer group suspects that the proportion of households that have three cell phones is not known to be 30%. A cell phone company has reason to believe that the proportion is 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Set up the Hypothesis Test:

H_0: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N\left(0.30, \sqrt{\frac{0.30 \cdot 0.70}{150}}\right)$

The value that helps determine the p-value is p'. Calculate p'.

$p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43$, $n = 150$

$p' = \frac{43}{150}$

What is a **success** for this problem?

A success is having three cell phones in a household.

What is the level of significance?

The level of significance is the preset α. Since α is not given, assume that α = 0.05.

Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Calculate the p-value.

p-value = 0.7216

Make a decision. _____________(Reject/Do not reject) H_0 because____________.

Assuming that α = 0.05, α < p-value. The Decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter *p*. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived α = 0.01, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Hypothesis testing problems consist of multiple steps. To help you do the problems, solution sheets are provided for your use. Look in the Table of Contents Appendix for the topic "Solution Sheets." If you like, use copies of the appropriate solution sheet for homework problems.

Example 9.16

```
My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had give up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Use the new shampoo or go without?
```

Set up the Hypothesis Test:

H_o: $p = 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\overline{X}$ or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

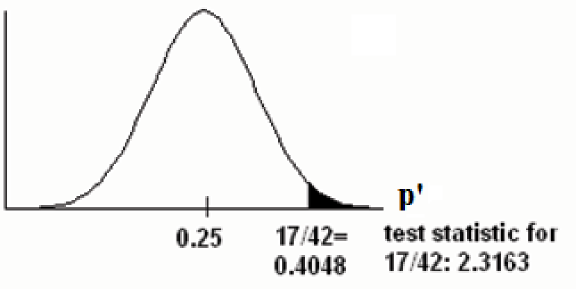
Calculate the p-value using the normal distribution for proportions:

p-value = 0.0103

In 1 – 2 complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



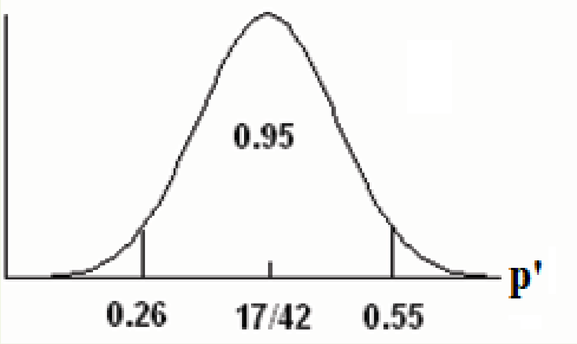
Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using COMPLETE SENTENCES.

alpha	decision	reason for decision
0.01	Do not reject H_o	α<p-value

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.



Confidence Interval: (0.26, 0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

9.12 Summary of Formulas

H_o and ***H_a*** are contradictory.

Table 9.3

If H_o has:	equal (=)	greater than or equal to (≥)	less than or equal to (≤)
then H_a has:	not equal (≠) **or** greater than (>) **or** less than (<)	less than (<)	greater than (>)

If α ≤ p-value, then do not reject ***H_o***.

If α > p-value, then reject ***H_o*** .

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

If there is no given preconceived α, then use α = 0.05.

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student-t test**.
- Single population proportion: **Normal test**.

9.13 Practice 1: Single Mean, Known Population Standard Deviation

Student Learning Outcomes

- The student will explore hypothesis testing with single mean and known population standard deviation.

Given

Suppose that a recent article stated that the average time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the average time has increased in the new century. A random sample of 26 first–time convicted burglars in a recent year was picked. The average length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the average length of jail time has increased.

Hypothesis Testing: Single Mean (Average)

Discussion Questions

9.14 Practice 2: Single Mean, Unknown Population Standard Deviation

Student Learning Outcomes

- The student will explore the properties of hypothesis testing with a single mean and unknown population standard deviation.

Given

A random survey of 75 death row inmates revealed that the average length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population average time on death row could likely be 15 years.

Hypothesis Testing: Single Average

Discussion Question

Does it appear that the average time on death row could be 15 years? Why or why not?

9.15 Practice 3: Single Proportion

Student Learning Outcomes

- The student will explore the properties of hypothesis testing with a single proportion.

Given

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. (http://www.nimh.nih.gov/publicat/depression.cfm) Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

Hypothesis Testing: Single Proportion

Discusion Question

9.16 Homework

Directions

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in the Appendix. Please feel free to make copies of it. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

If you are using a student-t distribution for a homework problem below, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

The following questions were written by past students. They are excellent problems!

Try these multiple choice questions.

The next three questions refer to the following information: A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of attended the midnight showing.

The next two questions refer to the following information:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than 7 hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated an average of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than 7 hours of sleep per night, on average?

The next three questions refer to the following information: An organization in 1995 reported that teenagers spent an average of 4.5 hours per week on the telephone. The organization thinks that, in 2007, the average is higher. Fifteen (15) randomly chosen teenagers were asked how many hours per week they spend on the telephone. The sample mean was 4.75 hours with a sample standard deviation of 2.0.

9.17 Review

The next three exercises refer to the following information: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. X = the number of estimates.

Table 9.4

X	Rel. Freq.	Cumulative Rel. Freq.
1	0.3	
2	0.2	
4	0.4	
5	0.1	

Complete the cumulative relative frequency column.

The next three questions refer to the following table: Seventy 5th and 6th graders were asked their favorite dinner.

Table 9.5

	Pizza	Hamburgers	Spaghetti	Fried shrimp
5th grader	15	6	9	0
6th grader	15	7	10	8

9.18 Lab: Hypothesis Testing of a Single Mean and Single Proportion

Class Time:

Names:

Student Learning Outcomes:

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that σ = 2. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_o:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately.Shade the actual level of significance.
   a. Graph:

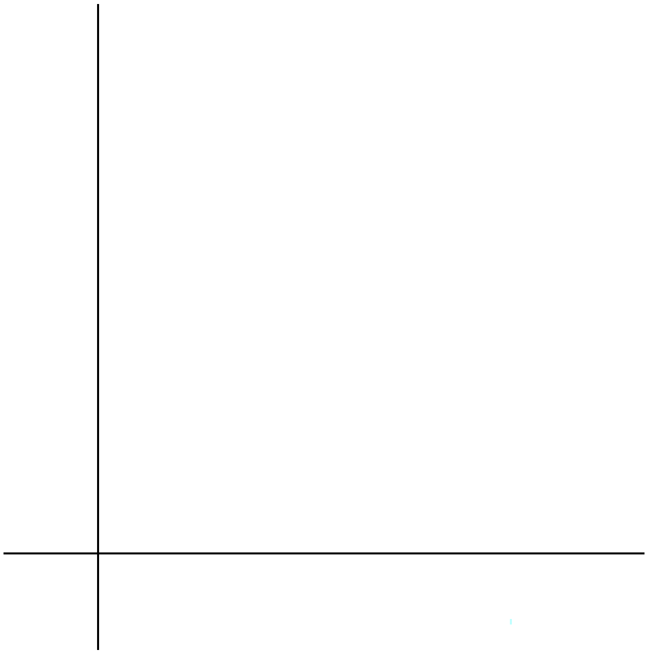


Figure 9.2

   b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

According to the 2000 Census, about 39.5% of Californians and 17.9% of all Americans speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school that speak a language other than English at home is different from 39.5%.

1. H_o:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

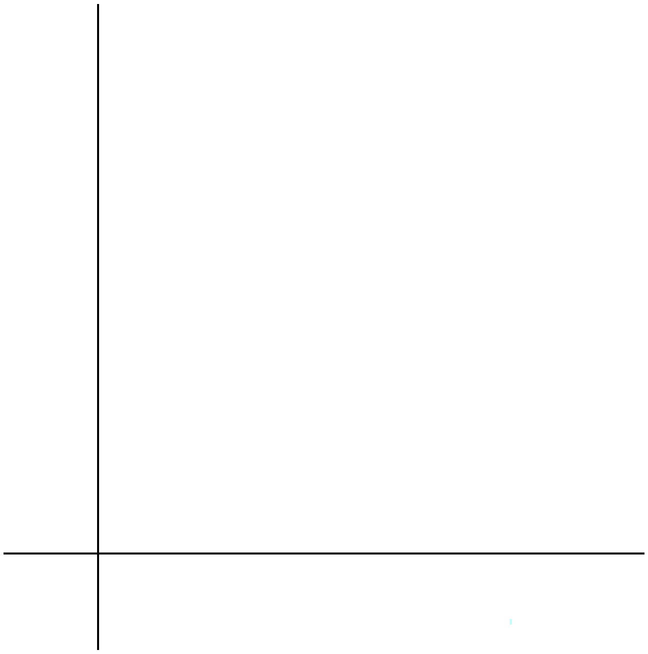


Figure 9.3

b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

Suppose that young adults own an average of 3 pairs of jeans. Survey 8 people from your class to determine if the average is higher than 3.

1. H_o:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:

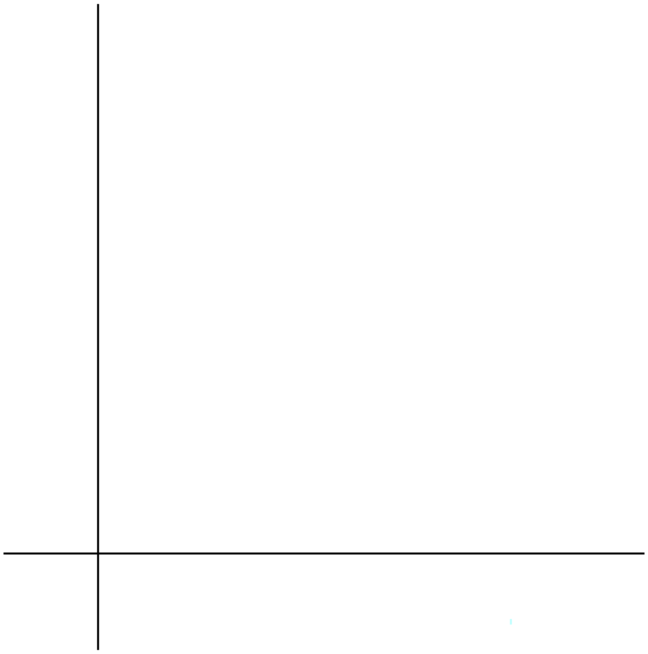


Figure 9.4

b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Glossary

Binomial Distribution: A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, *n*, of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: **$X \sim B(n, p)$**. The mean is μ = np and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly *x* successes in *n* trials is $P\left(X = x\right) = \binom{n}{x} p^{x} q^{n - x}$.

Central Limit Theorem: Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Confidence Interval (CI): An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing: Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

Hypothesis Testing: Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

Hypothesis Testing: Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

Hypothesis: A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Hypothesis: A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Hypothesis: A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Hypothesis: A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Level of Significance of the Test : Probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

Normal Distribution: A continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called **the standard normal distribution**.

p-value: The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

p-value: The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

p-value: The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

p-value: The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student-t Distribution: Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Student-t Distribution: Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Type 1 Error: The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Type 1 Error: The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Type 2 Error: The decision is to not reject the Null hypothesis when, in fact, the Null hypothesis is false.

10 HYPOTHESIS TESTING: TWO MEANS, PAIRED DATA, TWO PROPORTIONS

10.1 Hypothesis Testing: Two Population Means and Two Population Proportions

Student Learning Objectives

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

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Introduction

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported about various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

In the previous chapter, you learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two averages or two proportions to each other. The general procedure is still the same, just expanded.

To compare two averages or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** mean that the two samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and p-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. Because of technology, we do not need to separate two population means, independent groups, population variances unknown into large and small sample sizes.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Becomes a test of one population mean.

10.2 Comparing Two Independent Population Means with Unknown Population Standard Deviations

1. The two independent samples are simple random samples from two distinct populations.
2. Both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30. In that case, the populations need not be normally distributed.

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

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

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \tag{10.1}$$

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

T-score

$$\frac{\left(\overline{x_1}-\overline{x_2}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\left(s_1\right)^2}{n_1}+\frac{\left(s_2\right)^2}{n_2}}} \tag{10.2}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\overline{x_1}$ and $\overline{x_2}$ are the sample means. μ_1 and μ_2 are the population means.

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

Degrees of freedom

$$\text{df} = \frac{\left[\frac{\left(s_1\right)^2}{n_1}+\frac{\left(s_2\right)^2}{n_2}\right]^2}{\frac{1}{n_1-1}\cdot\left[\frac{\left(s_1\right)^2}{n_1}\right]^2+\frac{1}{n_2-1}\cdot\left[\frac{\left(s_2\right)^2}{n_2}\right]^2} \tag{10.3}$$

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

It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example 10.1 Independent groups

The average amount of time boys and girls ages 7 through 11 spend playing sports each day is believed to be the same. An experiment is done, data is collected, resulting in the table below:

Table 10.1

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2 hours	$\sqrt{0.75}$
Boys	16	3.2 hours	1.00

Is there a difference in the average amount of time boys and girls ages 7 through 11 play sports each day? Test at the 5% level of significance.

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

Random variable: $\overline{X_g}-\overline{X_b}$ = difference in the average amount of time girls and boys play sports each day.

H_o: $\mu_g = \mu_b\left(\mu_g - \mu_b = 0\right)$

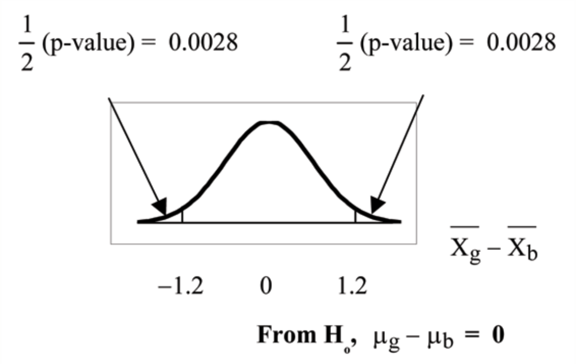
H_a: $\mu_g \neq \mu_b\left(\mu_g - \mu_b \neq 0\right)$

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

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

Calculate the p-value using a Student-t distribution: p-value = 0.0054

Graph:



$s_g = \sqrt{0.75}$

$s_b = 1$

So, $\overline{x_g} - \overline{x_b} = 2 - 3.2 = -1.2$

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

Make a decision: Since α > p-value, reject H_o.

This means you reject $\mu_g = \mu_b$. The means are different.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the average number of hours that girls and boys aged 7 through 11 play sports per day is different.

TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `4:2-SampTTest`. Arrow over to Stats and press `ENTER`. Arrow down and enter `2` for the first sample mean, `√0.75` for Sx1, `9` for n1, `3.2` for the second sample mean, `1` for Sx2, and `16` for n2. Arrow down to μ1: and arrow to `does not equal` μ2. Press `ENTER`. Arrow down to Pooled: and `No`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.0054, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Example 10.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples 9 graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from college A **has taken more math classes,** on the average. Test at a 1% significance level. Answer the following questions.

Is this a test of two means or two proportions?

two means

Are the populations standard deviations known or unknown?

unknown

Which distribution do you use to perform the test?

Student-t

What is the random variable?

$\overline{X_A} - \overline{X_B}$

What are the null and alternate hypothesis?

- $H_o : \mu_A \le \mu_B$
- $H_a : \mu_A > \mu_B$

Is this test right, left, or two tailed?

right

What is the p-value?

0.1928

Do you reject or not reject the null hypothesis?

Do not reject.

Conclusion:

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

10.3 Comparing Two Independent Population Means with Known Population Standard Deviations

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The distribution is Normal and is for the difference of sample means, $\overline{X_1} - \overline{X_2}$. The normal distribution has the following format:

Normal distribution

$$\overline{X_1} - \overline{X_2} \sim N\left[u_1 - u_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{10.4}$$

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \tag{10.5}$$

The test statistic (z-score) is:

$$z = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{10.6}$$

Example 10.3

independent groups, population standard deviations known: The mean lasting time of 2 competing floor waxes is to be compared. **Twenty floors** are randomly assigned **to test each wax**. The following table is the result.

Table 10.2

Wax	Sample Mean Number of Months Floor Wax Last	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\overline{X_1} - \overline{X_2} =$ difference in the average number of months the competing floor waxes last.

$H_o : \mu_1 \leq \mu_2$

$H_a : \mu_1 > \mu_2$

The words **"is more effective"** says that **wax 1 lasts longer than wax 2**, on the average. "Longer" is a " > " symbol and goes into H_a. Therefore, this is a right-tailed test.

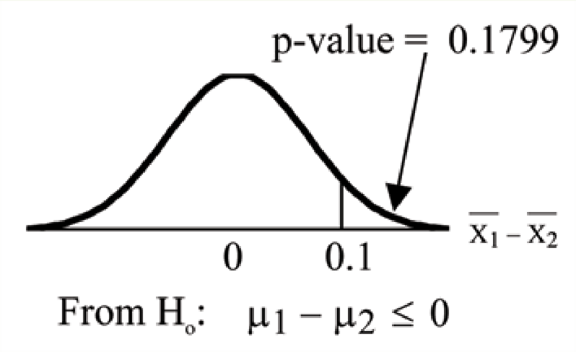
Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula above, the distribution is:

$$\overline{X_1} - \overline{X_2} \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right)$$

Since $\mu_1 \leq \mu_2$ then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is 0.

Calculate the p-value using the normal distribution: p-value = 0.1799

Graph:



$\overline{x_1} - \overline{x_2} = 3 - 2.9 = 0.1$

Compare α and the p-value: α = 0.05 and p-value = 0.1799. Therefore, α < p-value.

Make a decision: Since α < p-value, do not reject H_o.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that wax 1 lasts longer (wax 1 is more effective) than wax 2.

TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `3:2-SampZTest`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter `.33` for sigma1, `.36` for sigma2, `3` for the first sample mean, `20` for n1, `2.9` for the second sample mean, and `20` for n2. Arrow down to μ1: and arrow to > μ2. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.1799 and the test statistic is 0.9157. Do the procedure again but instead of `Calculate` do `Draw`.

10.4 Comparing Two Independent Population Proportions

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five and the number of failures is at least five for each of the samples.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions $\left(P'_A - P'_B\right)$ reflects a difference in the populations.

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

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{10.7}$$

The distribution for the differences is:

$$P'_A - P'_B \sim N\left[0, \sqrt{p_c \cdot \left(1 - p_c\right) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{10.8}$$

The test statistic (z-score) is:

$$z = \frac{\left(p'_A - p'_B\right) - \left(p_A - p_B\right)}{\sqrt{p_c \cdot \left(1 - p_c\right) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{10.9}$$

Example 10.4 Two population proportions

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

Determining the solution

This is a test of 2 population proportions.

How do you know?

The problem asks for a difference in percentages.

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

Random Variable:

$P'_A - P'_B =$ difference in the percentages of adult patients who did not react after 30 minutes to medication A and medication B.

$H_o : p_A = p_B$ $\qquad$ $p_A - p_B = 0$

$H_a : p_A \neq p_B$ $\qquad$ $p_A - p_B \neq 0$

The words **"is a difference"** tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 0.08 \qquad 1 - p_c = 0.92$$

Therefore, $\quad P'_A - P'_B \sim N\left[0, \sqrt{\left(0.08\right) \cdot \left(0.92\right) \cdot \left(\frac{1}{200} + \frac{1}{200}\right)}\right]$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution: p-value = 0.1404.

Estimated proportion for group A: $\quad p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $\quad p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

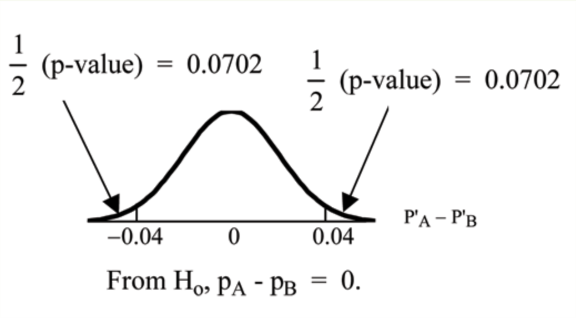


Figure 10.1

$P'_A - P'_B = 0.1 - 0.06 = 0.04$.

Half the p-value is below -0.04 and half is above 0.04.

Compare α and the p-value: α = 0.01 and the p-value = 0.1404. α < p-value.

Make a decision: Since α < p-value, you cannot reject H_0.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the percentages of adult patients who did not react after 30 minutes to medication A and medication B.

TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `6:2-PropZTest`. Arrow down and enter `20` for x1, `200` for n1, `12` for x2, and `200` for n2. Arrow down to `p1`: and arrow to `not equal p2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.1404 and the test statistic is 1.47. Do the procedure again but instead of `Calculate` do `Draw`.

10.5 Matched or Paired Samples

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. The matched pairs have differences that either come from a population that is normal or the number of differences is greater than 30 or both.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student-t test for a single population mean with $n - 1$ degrees of freedom where n is the number of differences.

The test statistic (t-score) is:

$$t = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{10.10}$$

Example 10.5 Matched or paired samples

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table. The "before" value is matched to an "after" value.

Table 10.3

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Corresponding "before" and "after" values form matched pairs.

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data **for the test** are the differences: {0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6}

The sample mean and sample standard deviation of the differences are: $\overline{x}_d$ = -3.13 and s_d = 2.91 Verify these values.

Let μ_d be the population mean for the differences. We use the subscript *d* to denote "differences."

Random Variable: $\overline{X}_d$ = the average difference of the sensory measurements

$$H_o : \mu_d \geq 0$$

There is no improvement. ($\mu_d$is the population mean of the differences.)

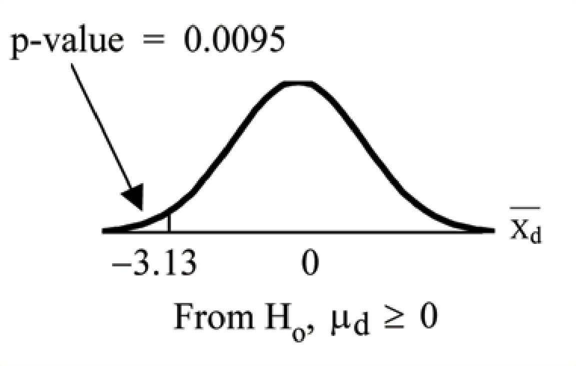
$$H_a : \mu_d < 0$$

There is improvement. The score should be lower after hypnotism so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a student-t with df = $n - 1 = 8 - 1 = 7$. Use t_7. **(Notice that the test is for a single population mean.)**

Calculate the p-value using the Student-t distribution: p-value = 0.0095

Graph:



$\overline{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\overline{x}_d$ = -3.13

$\overline{s}_d$ = 2.91

Compare α and the p-value: α = 0.05 and p-value = 0.0095. α > p-value.

Make a decision: Since α > p-value, reject H_o.

This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction and you will have the differences in the third list.

TI-83+ and TI-84: Use your list of differences as the data. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to μ: and arrow over to < μ_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is 0.0094 and the test statistic is -3.04. Do these instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`.

Example 10.6

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked 4 of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Table 10.4

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weighted lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}

Using the differences data, calculate the sample mean and the sample standard deviation.

$\bar{x}_d = 21.3$ $s_d = 46.7$

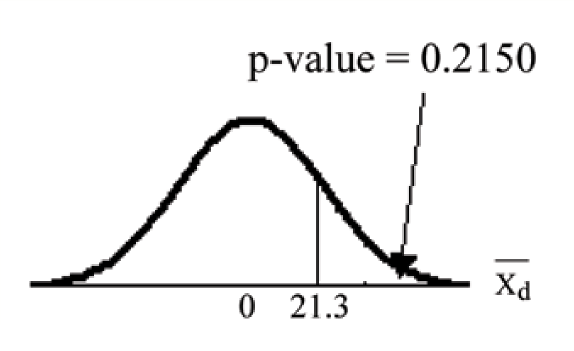
Using the difference data, this becomes a test of a single __________ (fill in the blank).

Define the random variable: $\bar{X}_d$ = average difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

$H_o : \mu_d \leq 0$ $H_a : \mu_d > 0$

Graph:



Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is to not reject the null hypothesis because α < p-value.

What is the conclusion?

means; At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Example 10.7

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The following data was collected.

Table 10.5

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Conduct a hypothesis test to determine whether the differences in distances between the children's dominant versus weaker hands is significant.

Hint

use a t-test on the difference data.

Check

The test statistic is 2.18 and the p-value is 0.0716.

What is your conclusion?

H_0: μ_d equals 0; H_a: μ_d does not equal 0; Do not reject the null; At a 5% significance level, from the sample data, there is not sufficient evidence to conclude that the differences in distances between the children's dominant versus weaker hands is significant (there is not sufficient evidence to show that the children could push the shot-put further with their dominant hand). Alpha and the p-value are close so the test is not strong.

10.6 Summary of Types of Hypothesis Tests

Two Population Means

- Populations are independent and population standard deviations are unknown.
- Populations are independent and population standard deviations are known (not likely).

Matched or Paired Samples

- Two samples are drawn from the same set of objects.
- Samples are dependent.

Two Population Proportions

- Populations are independent.

10.7 Practice 1: Hypothesis Testing for Two Proportions

Student Learning Outcomes

- The student will explore the properties of hypothesis testing with two proportions.

Given

In the 2000 Census, 2.4 percent of the U.S. population reported being two or more races. However, the percent varies tremendously from state to state. (http://www.census.gov/prod/2001pubs/c2kbr01-6.pdf) Suppose that two random surveys are conducted. In the first random survey, out of 1000 North Dakotans, only 9 people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Hypothesis Testing: Two Proportions

Discussion Question

10.8 Practice 2: Hypothesis Testing for Two Averages

Student Learning Outcome

- The student will explore the properties of hypothesis testing with two averages.

Given

The U.S. Center for Disease Control reports that the average life expectancy for whites born in 1900 was 47.6 years and for nonwhites it was 33.0 years. (http://www.cdc.gov/nchs/data/dvs/nvsr53_06t12.pdf) Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the average life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the average life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the average life spans in the county were the same for whites and nonwhites.

Hypothesis Testing: Two Averages

Discussion Question

10.9 Homework

For questions **Exercise 0.0** - **Exercise 0.0**, indicate which of the following choices best identifies the hypothesis test.

A. Independent group means, population standard deviations and/or variances known

B. Independent group means, population standard deviations and/or variances unknown

C. Matched or paired samples

D. Single mean

E. 2 proportions

F. Single proportion

For each problem below, fill in a hypothesis test solution sheet. The solution sheet is in the Appendix and can be copied. For the online version of the book, it is suggested that you copy the .doc or .pdf files.

> If you are using a student-t distribution for a homework problem below, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Questions **Exercise 0.0** – **Exercise 0.0** refer to the Terri Vogel's data set (see Table of Contents).

Try these multiple choice questions.

For questions **Exercise 0.0** – **Exercise 0.0**, use the following information.

A new AIDS prevention drugs was tried on a group of 224 HIV positive patients. Forty-five (45) patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t= treated patient and ut= untreated patient.

For questions **Exercise 0.0** – **Exercise 0.0**, use the following information.

An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a "biofeedback exercise program." Six (6) subjects were randomly selected and the blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated $\left(after - before \right)$ producing the following results: $\overline{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training,

For questions **Exercise 0.0**– **Exercise 0.0**, use the following information.

The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. As of May 25, 2005, the Reserve Division teams scored the following number of goals for 2005.

Table 10.6

Western	Eastern
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to determine if the Western Reserve Division teams score, on average, fewer goals than the Eastern Reserve Division teams. Subscripts: **1** Western Reserve Division (**W**); **2** Eastern Reserve Division (**E**)

Questions **Exercise 0.0** – **Exercise 0.0** refer to the following.

A researcher is interested in determining if a certain drug vaccine prevents West Nile disease. The vaccine with the drug is administered to 36 people and another 36 people are given a vaccine that does not contain the drug. Of the group that gets the vaccine with the drug, one (1) gets West Nile disease. Of the group that gets the vaccine without the drug, three (3) get West Nile disease. Conduct a hypothesis test to determine if the proportion of people that get the vaccine without the drug and get West Nile disease is more than the proportion of people that get the vaccine with the drug and get West Nile disease.

- "Drug" subscript: group who get the vaccine with the drug.
- "No Drug" subscript: group who get the vaccine without the drug

Questions **Exercise 0.0** and **Exercise 0.0** refer to the following:

A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four (4) new students. She records their 18-holes scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

Table 10.7

	Player 1	Player 2	Player 3	Player 4
Average score before class	83	78	93	87
Average score after class	80	80	86	86

Questions **Exercise 0.0** and **Exercise 0.0** refer to the following:

Suppose a statistics instructor believes that there is no significant difference between the average class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The average and standard deviation for 35 statistics day students were 75.86 and 16.91. The average and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students.

10.10 Review

The next three questions refer to the following information:

In a survey at Kirkwood Ski Resort the following information was recorded:

Table 10.8 Sport Participation by Age

	0 – 10	11 - 20	21 - 40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Suppose that one person from of the above was randomly selected.

The next three questions refer to the following information:

A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

The next three questions refer to the following information:

At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of 5 minutes and a standard deviation of 2 minutes.

10.11 Lab: Hypothesis Testing for Two Means and Two Proportions

Class Time:

Names:

Student Learning Outcomes:

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Supplies:

- The business section from two consecutive days’ newspapers
- 3 small packages of M&Ms®
- 5 small packages of Reese's Pieces®

Increasing Stocks Survey

Look at yesterday’s newspaper business section. Conduct a hypothesis test to determine if the proportion of New York Stock Exchange (NYSE) stocks that increased is greater than the proportion of NASDAQ stocks that increased. As randomly as possible, choose 40 NYSE stocks and 32 NASDAQ stocks and complete the following statements.

1. H_o
2. H_a
3. In words, define the Random Variable. ____________=
4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

   a. Graph:

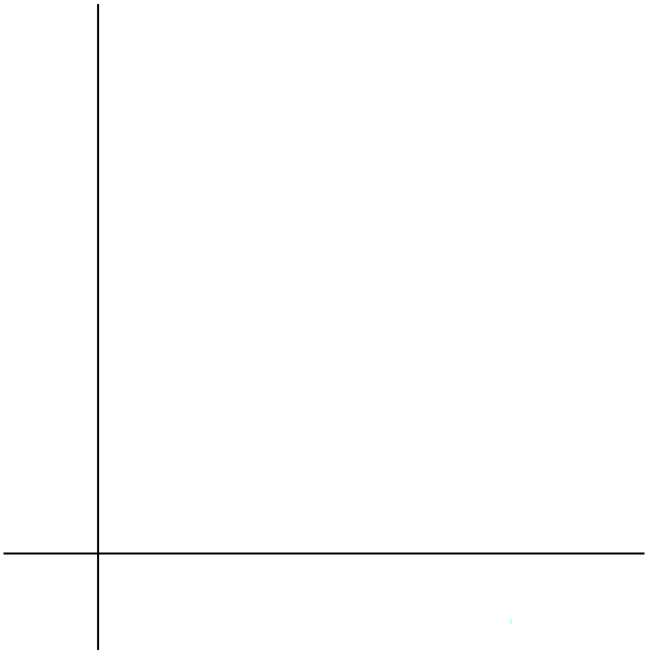


Figure 10.2

b. Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Decreasing Stocks Survey

Randomly pick 8 stocks from the newspaper. Using two consecutive days' business sections, test whether the stocks went down, on average, for the second day.

1. H_o
2. H_a
3. In words, define the Random Variable. ____________=
4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:

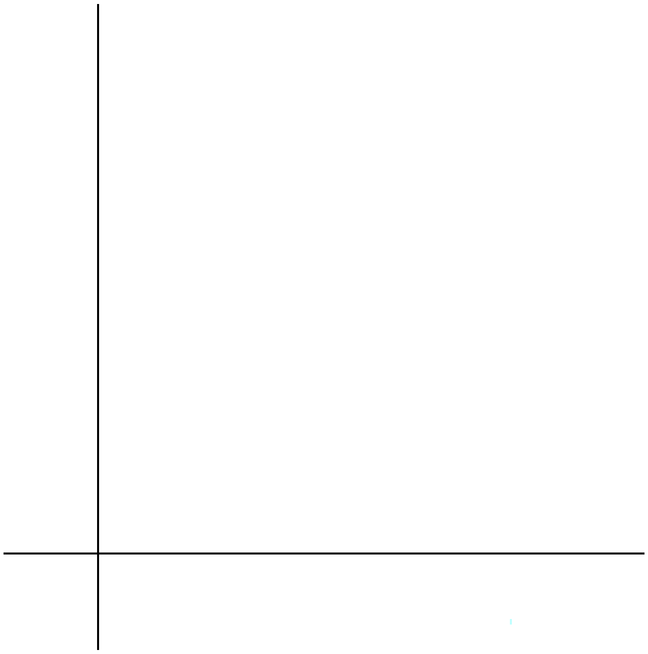


Figure 10.3

b. Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Candy Survey

Buy three small packages of M&Ms and 5 small packages of Reese's Pieces (same net weight as the M&Ms). Test whether or not the average number of candy pieces per package is the same for the two brands.

1. H_o:
2. H_a:
3. In words, define the random variable. __________=
4. What distribution should be used for this test?
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:

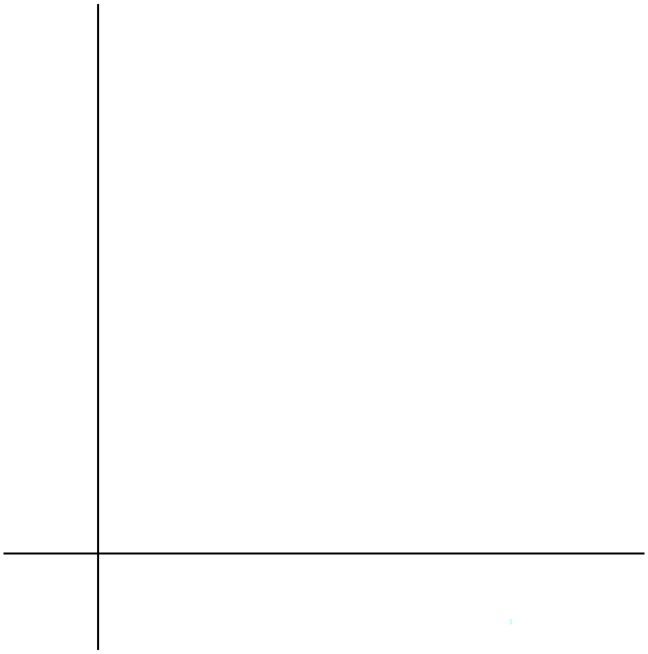


Figure 10.4

b. Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Shoe Survey

Test whether women have, on average, more pairs of shoes than men. Include all forms of sneakers, shoes, sandals, and boots. Use your class as the sample.

1. H_o
2. H_a
3. In words, define the Random Variable. ____________=
4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

   a. Graph:

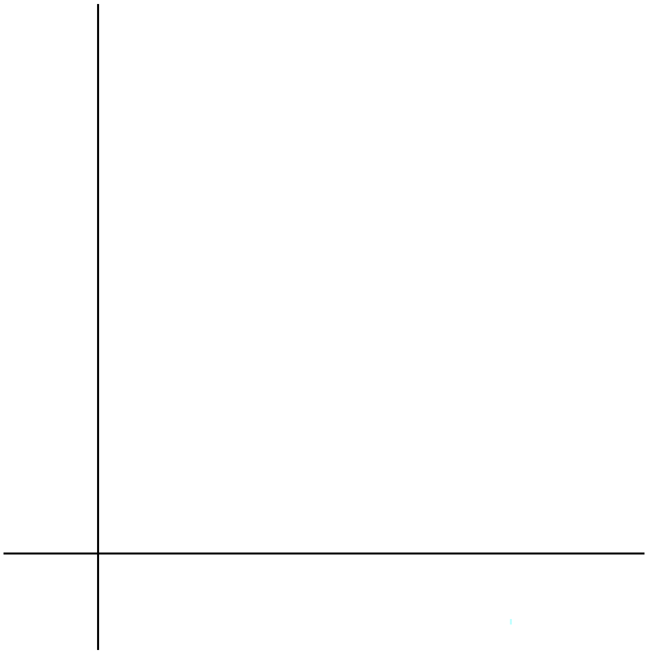


Figure 10.5

b. Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Glossary

Degrees of Freedom (df): The number of objects in a sample that are free to vary.

Standard Deviation: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variable (Random Variable): A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

11 THE CHI-SQUARE DISTRIBUTION

11.1 The Chi-Square Distribution

Student Learning Objectives

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

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests (optional).

Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to the above examples. This distribution is called the Chi-square distribution.

In this chapter, you will learn the three major applications of the Chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as with the lottery example
- The test of independence, which determines if events are independent, such as with the movie example
- The test of a single variance, which tests variability, such as with the coffee example

> Though the Chi-square calculations depend on calculators or computers for most of the calculations, there is a table available (see the Table of Contents **15. Tables**). TI-83+ and TI-84 calculator instructions are included in the text.

Optional Collaborative Classroom Activity

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, etc.). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

11.2 Notation

The notation for the chi-square distribution is:

$$\chi^2 \sim \chi^2_{df}$$

where df = degrees of freedom depend on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use df = $n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2 \cdot df}$.

The random variable is shown as χ^2 but may be any upper case letter.

The random variable for a chi-square distribution with k degrees of freedom is the sum of ***k*** independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + ... + (Z_k)^2$$

11.3 Facts About the Chi-Square Distribution

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.

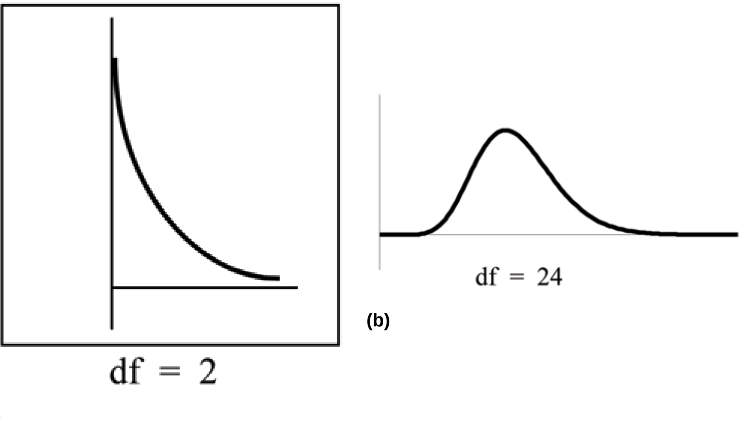


Figure 11.1

3. The test statistic for any test is always greater than or equal to zero.
4. When df > 90, the chi-square curve approximates the normal. For $X \sim \chi^2_{1000}$ the mean, μ = df = 1000 and the standard deviation, $\sigma = \sqrt{2 \cdot 1000} = 44.7$. Therefore, $X \sim N(1000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.

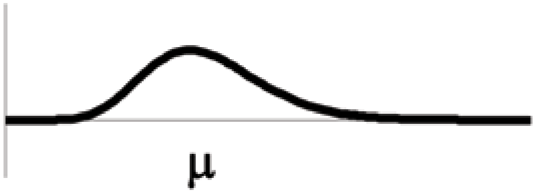


Figure 11.2

11.4 Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:

$$\sum_{n} \frac{(O-E)^2}{E} \tag{11.1}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- n = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true.

There are n terms of the form $\frac{(O-E)^2}{E}$.

The degrees of freedom are df = (number of categories - 1).

The goodness-of-fit test is almost always right tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Example 11.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Three statistics instructors wondered whether the absentee rate was the **same** for every day of the school week. They took a sample of absent students from three of their statistics classes during one week of the term. The results of the survey appear in the table.

Table 11.1

	Monday	Tuesday	Wednesday	Thursday	Friday
# of students absent	28	22	18	20	32

Determine the null and alternate hypotheses needed to run a goodness-of-fit test.

Since the instructors wonder whether the absentee rate is the same for every school day, we could say in the null hypothesis that the data **"fit"** a uniform distribution.

***H_o*: The rate at which college students are absent from their statistics class fits a uniform distribution.

The alternate hypothesis is the opposite of the null hypothesis.

***H_a*: The rate at which college students are absent from their statistics class does not fit a uniform distribution.

How many students do you **expect** to be absent on any given school day?

The total number of students in the sample is 120. **If the null hypothesis were true,** you would divide 120 by 5 to get 24 absences expected per day. **The expected number is based on a true null hypothesis.**

What are the degrees of freedom (df)?

There are 5 days of the week or 5 "cells" or categories.

df = no. cells - 1 = 5 - 1 = 4

Example 11.2

Employers particularly want to know which days of the week employees are absent in a five day work week. Most employers would like to believe that employees are absent equally during the week. That is, the average number of times an employee is absent is the same on Monday, Tuesday, Wednesday, Thursday, or Friday. Suppose a sample of 20 absent days was taken and the days absent were distributed as follows:

Table 11.2 Day of the Week Absent

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	5	4	2	3	6

For the population of employees, do the absent days occur with equal frequencies during a five day work week? Test at a 5% significance level.

The null and alternate hypotheses are:

- H_o: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 20 absent days, there would be 4 absences on Monday, 4 on Tuesday, 4 on Wednesday, 4 on Thursday, and 4 on Friday. These numbers are the **expected** (*E*) values. The values in the table are the **observed** (*O*) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings:

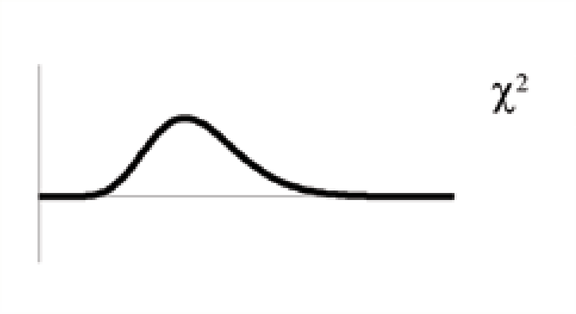
- Expected (*E*) values
- Observed (*O*) values
- $(O-E)$
- $(O-E)^2$
- $\frac{(O-E)^2}{E}$

Now add (sum) the last column. Verify that the sum is 2.5. This is the χ^2 test statistic.

To find the p-value, calculate $P\left(\chi^2 > 2.5\right)$. This test is right-tailed.

The dfs are the number of cells−1 = 4.

Next, complete a graph like the one below with the proper labeling and shading. (You should shade the right tail. It will be a "large" right tail for this example because the p-value is "large.")



Use a computer or calculator to find the p-value. You should get p-value = 0.6446.

The decision is to not reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

TI-83+ and TI-84: Press `2nd DISTR`. Arrow down to χ^2`cdf`. Press `ENTER`. Enter `(2.5,1E99,4)`. Rounded to 4 places, you should see 0.6446 which is the p-value.

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example (Example 11-3) has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter whatever else is asked and press `calculate` or `draw`. Make sure you clear any lists before you start. See below.

To Clear Lists in the calculators: Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Or, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER`.

Example 11.3

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Table 11.3

Number of Televisions	Percent
0	10
1	16
2	55
3	11
over 3	8

The table contains expected (*E*) percents.

A random sample of 600 families in the far western United States resulted in the following data:

Table 11.4

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
over 3	15
	Total = 600

The table contains observed (O) frequency values.

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are:

Number of Televisions	Percent	Expected Frequency
0	10	$(0.10) \cdot (600) = 60$
1	16	$(0.16) \cdot (600) = 96$
2	55	$(0.55) \cdot (600) = 330$
3	11	$(0.11) \cdot (600) = 66$
over 3	8	$(0.08) \cdot (600) = 48$

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter .10*600.

H_o: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

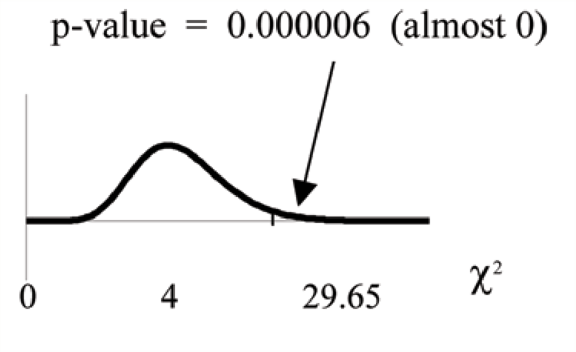
H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where df = (the number of cells) − 1 = 5 − 1 = 4.

df ≠ 600 − 1

Calculate the test statistic: $\chi^2 = 29.65$

Graph:



Probability statement: p-value = $P(\chi^2 > 29.65) = 0.000006$.

Compare α and the p-value:

- α = 0.01
- p-value = 0.000006

So, α > p-value.

Make a decision: Since α > p-value, reject H_o.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

TI-83+ and some TI-84 calculators: Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example 11-2). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600, .16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:χ2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to 4 places, you should see `5.77E-6 = .000006` (rounded to 6 decimal places) which is the p-value.

Example 11.4

Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let *X* = the number of heads in one flip of the two coins. *X* takes on the value 0, 1, 2. (There are 0, 1, or 2 heads in the flip of 2 coins.) Therefore, the **number of cells is 3**. Since *X* = the number of heads, the observed frequencies are 20 (for 2 heads), 57 (for 1 head), and 23 (for 0 heads or both tails). The expected frequencies are 25 (for 2 heads), 50 (for 1 head), and 25 (for 0 heads or both tails). This test is right-tailed.

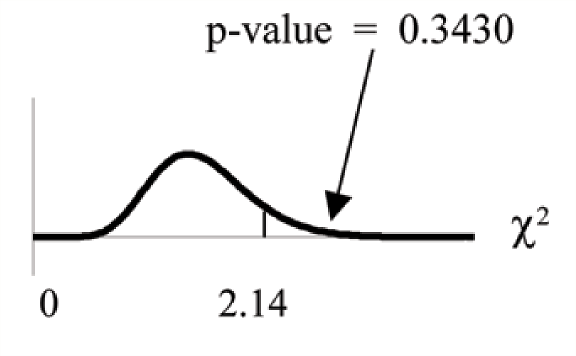
H_o: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where df = 3 − 1 = 2.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



Probability statement: p-value = $P\left(\chi^2 > 2.14\right) = 0.3430$

Compare α and the p-value:

- α = 0.05
- p-value = 0.3430

So, α < p-value.

Make a decision: Since α < p-value, do not reject H_o.

Conclusion: The coins are fair.

TI-83+ and some TI- 84 calculators: Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum".Enter L3`. Rounded to 2 decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press 7). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to 4 places, you should see `.3430` which is the p-value.

For the newer TI-84 calculators, check `STAT TESTS` to see if you have `Chi2 GOF`. If you do, see the calculator instructions (a NOTE) before Example 11-3

11.5 Test of Independence

Tests of independence involve using a **contingency table** of observed (data) values. You first saw a contingency table when you studied probability in the **Probability Topics** chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{11.2}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

Example 11.5

Suppose A = a speeding violation in the last year and B = a car phone user. If A and B are independent then $P(A\text{AND}B) = P(A)P(B)$. $A\text{AND}B$ is the event that a driver received a speeding violation last year and is also a car phone user. Suppose, in a study of drivers who received speeding violations in the last year and who use car phones, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were car phone users and 450 were not.

Let y = expected number of car phone users who received speeding violations.

If A and B are independent, then $P(A\text{AND}B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$$

Solve for y : $y = \frac{70 \cdot 305}{755} = 28.3$

About 28 people from the sample are expected to be car phone users and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example above, then the null hypothesis is:

H_o: Being a car phone user and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be car phone users and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

The degrees of freedom for the test of independence are:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the **expected number** (E):

$$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}}$$

Example 11.6

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table 11.5 Number of Hours Worked Per Week by Volunteer Type (Observed) The table contains **observed (O)** values (data).

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Are the number of hours volunteered **independent** of the type of volunteer?

The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_o: The number of hours volunteered is **independent** of the type of volunteer.

H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected table is:

Number of Hours Worked Per Week by Volunteer Type (Expected) The table contains **expected** (*E*) values (data).

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

For example, the calculation for the expected frequency for the top left cell is

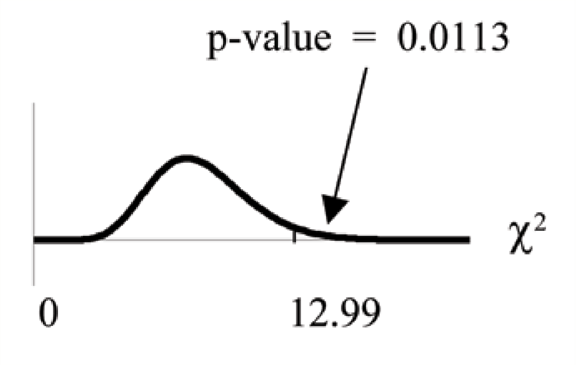
$$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}} = \frac{255 \cdot 298}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

df = (3 columns − 1)(3 rows − 1) = (2)(2) = 4

Graph:



Probability statement: p-value = $P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume α = 0.05. p-value = 0.0113. α > p-value.

Make a decision: Since α > p-value, reject H_0. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Example 11-6. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the p-value = 0.0113. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Example 11.7

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Table 11.6 Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

How many high anxiety level students are expected to have a high need to succeed in school?

The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

a. $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} =$

b. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

a. $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} = 8.19$

b. 8

11.6 Test of a Single Variance (Optional)

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternate hypotheses are stated in terms of the **population variance** (or population standard deviation). The test statistic is:

$$\frac{\left(n-1\right)\cdot s^2}{\sigma^2} \tag{11.3}$$

where:

- n = the total number of data
- s^2 = sample variance
- σ^2 = population variance

You may think of s as the random variable in this test. The degrees of freedom are df = $n - 1$.

A test of a single variance may be right-tailed, left-tailed, or two-tailed.

The following example will show you how to set up the null and alternate hypotheses. The null and alternate hypotheses contain statements about the population variance.

Example 11.8

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is 5 points. One of his best students thinks otherwise. The student claims that the standard deviation is more than 5 points. If the student were to conduct a hypothesis test, what would the null and alternate hypotheses be?

Even though we are given the population standard deviation, we can set the test up using the population variance as follows.

- H_o: $\sigma^2 = 5^2$
- H_a: $\sigma^2 > 5^2$

Example 11.9

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5%, test the claim that **a single line causes lower variation among waiting times (shorter waiting times) for customers**.

Since the claim is that a single line causes lower variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

- H_o: $\sigma^2 = 7.2^2$
- H_a: $\sigma^2 < 7.2^2$

The word **"lower"** tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where:

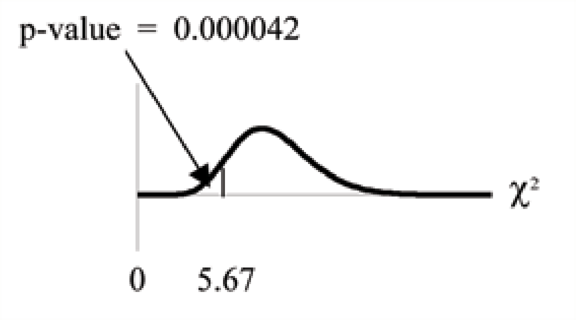
- n = the number of customers sampled
- df = $n - 1 = 25 - 1 = 24$

Calculate the test statistic:

$$\chi^2 = \frac{\left(n-1\right)\cdot s^2}{\sigma^2} = \frac{\left(25-1\right)\cdot 3.5^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph:



Probability statement: p-value = $P\left(\chi^2 < 5.67\right) = 0.000042$

Compare α and the p-value: α = 0.05   p-value = 0.000042   α > p-value

Make a decision: Since α > p-value, reject H_o.

This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes, but lower.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times **or** with a single line, the customer waiting times vary less than 7.2 minutes.

TI-83+ and TI-84 calculators: In `2nd DISTR`, use `7:χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example 11-9, `χ2cdf(-1E99,5.67,24)`. The p-value = 0.000042.

11.7 Summary of Formulas

Formula

$\mu = \text{df}$ and $\sigma = \sqrt{2 \cdot \text{df}}$

Formula

- Use goodness-of-fit to test whether a data set fits a particular probability distribution.
- The degrees of freedom are number of cells or categories - 1.
- The test statistic is $\sum_{n} \frac{(O-E)^2}{E}$, where O = observed values (data), E = expected values (from theory), and n = the number of different data cells or categories.
- The test is right-tailed.

Formula

- Use the test of independence to test whether two factors are independent or not.
- The degrees of freedom are equal to (number of columns - 1)(number of rows - 1).
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

Formula

- Use the test to determine variation.
- The degrees of freedom are the number of samples - 1.
- The test statistic is $\frac{\left(n-1\right) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left, right, or two-tailed.

11.8 Practice 1: Goodness-of-Fit Test

Student Learning Outcomes

- The student will explore the properties of goodness-of-fit test data.

Given

The following data are real. The cumulative number of AIDS cases reported for Santa Clara County through December 31, 2003, is broken down by ethnicity as follows:

Table 11.7

Ethnicity	Number of Cases
White	2032
Hispanic	897
African-American	372
Asian, Pacific Islander	168
Native American	20
	Total = 3489

The percentage of each ethnic group in Santa Clara County is as follows:

Table 11.8

Ethnicity	Percentage of total county population	Number expected (round to 2 decimal places)
White	47.79%	1667.39
Hispanic	24.15%	
African-American	3.55%	
Asian, Pacific Islander	24.21%	
Native American	0.29%	
	Total = 100%	

Expected Results

If the ethnicity of AIDS victims followed the ethnicity of the total county population, fill in the expected number of cases per ethnic group.

Goodness-of-Fit Test

Perform a goodness-of-fit test to determine whether the make-up of AIDS cases follows the ethnicity of the general population of Santa Clara County.

Discussion Question

11.9 Practice 2: Contingency Tables

Student Learning Outcomes

- The student will explore the properties of contingency tables.

Conduct a hypothesis test to determine if smoking level and ethnicity are independent.

Collect the Data

Copy the data provided in **Probability Topics Practice 1: Calculating Probabilities** into the table below.

Table 11.9 Smoking Levels by Ethnicity (Observed)

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Hypothesis

State the hypotheses.

- H_o:
- H_a:

Expected Values

Enter expected values in the above below. Round to two decimal places.

Analyze the Data

Calculate the following values:

Graph the Data

Conclusions

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α .

11.10 Practice 3: Test of a Single Variance

Student Learning Outcomes

- The student will explore the properties of data with a test of a single variance.

Given

Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

Sample Variance

Hypothesis Test

Perform a hypothesis test on the consistency part of the claim.

Discussion Questions

11.11 Homework

Word Problems

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to The Table of Contents 14. Appendix for the solution sheet. Round expected frequency to two decimal places.

The next two questions refer to the following information. The columns in the chart below contain the Race/Ethnicity of U.S. Public Schools: High School Class of 2009, the percentages for the Advanced Placement Examinee Population for that class and the Overall Student Population. (*Source: http://www.collegeboard.com*). Suppose the right column contains the result of a survey of 1000 local students from the Class of 2009 who took an AP Exam.

Table 11.10

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American or Pacific Islander	10.2%	5.4%	113
Black or African American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

The next two questions refer to the following information: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of student expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5000 graduating females and 5000 graduating males was done as a follow-up in 2010 to determine what their actual major was. The results are shown in the tables for Exercises 7 and 8. The second column in each table does not add to 100% because of rounding.

Try these true/false questions.

11.12 Review

The next two questions refer to the following real study:

A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12 - 19. 6% of the girls surveyed said they have been pregnant. (*Parade Magazine*) We are interested in the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

The next four questions refer to the following information:

Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of 7 years and a standard deviation of 2 years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

The next five questions refer to the following information:

We are interested in the checking account balance of a twenty-year-old college student. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of $640 and a sample standard deviation of $150. Let X = checking account balance of an individual twenty year old college student.

The next two questions refer to the following information:

The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30 . Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

The next two questions refer to the following information:

70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:

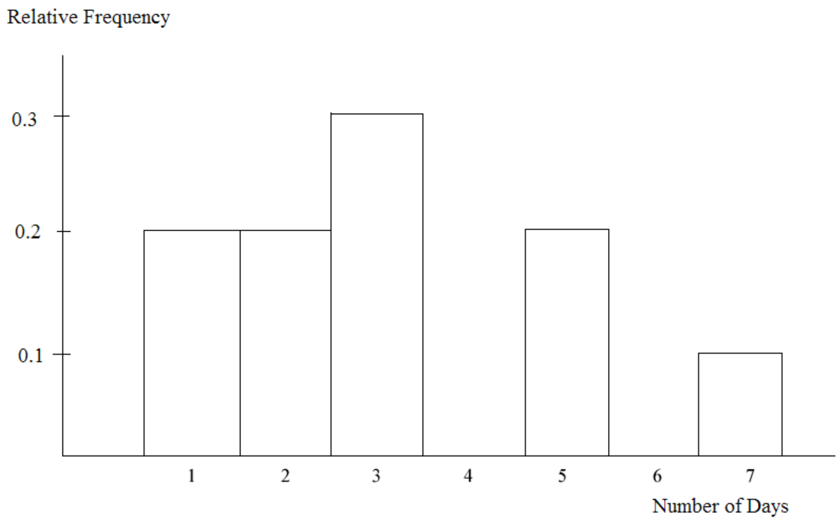


Figure 11.3

The next two questions refer to the following information:

A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

11.13 Lab 1: Chi-Square Goodness-of-Fit

Class Time:

Names:

Student Learning Outcome:

- The student will evaluate data collected to determine if they fit either the uniform or exponential distributions.

Collect the Data

Go to your local supermarket. Ask 30 people as they leave for the total amount on their grocery receipts. (Or, ask 3 cashiers for the last 10 amounts. Be sure to include the express lane, if it is open.)

1. Record the values.

Table 11.11

2. Construct a histogram of the data. Make 5 - 6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

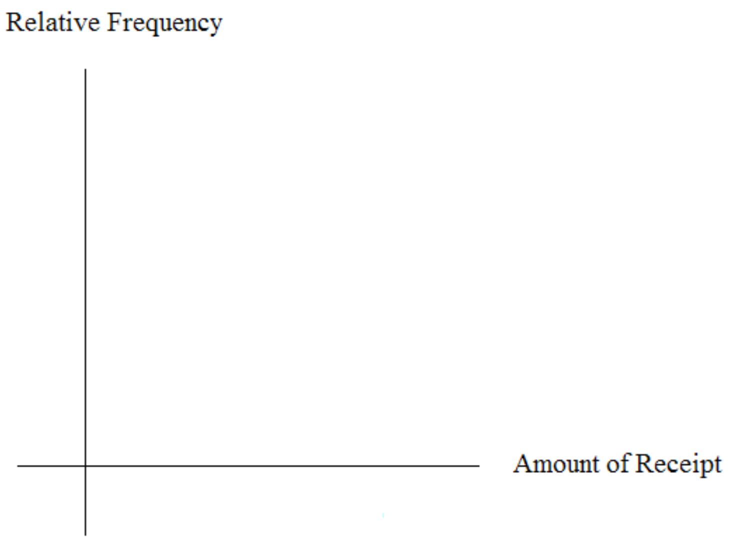


Figure 11.4

3. Calculate the following:

   a. $\overline{X}$ =

   b. S =

   c. S^2 =

Uniform Distribution

Test to see if grocery receipts follow the uniform distribution.

1. Using your lowest and highest values, X ~ U(_______,_______)
2. Divide the distribution above into fifths.
3. Calculate the following:

   a. Lowest value =

   b. 20th percentile =

   c. 40th percentile =

   d. 60th percentile =

   e. 80th percentile =

   f. Highest value =
4. For each fifth, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Table 11.12

Fifth	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		

5. H_o:
6. H_a:
7. What distribution should you use for a hypothesis test?
8. Why did you choose this distribution?
9. Calculate the test statistic.
10. Find the p-value.
11. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

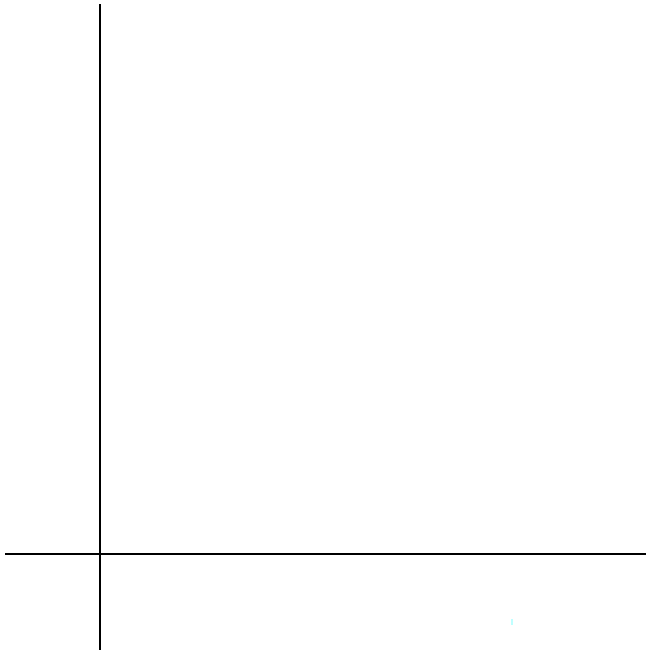


Figure 11.5

12. State your decision.
13. State your conclusion in a complete sentence.

Exponential Distribution

Test to see if grocery receipts follow the exponential distribution with decay parameter $\frac{1}{\bar{x}}$.

1. Using $\frac{1}{\bar{x}}$ as the decay parameter, $X \sim Exp($_______$)$.
2. Calculate the following:
   a. Lowest value =
   b. First quartile =
   c. 37th percentile =
   d. Median =
   e. 63rd percentile =
   f. 3rd quartile =
   g. Highest value =
3. For each cell, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Table 11.13

Cell	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		
6th		

4. H_o
5. H_a
6. What distribution should you use for a hypothesis test?
7. Why did you choose this distribution?

8. Calculate the test statistic.
9. Find the p-value.
10. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

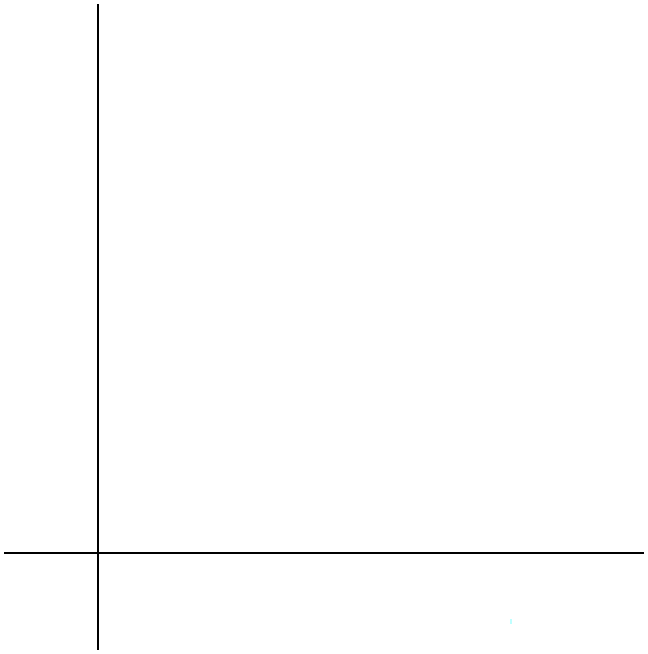


Figure 11.6

11. State your decision.
12. State your conclusion in a complete sentence.

Discussion Questions

1. Did your data fit either distribution? If so, which?
2. In general, do you think it's likely that data could fit more than one distribution? In complete sentences, explain why or why not.

11.14 Lab 2: Chi-Square Test for Independence

Class Time:

Names:

Student Learning Outcome:

- The student will evaluate if there is a significant relationship between favorite type of snack and gender.

Collect the Data

1. Using your class as a sample, complete the following chart.

Table 11.14 Favorite type of snack

	sweets (candy & baked goods)	ice cream	chips & pretzels	fruits & vegetables	Total
male					
female					
Total					

2. Looking at the above chart, does it appear to you that there is dependence between gender and favorite type of snack food? Why or why not?

Hypothesis Test

Conduct a hypothesis test to determine if the factors are independent

1. H_o:
2. H_a:

3. What distribution should you use for a hypothesis test?
4. Why did you choose this distribution?
5. Calculate the test statistic.
6. Find the p-value.
7. Sketch a graph of the situation. Label and scale the x-axis. Shade the area corresponding to the p-value.

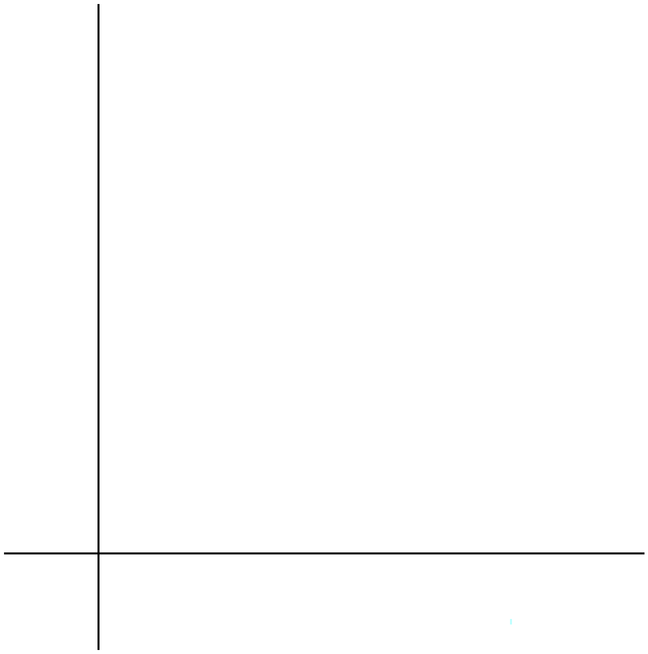


Figure 11.7

8. State your decision.
9. State your conclusion in a complete sentence.

Discussion Questions

1. Is the conclusion of your study the same as or different from your answer to (I2) above?
2. Why do you think that occurred?

Glossary

Contingency Table: The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

12 LINEAR REGRESSION AND CORRELATION

12.1 Linear Regression and Correlation

Student Learning Objectives

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

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Introduction

Professionals often want to know how two or more variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is it and how strong is the relationship?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee. These are all examples in which regression can be used.

The type of data described in the examples is **bivariate** data - "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (*x*). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

12.2 Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:

$$y = a + \text{b}x \tag{12.1}$$

where *a* and *b* are constant numbers.

***x* is the independent variable, and *y* is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 12.1

The following examples are linear equations.

$$y = 3 + 2x \tag{12.2}$$

$$y = -0.01 + 1.2x \tag{12.3}$$

The graph of a linear equation of the form $y = a + \text{b}x$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example 12.2

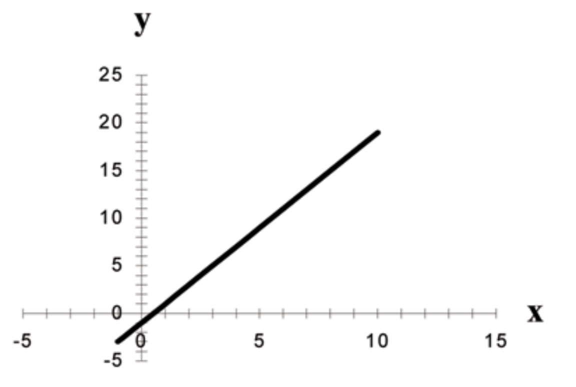


Figure 12.1 Graph of the equation $y = -1 + 2x$.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example 12.3

Aaron's Word Processing Service (AWPS) does word processing. Its rate is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

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

Let y = the total cost to the customer.

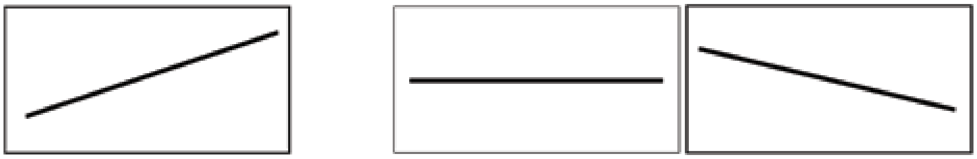
The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is:

$y = 31.50 + 32x$

12.3 Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point (0, a) where the line crosses the y-axis.



(a) If $b > 0$, the line slopes upward to the right. **(b)** If $b = 0$, the line is horizontal. **(c)** If $b < 0$, the line slopes downward to the right.

Figure 12.2 Three possible graphs of $y = a + bx$.

Example 12.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

12.4 Scatter Plots

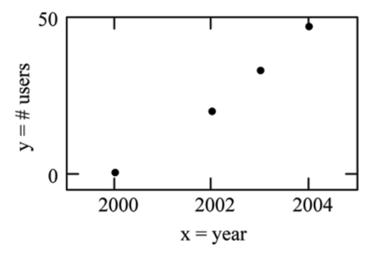
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example 12.5

From an article in the *Wall Street Journal*: In Europe and Asia, m-commerce is becoming more popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. In the next few years, will there be a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table 12.1

x (year)	*y* (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0



(a) Table showing the number of m-commerce users (in millions) by year. **(b)** Scatter plot showing the number of m-commerce users (in millions) by year.

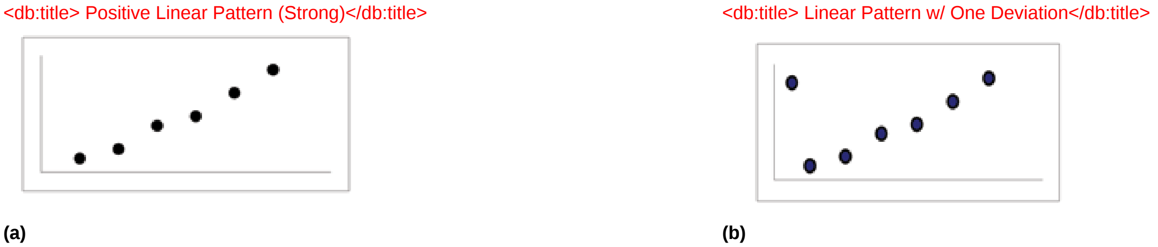
Figure 12.3

A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

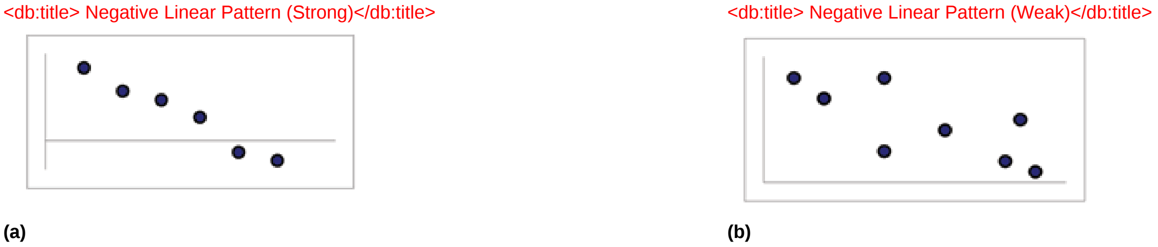


<db:title> Positive Linear Pattern (Strong)</db:title>

<db:title> Linear Pattern w/ One Deviation</db:title>

(a) **(b)**

Figure 12.4

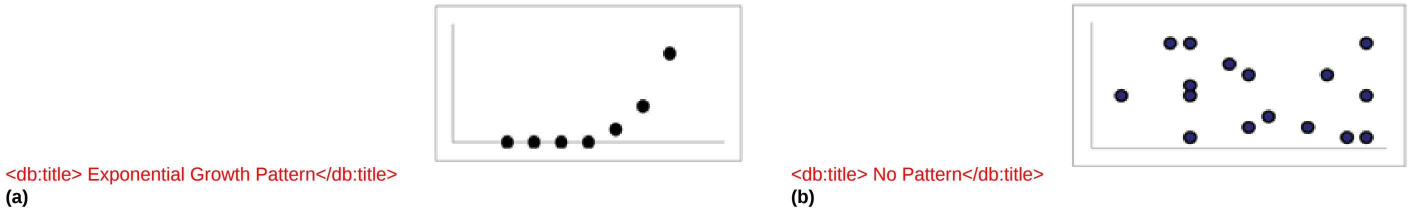


<db:title> Negative Linear Pattern (Strong)</db:title>

<db:title> Negative Linear Pattern (Weak)</db:title>

(a) **(b)**

Figure 12.5



<db:title> Exponential Growth Pattern</db:title>
(a)

<db:title> No Pattern</db:title>
(b)

Figure 12.6

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If *x* is the independent variable and *y* the dependent variable, then we can use a regression line to predict *y* for a given value of *x*.

12.5 The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line**.

Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, *x*, is pinky finger length and the dependent variable, *y*, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

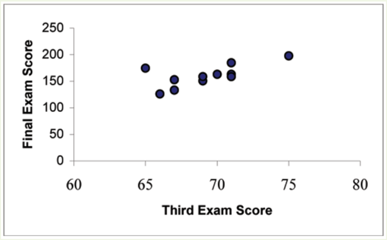
Example 12.6

A random sample of 11 statistics students produced the following data where *x* is the third exam score, out of 80, and *y* is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table 12.2

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

(a) Table showing the scores on the final exam based on scores from the third exam.



(b) Scatter plot showing the scores on the final exam based on scores from the third exam.

Figure 12.7

The third exam score, *x*, is the independent variable and the final exam score, *y*, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (*x*, *y*)and each point of the line of best fit using least-squares linear regression has the form $\left(x, \hat{y}\right)$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of *y***. It is the value of *y* obtained using the regression line. It is not generally equal to *y* from data.

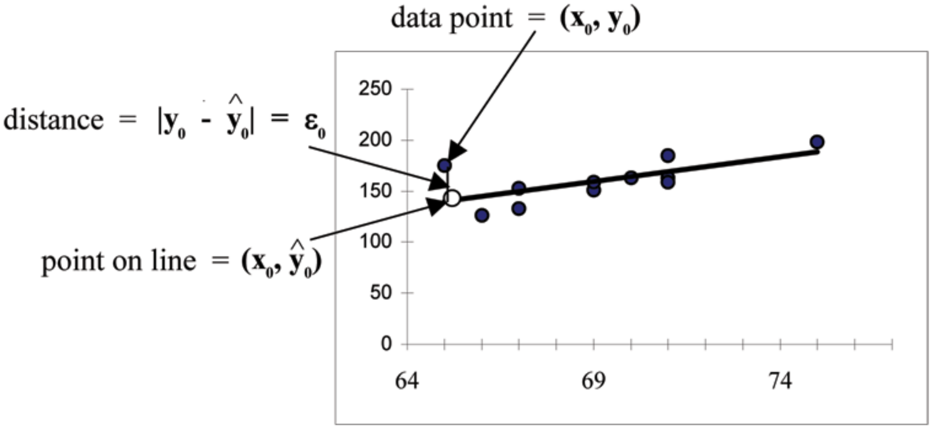


Figure 12.8

The term $\left| y_0 - \hat{y}_0 \right| = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake, but measures the vertical distance between the actual value of *y* and the estimated value of *y*. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for *y*. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for *y*.

In the diagram above, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $\left| y_i - \hat{y}_i \right| = \varepsilon_i$ for i = 1, 2, 3, ..., 11.

Each ε is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$\left(\varepsilon_1\right)^2 + \left(\varepsilon_2\right)^2 + \ldots + \left(\varepsilon_{11}\right)^2 = \sum_{i=1}^{11} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx \qquad (12.4)$$

where $a = \bar{y} - b \cdot \bar{x}$ and $b = \dfrac{\Sigma\left(x - \bar{x}\right) \cdot \left(y - \bar{y}\right)}{\Sigma\left(x - \bar{x}\right)^2}$.

$\bar{x}$ and $\bar{y}$ are the averages of the x values and the y values, respectively. The best fit line always passes through the point $\left(\bar{x}, \bar{y}\right)$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line** .

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:

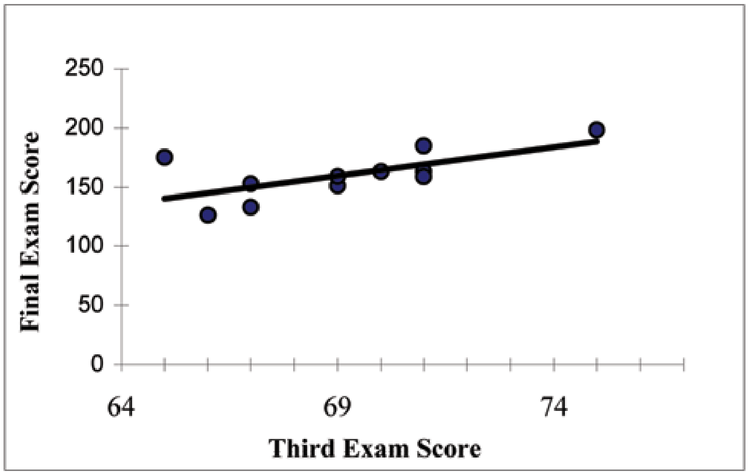


Figure 12.9

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

$$\hat{y} = -173.51 + 4.83x \qquad (12.5)$$

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for *y* given *x* within the domain of *x*-values in the sample data, **but not necessarily for *x*-values outside that domain.**

You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.

You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

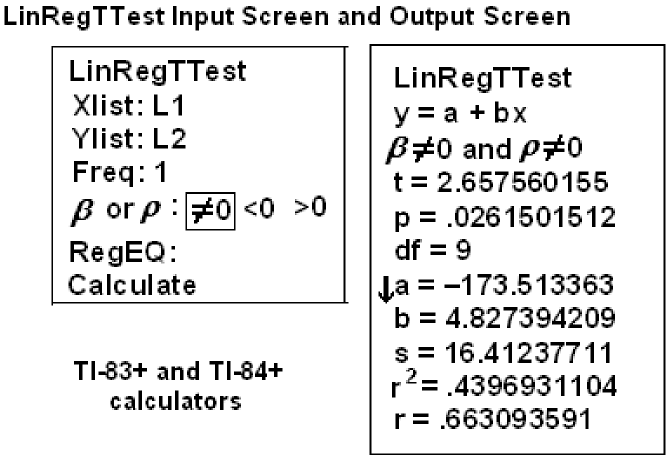
Slope: The slope of the line is b = 4.83.

Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)
3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
4. On the next line, at the prompt β or ρ, highlight "≠ 0" and press ENTER
5. Leave the line for "RegEq:" blank
6. Highlight Calculate and press ENTER.


LinRegTTest Input Screen and Output Screen

```
LinRegTTest
Xlist: L1
Ylist: L2
Freq: 1
β or ρ : ≠0  <0  >0
RegEQ:
Calculate
```

TI-83+ and TI-84+ calculators

```
LinRegTTest
y = a + bx
β≠0 and ρ≠0
t = 2.657560155
p = .0261501512
df = 9
↓a = −173.513363
b = 4.827394209
s = 16.41237711
r² = .4396931104
r = .663093591
```

Figure 12.10

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says y=a+bx. Scroll down to find the values a=-173.513, and b=4.8273 ; the equation of the best fit line is $\hat{y}$ = -173.51 + 4.83*x*

The two items at the bottom are r^2 = .43969 and *r*=.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1
3. On the input screen for PLOT 1, highlight **On** and press ENTER
4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
5. Indicate Xlist: L1 and Ylist: L2
6. For Mark: it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
8. To graph the best fit line, press the "Y=" key and type the equation -173.5+4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

**With contributions from Roberta Bloom

12.6 Correlation Coefficient and Coefficient of Determination

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is a numerical measure of the strength of association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as

$$r = \frac{n \cdot \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y)}{\sqrt{\left[n \cdot \Sigma x^2 - (\Sigma x)^2\right] \cdot \left[n \cdot \Sigma y^2 - (\Sigma y)^2\right]}} \qquad (12.6)$$

where *n* = the number of data points.

If you suspect a linear relationship between *x* and *y*, then *r* can measure how strong the linear relationship is.

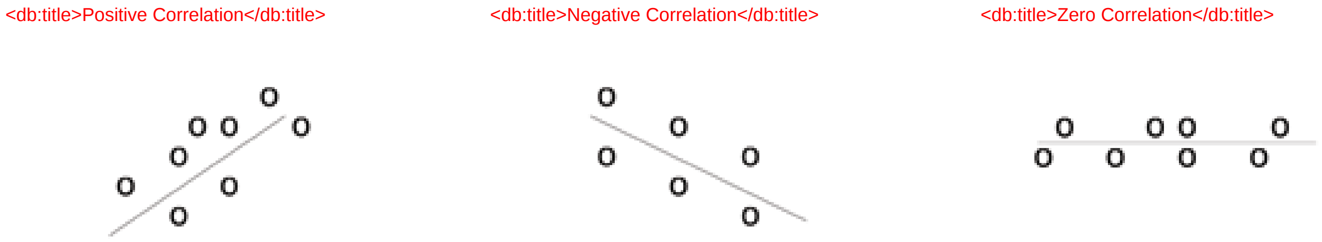
What the VALUE of r tells us:

- The value of *r* is always between -1 and +1: $-1 \le r \le 1$.
- The closer the correlation coefficient *r* is to -1 or 1 (and the further from 0), the stronger the evidence of a significant linear relationship between *x* and *y*; this would indicate that the observed data points fit more closely to the best fit line. Values of *r* further from 0 indicate a stronger linear relationship between *x* and *y*. Values of *r* closer to 0 indicate a weaker linear relationship between *x* and *y*.
- If r=0 there is absolutely no linear relationship between *x* and *y* **(no linear correlation)**.
- If *r* = 1, there is perfect positive correlation. If *r* = -1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of *r* means that when *x* increases, *y* increases and when *x* decreases, *y* decreases **(positive correlation)**.
- A negative value of *r* means that when *x* increases, *y* decreases and when *x* decreases, *y* increases **(negative correlation)**.
- The sign of *r* is the same as the sign of the slope, *b*, of the best fit line.

Strong correlation does not suggest that *x* causes *y* or *y* causes *x*. We say **"correlation does not imply causation."** For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!


<db:title>Positive Correlation</db:title>

<db:title>Negative Correlation</db:title>

<db:title>Zero Correlation</db:title>

(a) A scatter plot showing data with a positive correlation. 0<r<1

(b) A scatter plot showing data with a negative correlation. -1<r<0

(c) A scatter plot showing data with zero correlation. r =0

Figure 12.11

The formula for *r* looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate *r*. The correlation coefficient *r* is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

r^2 is called the coefficient of determination. r^2 is the square of the correlation coefficient , but is usually stated as a percent, rather than in decimal form. r^2 has an interpretation in the context of the data

- r^2, when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- 1-r^2, when expressed as a percent, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

The line of best fit is: $\hat{y} = -173.51 + 4.83x$

The correlation coefficient is $r = 0.6631$

The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$

Interpretation of r^2 in the context of this example:

Approximately 44% of the variation in the final exam grades can be explained by the variation in the grades on the third exam, using the best fit regression line.

Therefore approximately 56% of the variation in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best fit regression line. (This is seen as the scattering of the points about the line.)

**With contributions from Roberta Bloom.

12.7 Testing the Significance of the Correlation Coefficient

Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough and reliable enough to use to model the relationship in the population.

The sample data is used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we only have sample data, we can not calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter "rho".

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to 0" or "significantly different from 0". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is "significant".

- Conclusion: "The correlation coefficient IS SIGNIFICANT"
- What the conclusion means: We believe that there is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".

- Conclusion: "The correlation coefficient IS NOT SIGNIFICANT."
- What the conclusion means: We do NOT believe that there is a significant linear relationship between x and y. Therefore we can NOT use the regression line to model a linear relationship between x and y in the population.

- If *r* is significant and the scatter plot shows a reasonable linear trend, the line can be used to predict the value of *y* for values of *x* that are within the domain of observed *x* values.
- If *r* is not significant OR if the scatter plot does not show a reasonable linear trend, the line should not be used for prediction.
- If *r* is significant and if the scatter plot shows a reasonable linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed *x* values in the data.

PERFORMING THE HYPOTHESIS TEST

SETTING UP THE HYPOTHESES:

- **Null Hypothesis: Ho: ρ=0**
- **Alternate Hypothesis: Ha: ρ≠0**

What the hypotheses mean in words:

- **Null Hypothesis Ho:** The population correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis Ha:** The population correlation coefficient IS significantly DIFFERENT FROM 0. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

There are two methods to make the decision. Both methods are equivalent and give the same result.

Method 1: Using the p-value

Method 2: Using a table of critical values

In this chapter of this textbook, we will always use a significance level of 5%, α = 0.05

Note: Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using α = 0.05. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, α = 0.05. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

The linear regression t-test LinRegTTEST on the TI-83+ or TI-84+ calculators calculates the p-value.

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "**≠ 0**"

The output screen shows the p-value on the line that reads "p=".

(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level (α = 0.05):

- Decision: REJECT the null hypothesis.
- Conclusion: "The correlation coefficient IS SIGNIFICANT."
- We believe that there IS a significant linear relationship between x and y. because the correlation coefficient is significantly different from 0.

If the p-value is NOT less than the significance level (α = 0.05)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "The correlation coefficient is NOT significant."
- We believe that there is NOT a significant linear relationship between x and y. because the correlation coefficient is NOT significantly different from 0.

You will use technology to calculate the p-value. The following describe the calculations to compute the test statistics and the p-value:

The p-value is calculated using a *t*-distribution with n-2 degrees of freedom.

The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, *t*, is shown in the computer or calculator output along with the p-value.

The test statistic *t* has the same sign as the correlation coefficient *r*.

The p-value is the probability (area) in both tails further out beyond the values -*t* and *t*.

For the TI-83+ and TI-84+ calculators, the command 2*tcdf(abs(t),10^99, n-2) computes the p-value given by the LinRegTTest; abs(t) denotes absolute value: |*t*|

THIRD EXAM vs FINAL EXAM EXAMPLE: p value method

- Consider the **third exam/final exam example (http://cnx.org/content/m17092/1.11/#element-22)** .
- The line of best fit is: $\hat{y}$ = -173.51 + 4.83x with *r* = 0.6631 and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (*x* value), can we use the line to predict the final exam score (predicted *y* value)?**

Ho: ρ = 0

Ha: ρ ≠ 0

α = 0.05

The p-value is 0.026 (from LinRegTTest on your calculator or from computer software)

The p-value, 0.026, is less than the significance level of α = 0.05

Decision: Reject the Null Hypothesis Ho

Conclusion: The correlation coefficient IS SIGNIFICANT.

Because *r* is significant and the scatter plot shows a reasonable linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table at the end of this chapter (before the Summary) may be used to give you a good idea of whether the computed value of ***r* is significant or not**. Compare *r* to the appropriate critical value in the table. If *r* is not between the positive and negative critical values, then the correlation coefficient is significant. If *r* is significant, then you may want to use the line for prediction.

Example 12.7

Suppose you computed *r* = 0.801 using *n* = 10 data points. df = *n* − 2 = 10 − 2 = 8. The critical values associated with df = 8 are -0.632 and + 0.632. If *r*<negative critical value or *r* > positive critical value, then *r* is significant. Since *r* = 0.801 and 0.801 > 0.632, *r* is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

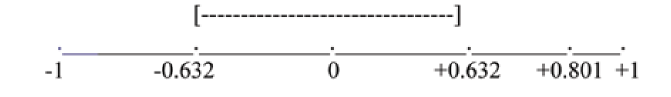


Figure 12.12 *r* is not significant between -0.632 and +0.632. *r* = 0.801 > +0.632. Therefore, *r* is significant.

Example 12.8

Suppose you computed *r* = -0.624 with 14 data points. df = 14 − 2 = 12. The critical values are -0.532 and 0.532. Since -0.624<-0.532, *r* is significant and the line may be used for prediction

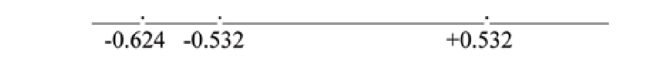


Figure 12.13 *r* = -0.624<-0.532. Therefore, *r* is significant.

Example 12.9

Suppose you computed *r* = 0.776 and *n* = 6. df = 6 − 2 = 4. The critical values are -0.811 and 0.811. Since -0.811< 0.776 < 0.811, *r* is not significant and the line should not be used for prediction.

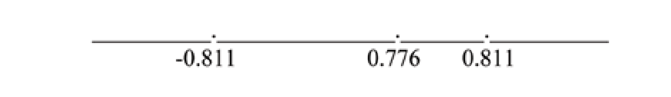


Figure 12.14 -0.811<*r* = 0.776<0.811. Therefore, *r* is not significant.

THIRD EXAM vs FINAL EXAM EXAMPLE: critical value method

- Consider the **third exam/final exam example (http://cnx.org/content/m17092/1.11/#element-22)** .
- The line of best fit is: $\hat{y}$ = -173.51 + 4.83x with *r* = 0.6631 and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (*x* value), can we use the line to predict the final exam score (predicted *y* value)?**

Ho: ρ = 0

Ha: ρ ≠ 0

α = 0.05
Use the "95% Critical Value" table for *r* with df = n − 2 = 11 − 2 = 9
The critical values are -0.602 and +0.602
Since 0.6631 > 0.602, *r* is significant.
Decision: Reject Ho
Conclusion: The correlation coefficient is significant
Because *r* is significant and the scatter plot shows a reasonable linear trend, the regression line can be used to predict final exam scores.

Example 12.10 Additional Practice Examples using Critical Values

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if *r* is significant and the line of best fit associated with each *r* can be used to predict a *y* value. If it helps, draw a number line.

1. *r* = -0.567 and the sample size, *n*, is 19. The df = *n* − 2 = 17. The critical value is -0.456. -0.567<-0.456 so *r* is significant.
2. *r* = 0.708 and the sample size, *n*, is 9. The df = *n* − 2 = 7. The critical value is 0.666. 0.708 > 0.666 so *r* is significant.
3. *r* = 0.134 and the sample size, *n*, is 14. The df = 14 − 2 = 12. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so *r* is not significant.
4. *r* = 0 and the sample size, *n*, is 5. No matter what the dfs are, *r* = 0 is between the two critical values so *r* is not significant.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best fit line for our particular sample. We want to use this best fit line for the sample as an estimate of the best fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the **average of the y values for each particular x value** lie on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) above implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line the equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.

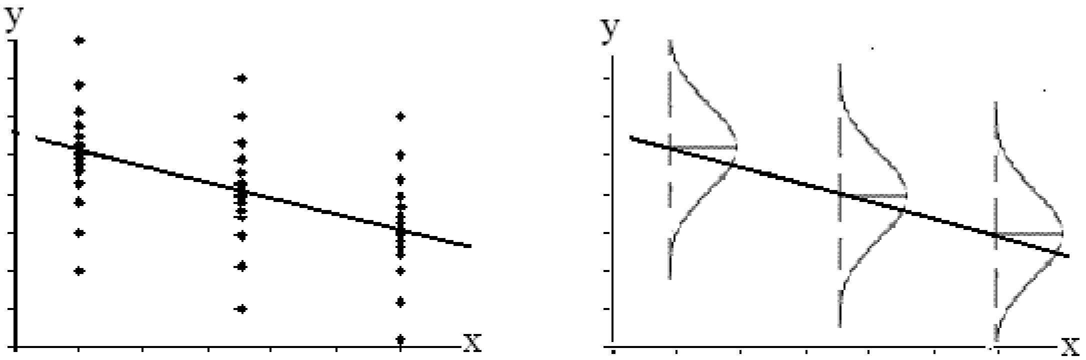


Figure 12.15 The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

**With contributions from Roberta Bloom

12.8 Prediction

Recall the **third exam/final exam example (http://cnx.org/content/m17092/1.11/#element-22)** .

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best fit line for the final exam grade as a function of the grade on the third exam. We can now use the least squares regression line for prediction.

Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The exam scores **(*x*-values)** range from 65 to 75. **Since 73 is between the *x*-values 65 and 75**, substitute *x* = 73 into the equation. Then:

$$\hat{y} = -173.51 + 4.83\left(73\right) = 179.08 \tag{12.7}$$

We predict that statistic students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 12.11

Recall the **third exam/final exam example (http://cnx.org/content/m17092/1.11/#element-22)** .

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

145.27

What would you predict the final exam score to be for a student who scored a 78 on the third exam?

The x values in the data are between 65 and 75. 78 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter x into the equation and calculate a y value, you should not do so!)

**With contributions from Roberta Bloom

12.9 Outliers

In some data sets, there are values **(observed data points)** called outliers. **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to carefully examine what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points but that greatly influence the line. As a result an influential point may be close to the line, even though it is far from the rest of the data. Because an influential point so strongly influences the best fit line, it generally will not have a large "error" or residual.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatterplot and best fit line. However we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best fit line as an outlier**. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatterplot by drawing an extra pair of lines that are two standard deviations above and below the best fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally only need to use one of these methods.

Example 12.12

In the **third exam/final exam example (http://cnx.org/content/m17092/1.11/#element-22)** , you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Graphical Identification of Outliers

With the TI-83,83+,84+ graphing calculators, it is easy to identify the outlier graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance was equal to 2*s* or farther, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find **s=16.412**

Line Y2=-173.5+4.83x-2(16.4) and line Y3=-173.5+4.83X+2(16.4)

where $\hat{y}$=-173.5+4.83x is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatterplot with the best fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the "Y="equation editor and press ZOOM 9. You will find that the only data point that is not between lines Y2 and Y3 is the point x=65, y=175. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than 2 standard deviations away from the best fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph more clear. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

Scatterplot of data and best fit line of the exam score data, showing the lines two standard deviations above and below the best fit line. The data value (65,175) lies slightly above the upper line identifying it as an outlier.

Numerical Identification of Outliers

In the table below, the first two columns are the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y}$=-173.5+4.83x. The residuals, or errors, have been calculated in the fourth column of the table: observed y value − predicted y value = $y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values where *n* = the total number of data points. If each residual is calculated and squared, and the results are added up, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

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

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, our calculator LinRegTTest found **s=16.4** as the standard deviation of the residuals: 35; -17; 16; -6; -19; 9; 3; -1; -10; -9; -1.

x	*y*	$\hat{y}$	$y - \hat{y}$
65	175	140	175 − 140 = 35
67	133	150	133 − 150 = -17
71	185	169	185 − 169 = 16
71	163	169	163 − 169 = -6
66	126	145	126 − 145 = -19
75	198	189	198 − 189 = 9
67	153	150	153 − 150 = 3
70	163	164	163 − 164 = -1
71	159	169	159 − 169 = -10
69	151	160	151 − 160 = -9
69	159	160	159 − 160 = -1

We are looking for all data points for which the residual is greater than 2s=2(16.4)=32.8 or less than -32.8. Compare these values to the residuals in column 4 of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line?

Numerically and graphically, we have identified the point (65,175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. **For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on to delete the outlier, so that we can explore how it affects the results, as a learning experience.**

Compute a new best-fit line and correlation coefficient using the 10 remaining points:

On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$\hat{y}$ = -355.19 + 7.39x and *r* = 0.9121

The new line with *r* = 0.9121 is a stronger correlation than the original (*r*=0.6631) because *r* = 0.9121 is closer to 1. This means that the new line is a better fit to the 10 remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating s and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, **square each** $\left| \boldsymbol{y - \hat{y}} \right|$ (See the TABLE above):

The squares are 35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2

Then, add (sum) all the $\left| \boldsymbol{y - \hat{y}} \right|$ **squared terms** using the formula

$$\sum_{i=1}^{11} \left(\left| y_i - \hat{y}_i \right| \right)^2 = \sum_{i=1}^{11} \varepsilon_i^2 \qquad \text{(Recall that } \left| y_i - \hat{y}_i \right| = \varepsilon_i\text{.)}$$

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

= 2440 = **SSE**. The result, **SSE** is the Sum of Squared Errors.

Next, calculate *s*, the standard deviation of all the $\left| \boldsymbol{y - \hat{y}} \right|$ **= ε values where *n* = the total number of data points.** (Calculate the standard deviation of 35; 17; 16; 6; 19; 9; 3; 1; 10; 9; 1.)

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

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$

Next, multiply *s* by 1.9:
$(1.9) \cdot (16.47) = 31.29$

31.29 is almost 2 standard deviations away from the mean of the $\left| y - \hat{y} \right|$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 1.9*s*, then we would consider the data point to be "too far" from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $\left| y - \hat{y} \right|$ values are **at least** 31.29, the corresponding (*x*, *y*) data point is a potential outlier.

For the third exam/final exam problem, all the $\left| y - \hat{y} \right|$'s are less than 31.29 except for the first one which is 35.

35 > 31.29 That is, $\left| y - \hat{y} \right| \geq (1.9) \cdot (s)$

The point which corresponds to $\left| y - \hat{y} \right| = 35$ is (65, 175). **Therefore, the data point (65, 175) is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

Example 12.13

Using this new line of best fit (based on the remaining 10 data points), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

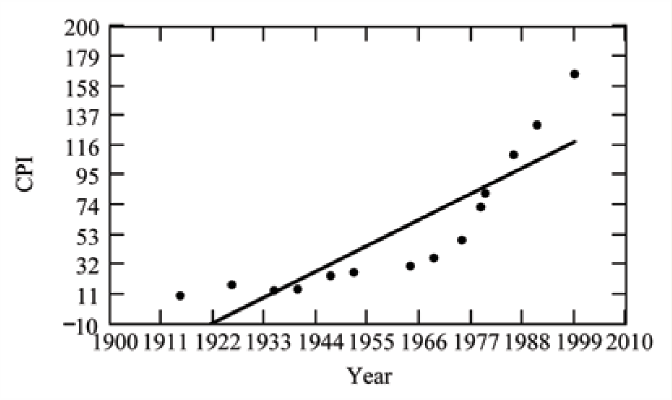
Example 12.14

(*From The Consumer Price Indexes Web site*) The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Table 12.3
Data:

x	*y*
1915	10.1
1926	17.7
1935	13.7
1940	14.7
1947	24.1
1952	26.5
1964	31.0
1969	36.7
1975	49.3
1979	72.6
1980	82.4
1986	109.6
1991	130.7
1999	166.6

- Make a scatterplot of the data.
- Calculate the least squares line. Write the equation in the form $\hat{y} = a + \text{b}x$.
- Draw the line on the scatterplot.
- Find the correlation coefficient. Is it significant?
- What is the average CPI for the year 1990?
- Scatter plot and line of best fit.
- $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
- $r = 0.8694$
- The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since 0.8694 > 0.532, *r* is significant.
- $\hat{y} = -3204 + 1.662\left(1990\right) = 103.4$ CPI
- Using the calculator LinRegTTest, we find that s = 25.4 ; graphing the lines Y2=-3204+1.662X-2(25.4) and Y3=-3204+1.662X+2(25.4) shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)



In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt ; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years 2004 : CPI=188.9 and 2008 : CPI=215.3 and see how it affects the model.

**With contributions from Roberta Bloom

12.10 95% Critical Values of the Sample Correlation Coefficient Table

Table 12.4

Degrees of Freedom: *n* − 2	Critical Values: (+ and −)
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100 and over	0.195

12.11 Summary

Bivariate Data: Each data point has two values. The form is (*x*, *y*).

Line of Best Fit or Least Squares Line (LSL): $\hat{y} = a + bx$

x = independent variable; *y* = dependent variable

Residual: Actual y value−predicted y value = $y - \hat{y}$

Correlation Coefficient r:

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer *r* is to 1 or -1, the closer the original points are to a straight line.
3. If *r* is negative, the slope is negative. If *r* is positive, the slope is positive.
4. If *r* = 0, then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the **SSE**, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

12.12 Practice: Linear Regression

Student Learning Outcomes

- The student will explore the properties of linear regression.

Given

The data below are real. Keep in mind that these are only reported figures. (*Source: Centers for Disease Control and Prevention, National Center for HIV, STD, and TB Prevention, October 24, 2003*)

Table 12.5 Adults and Adolescents only, United States

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

We will use the columns “year” and “# AIDS cases diagnosed” for all questions unless otherwise stated.

Graphing

Graph “year” vs. “# AIDS cases diagnosed.” **Plot the points on the graph located below in the section titled "Plot"** . Do not include pre-1981. Label both axes with words. Scale both axes.

Data

Linear Equation

Write the linear equation below, rounding to 4 decimal places:

Solve

Plot

Plot the 2 above points on the graph below. Then, connect the 2 points to form the regression line.

Obtain the graph on your calculator or computer.

Discussion Questions

Look at the graph above.

12.13 Homework

The next two questions refer to the following data: The cost of a leading liquid laundry detergent in different sizes is given below.

Table 12.6

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

The next three questions use the following state information.

Table 12.7

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado		1876	38	104,100
Hawaii		1959	50	10,932
Iowa		1846	29	56,276
Maryland		1788	7	12,407
Missouri		1821	24	69,709
New Jersey		1787	3	8,722
Ohio		1803	17	44,828
South Carolina	13	1788	8	32,008
Utah		1896	45	84,904
Wisconsin		1848	30	65,499

The next two questions refer to the following information: The data below reflects the 1991-92 Reunion Class Giving. (Source: *SUNY Albany alumni magazine*)

Table 12.8

Class Year	Average Gift	Total Giving
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

Try these multiple choice questions

The next three questions refer to the following data: (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from **www.nhc.noaa.gov/gifs/table6.gif (http://www.nhc.noaa.gov/gifs/table6.gif)** A major hurricane is one with a strength rating of 3, 4 or 5.

Table 12.9

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	9	3

**Exercises 21 and 22 contributed by Roberta Bloom

12.14 Lab 1: Regression (Distance from School)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Use 8 members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Table 12.10

Distance from school	Cost of supplies this term

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph "distance" vs. "cost." Plot the points on the graph. Label both axes with words. Scale both axes.

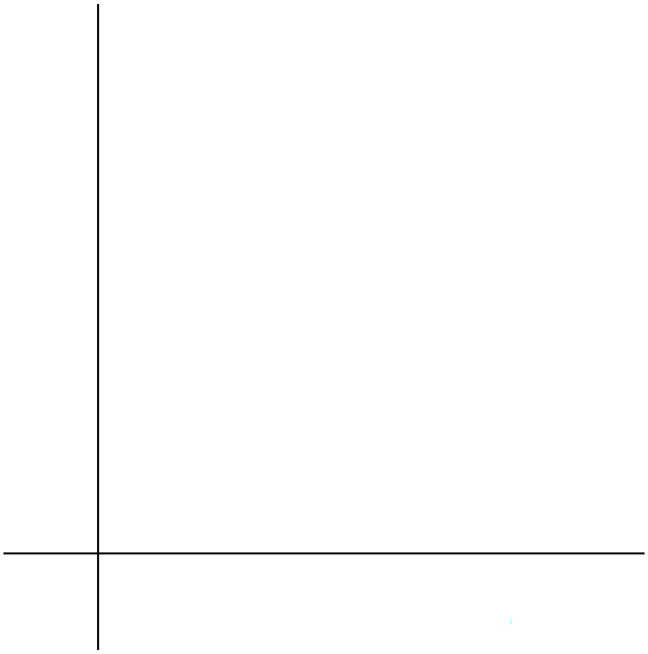


Figure 12.16

Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   a. a =

b. *b*=
c. correlation =
d. *n* =
e. equation: $\hat{y}$ =
f. Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)

2. Supply an answer for the following senarios:
   a. For a person who lives 8 miles from campus, predict the total cost of supplies this term:
   b. For a person who lives 80 miles from campus, predict the total cost of supplies this term:
3. Obtain the graph on your calculator or computer. Sketch the regression line below.

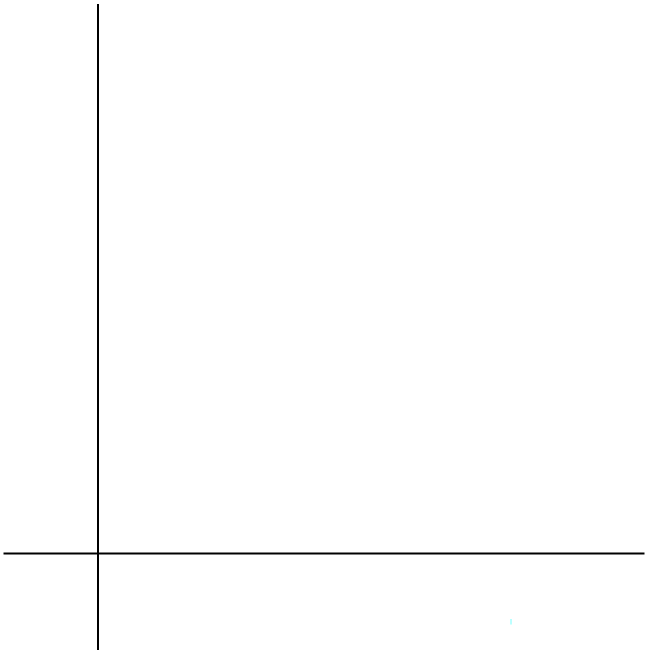


Figure 12.17

Discussion Questions

1. Answer each with 1-3 complete sentences.
   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between the distance and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.15 Lab 2: Regression (Textbook Cost)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Survey 10 textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Table 12.11

Number of pages	Cost of textbook

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “distance” vs. “cost.” Plot the points on the graph in "Analyze the Data". Label both axes with words. Scale both axes.

Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   a. *a* =
   b. *b* =
   c. correlation =
   d. *n* =
   e. equation: *y* =
   f. Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)
2. Supply an answer for the following senarios:
   a. For a textbook with 400 pages, predict the cost:
   b. For a textbook with 600 pages, predict the cost:
3. Obtain the graph on your calculator or computer. Sketch the regression line below.

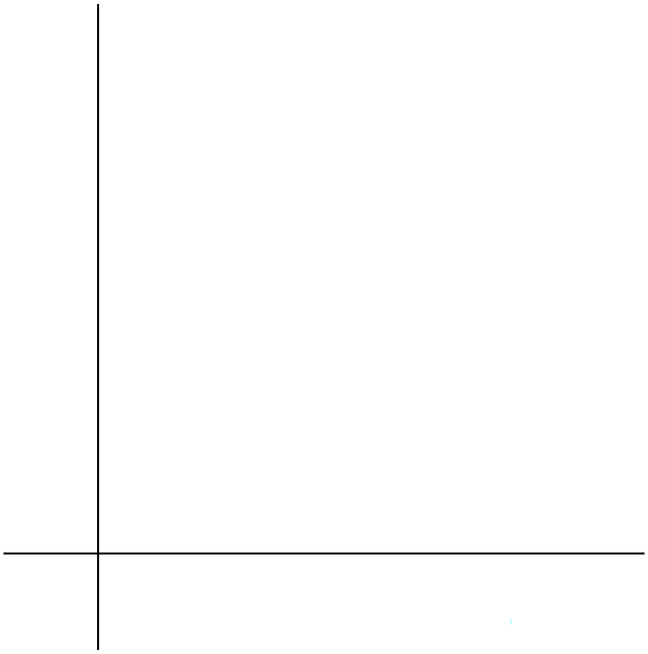


Figure 12.18

Discussion Questions

1. Answer each with 1-3 complete sentences.
   a. Does the line seem to fit the data? Why?

b. What does the correlation imply about the relationship between the number of pages and the cost?

2. Are there any outliers? If so, which point(s) is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.16 Lab 3: Regression (Fuel Efficiency)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Use the most recent April issue of Consumer Reports. It will give the total fuel efficiency (in miles per gallon) and weight (in pounds) of new model cars with automatic transmissions. We will use this data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Which variable should be the independent variable and which should be the dependent variable? Explain your answer in one or two complete sentences.
2. Using your random number generator, randomly select 20 cars from the list and record their weights and fuel efficiency into the table below.

Table 12.12

Weight	Fuel Efficiency

3. Which variable should be the dependent variable and which should be the independent variable? Why?
4. By hand, do a scatterplot of “weight” vs. “fuel efficiency”. Plot the points on graph paper. Label both axes with words. Scale both axes accurately.

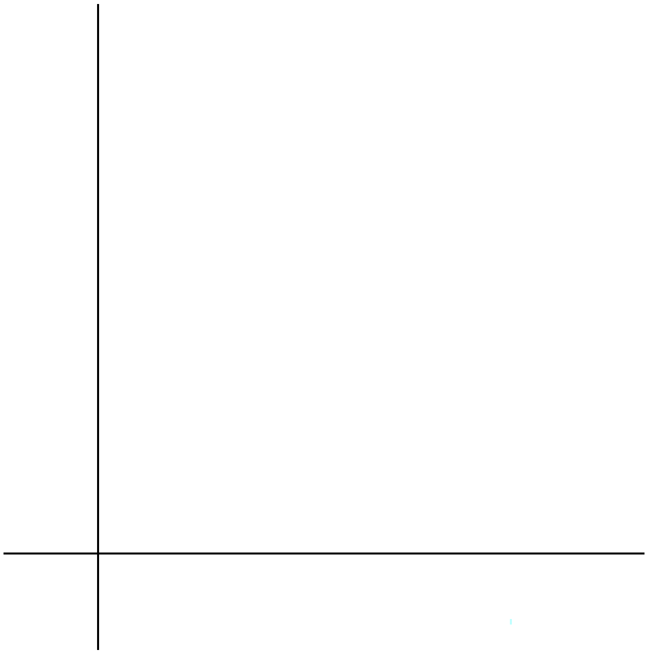


Figure 12.19

Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   a. *a* =
   b. *b* =
   c. correlation =
   d. *n* =
   e. equation: $\hat{y}$ =
2. Obtain the graph of the regression line on your calculator. Sketch the regression line on the same axes as your scatterplot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10000 pounds using the least squares line? Explain why or why not.
6. Questions. Answer each in 1 to 3 complete sentences.
   a. Does the line seem to fit the data? Why or why not?
   b. What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?
7. Are there any outliers? If so, which point is an outlier?

** This lab was designed and contributed by Diane Mathios.

Glossary

Coefficient of Correlation: A measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable. The formula is:

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

where n is the number of data points. The coefficient cannot be more then 1 and less then -1. The closer the coefficient is to ± 1, the stronger the evidence of a significant linear relationship between X and y.

Outlier: An observation that does not fit the rest of the data.

13 F DISTRIBUTION AND ANOVA

13.1 F Distribution and ANOVA

Student Learning Objectives

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

- Interpret the F probability distribution as the number of groups and the sample size change.
- Discuss two uses for the F distribution, ANOVA and the test of two variances.
- Conduct and interpret ANOVA.
- Conduct and interpret hypothesis tests of two variances (optional).

Introduction

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests involving more than two averages, statisticians have developed a method called Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the F distribution, used for ANOVA, and the test of two variances. This is just a very brief overview of ANOVA. You will study this topic in much greater detail in future statistics courses.

- ANOVA, as it is presented here, relies heavily on a calculator or computer.
- For further information about ANOVA, use the online link **ANOVA (http://en.wikipedia.org/wiki/Analysis_of_variance)** . Use the back button to return here. (The url is http://en.wikipedia.org/wiki/Analysis_of_variance.)

13.2 ANOVA

F Distribution and ANOVA: Purpose and Basic Assumption of ANOVA

The purpose of an **ANOVA** test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to help determine if the means are equal or not.

In order to perform an ANOVA test, there are three basic **assumptions** to be fulfilled:

- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.
- The populations are assumed to have **equal standard deviations (or variances).**

The Null and Alternate Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternate hypothesis is that at least one pair of means is different. For example, if there are *k* groups:

$H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_a : At least two of the group means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal.

13.3 The F Distribution and the F Ratio

The distribution used for the hypothesis test is a new one. It is called the F distribution, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if *F* follows an *F* distribution and the degrees of freedom for the numerator are 4 and the degrees of freedom for the denominator are 10, then $F \sim F_{4, 10}$.

To calculate the *F* ratio, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means. If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- $SS_{between}$ = the sum of squares that represents the variation among the different samples.
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities which, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in **Descriptive Statistics**.

MS means "mean square." $\text{MS}_{\text{between}}$ is the variance between groups and $\text{MS}_{\text{within}}$ is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the jth group
- s_j= the sum of the values in the jth group
- N = total number of all the values combined. (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{total} = \sum x^2 - \frac{(\sum x)^2}{N}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{N}$
- Explained variation- sum of squares representing variation among the different samples $SS_{between} = \sum\left[\frac{(sj)^2}{n_j}\right] - \frac{(\sum s_j)^2}{N}$
- Unexplained variation- sum of squares representing variation within samples due to chance: $SS_{within} = SS_{total} - SS_{between}$
- df's for different groups (df's for the numerator): $df_{between} = k - 1$
- Equation for errors within samples (df's for the denominator): $df_{within} = N - k$
- Mean square (variance estimate) explained by the different groups: $MS_{between} = \frac{SS_{between}}{df_{between}}$
- Mean square (variance estimate) that is due to chance (unexplained): $MS_{within} = \frac{SS_{within}}{df_{within}}$

$\text{MS}_{\text{between}}$ and $\text{MS}_{\text{within}}$ can be written as follows:

- $\text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{\text{df}_{\text{between}}} = \frac{\text{SS}_{\text{between}}}{k-1}$
- $\text{MS}_{\text{within}} = \frac{\text{SS}_{\text{within}}}{\text{df}_{\text{within}}} = \frac{\text{SS}_{\text{within}}}{N-\text{k}}$

The ANOVA test depends on the fact that $\text{MS}_{\text{between}}$ can be influenced by population differences among means of the several groups. Since $\text{MS}_{\text{within}}$ compares values of each group to its own group mean, the fact that group means might be different does not affect $\text{MS}_{\text{within}}$.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, $\text{MS}_{\text{between}}$ and $\text{MS}_{\text{within}}$ should both estimate the same value.

> The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} \tag{13.1}$$

If $\text{MS}_{\text{between}}$ and $\text{MS}_{\text{within}}$ estimate the same value (following the belief that H_o is true), then the F-ratio should be approximately equal to 1. Only sampling errors would contribute to variations away from 1. As it turns out, $\text{MS}_{\text{between}}$ consists of the population variance plus a variance produced from the differences between the samples. $\text{MS}_{\text{within}}$ is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, $\text{MS}_{\text{between}}$ will be larger than $\text{MS}_{\text{within}}$. The F-ratio will be larger than 1.

The above calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot \left(s_{\overline{x}}\right)^2}{\left(s_{\text{pooled}}\right)^2} \tag{13.2}$$

where ...

- $(s_{\overline{x}})^2 =$ the variance of the sample means
- $n =$ the sample size of each group

- $(s_{pooled})^2$ = the mean of the sample variances (pooled variance)
- $df_{numerator} = k - 1$
- $df_{denominator} = k(n - 1) = N - k$

The ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_o.

Notation

The notation for the F distribution is $F \sim F_{\text{df(num),df(denom)}}$

where df(num) = $\text{df}_{\text{between}}$ and df(denom) = $\text{df}_{\text{within}}$

The mean for the F distribution is $\mu = \dfrac{\text{df(num)}}{\text{df(denom)} - 1}$

13.4 Facts About the F Distribution

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of dfs.
3. The F statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
5. Other uses for the F distribution include comparing two variances and Two-Way Analysis of Variance. Comparing two variances is discussed at the end of the chapter. Two-Way Analysis is mentioned for your information only.

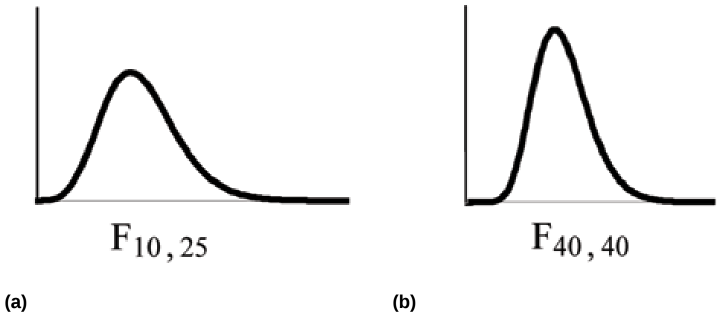


Figure 13.1

Example 13.1

One-Way ANOVA: Four sororities took a random sample of sisters regarding their grade averages for the past term. The results are shown below:

Table 13.1

GRADE AVERAGES FOR FOUR SORORITIES			
Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Using a significance level of 1%, is there a difference in grade averages among the sororities?

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each size 5.

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means μ_1, μ_2, μ_3, μ_4 are equal.

Distribution for the test: $F_{3,\,16}$

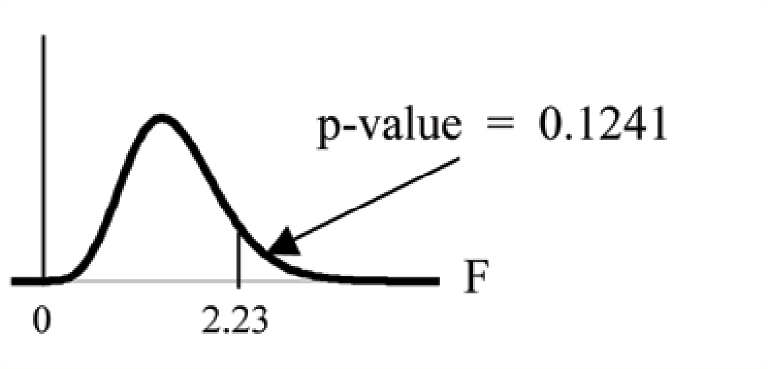
where k = 4 groups and N = 20 samples in total

df(num) = $k - 1 = 4 - 1 = 3$

df(denom) = $N - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph:



Probability statement: p-value = $P(F > 2.23) = 0.1241$

Compare α and the p-value: α = 0.01 p-value = 0.1242 α < p-value

Make a decision: Since α < p-value, you cannot reject H_o.

This means that the population averages appear to be the same.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the grade averages for the sororities.

TI-83+ or TI 84: Put the data into lists L1, L2, L3, and L4. Press `STAT` and arrow over to `TESTS`. Arrow down to `F:ANOVA`. Press `ENTER` and Enter (`L1,L2,L3,L4`). The F statistic is 2.2303 and the p-value is 0.1241. df(numerator) = 3 (under `"Factor"`) and df(denominator) = 16 (under `Error`).

Example 13.2

A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew 5 plants. At the end of the growing period, each plant was measured, producing the following data (in inches):

Table 13.2

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Does it appear that the three media in which the bean plants were grown produce the same average height? Test at a 3% level of significance.

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants so we will use the formula $F' = \dfrac{n \cdot \left(s_{\overline{x}}\right)^2}{\left(s_{\text{pooled}}\right)^2}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Next, calculate the variance of the three group means (Calculate the variance of 24.2, 25.4, and 24.4). **Variance of the group means = 0.413** = $\left(s_{\bar{x}}\right)^2$

Then $\text{MS}_{\text{between}} = n\left(s_{\bar{x}}\right)^2 = (5)(0.413)$ where n = 5 is the sample size (number of plants each child grew).

Calculate the average of the three sample variances (Calculate the average of 11.7, 18.3, and 16.3). **Average of the sample variances = 15.433** = $\left(s_{pooled}\right)^2$

Then $\text{MS}_{\text{within}} = \left(s_{pooled}\right)^2 = 15.433$.

The F statistic (or F ratio) is $F = \dfrac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} = \dfrac{n \cdot \left(s_{\bar{x}}\right)^2}{\left(s_{\text{pooled}}\right)^2} = \dfrac{(5)\cdot(0.413)}{15.433} = 0.134$

The dfs for the numerator = the number of groups−1 = 3 − 1 = 2

The dfs for the denominator = the total number of samples−the number of groups = 15 − 3 = 12

The distribution for the test is $F_{2,\,12}$ and the F statistic is $F = 0.134$

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since α = 0.03 and the p-value = 0.8759, do not reject H_o. (Why?)

Conclusion: With a 3% the level of significance, from the sample data, the evidence is not sufficient to conclude that the average heights of the bean plants are not different. Of the three media tested, it appears that it does not matter which one the bean plants are grown in.

(This experiment was actually done by three classmates of the son of one of the authors.)

Another fourth grader also grew bean plants but this time in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32.

Do an ANOVA test on the 4 groups. You may use your calculator or computer to perform the test. Are the heights of the bean plants different? Use a solution sheet.

- F = 0.9496
- p-value = 0.4401

The heights of the bean plants are the same.

Optional Classroom Activity

Randomly divide the class into four groups of the same size. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1% level of significance. Use one of the solution sheets at the end of the chapter (after the homework).

13.5 Test of Two Variances

Another of the uses of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

In order to perform a F test of two variances, it is important that the following are true:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}$$

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F-Ratio becomes $F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}$.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to 1. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than 1. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

Therefore, if F is close to 1, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than 1, then the evidence is against the null hypothesis.

A test of two variances may be left, right, or two-tailed.

Example 13.3

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9.

Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30$.

H_o: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

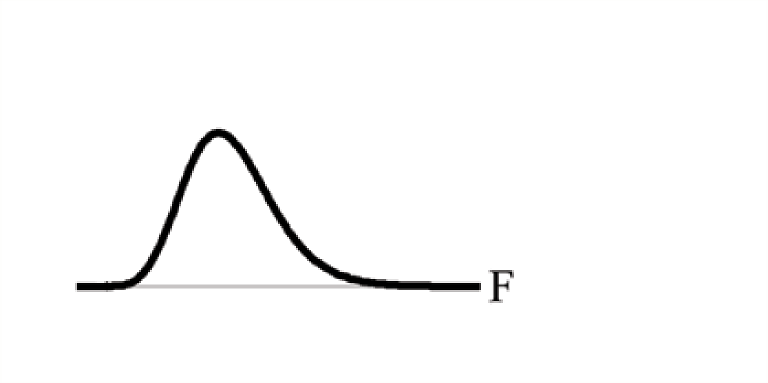
Calculate the test statistic: By the null hypothesis $\left(\sigma_1^2 = \sigma_2^2\right)$, the F statistic is

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.6$$

Distribution for the test: $F_{29, 29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph:This test is left tailed.

Draw the graph labeling and shading appropriately.



Probability statement: p-value = P ($F < 0.582$) = 0.0755

Compare α and the p-value: α = 0.10 α > p-value.

Make a decision: Since α > p-value, reject H_o.

Conclusion: With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

TI-83+ and TI-84: Press `STAT` and arrow over to `TESTS`. Arrow down to `D:2-SampFTest`. Press `ENTER`. Arrow to `Stats` and press `ENTER`. For `Sx1`, `n1`, `Sx2`, and `n2`, enter $\sqrt{(52.3)}$, `30`, $\sqrt{(89.9)}$, and `30`. Press `ENTER` after each. Arrow to `σ1:` and `<σ2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. F = 0.5818 and p-value = 0.0753. Do the procedure again and try `Draw` instead of `Calculate`.

13.6 Summary

- An **ANOVA** hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with 2 different degrees of freedom.
  Assumptions:
  1. Each population from which a sample is taken is assumed to be normal.
  2. Each sample is randomly selected and independent.
  3. The populations are assumed to have equal standard deviations (or variances)
- A **Test of Two Variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the F distribution with 2 different degrees of freedom.
  Assumptions:
  1. The populations from which the two samples are drawn are normally distributed.
  2. The two populations are independent of each other.

13.7 Practice: ANOVA

Student Learning Outcome

- The student will explore the properties of ANOVA.

Given

Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

Table 13.3

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\overline{x}$ =	________	________	________	________	________
s^2 =	________	________	________	________	________

Hypothesis

Data Entry

Enter the data into your calculator or computer.

Decisions and Conclusions

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α .

13.8 Homework

Directions

Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in the Table of Contents 14. Appendix.

For the next two problems, refer to the data from Terri Vogel's Log Book [link pending].

For the next four problems, refer to the following data.

The following table lists the number of pages in four different types of magazines.

Table 13.4

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

13.9 Review

The next two questions refer to the following situation:

Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

The next three questions refer to the following survey:

Table 13.5 Favorite Type of Pie by Gender

	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

The next two questions refer to the following situation:

Let's say that the probability that an adult watches the news at least once per week is 0.60.

The next three questions refer to the following situation:

The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

The next three questions refer to the following situation:

120 people were surveyed as to their favorite beverage (non-alcoholic). The results are below.

Table 13.6 Preferred Beverage by Age

	0 – 9	10 – 19	20 – 29	30 +		Totals
Milk	14	10	6	0	30	
Soda	3	8	26	15	52	
Juice	7	12	12	7	38	
Totals	24	30	44	22	120	

13.10 Lab: ANOVA

Class Time:

Names:

Student Learning Outcome:

- The student will conduct a simple ANOVA test involving three variables.

Collect the Data

1. Record the price per pound of 8 fruits, 8 vegetables, and 8 breads in your local supermarket.

Table 13.7

Fruits	Vegetables	Breads

2. Explain how you could try to collect the data randomly.

Analyze the Data and Conduct a Hypothesis Test

1. Compute the following:

   a. Fruit:

   i. $\overline{x}$ =

   ii. s_x=

   iii. n =

   a. Vegetables:

   i. $\overline{x}$ =

   ii. s_x =

   iii. n=

   a. Bread:

   i. $\overline{x}$ =

   ii. s_x=

   iii. n =

2. Find the following:

   a. df(num) =

   b. df(denom) =

3. State the approximate distribution for the test.
4. Test statistic: F =
5. Sketch a graph of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.
6. p-value =
7. Test at α = 0.05. State your decision and conclusion.
8. **a.** Decision: Why did you make this decision?

   b. Conclusion (write a complete sentence).

   c. Based on the results of your study, is there a need to further investigate any of the food groups' prices? Why or why not?

Glossary

Analysis of Variance: Also referred to as ANOVA. A method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- All populations of interest are normally distributed.
- The populations have equal standard deviations.
- Samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

Variance: Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \overline{x}$ where x is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

14 APPENDIX

14.1 Practice Final Exam 1

Questions 1-2 refer to the following:

An experiment consists of tossing two 12-sided dice (the numbers 1-12 are printed on the sides of each dice).

- Let Event *A* = both dice show an even number
- Let Event *B* = both dice show a number more than 8

Questions 4 - 5 refer to the following:

118 students were asked what type of color their bedrooms were painted: light colors, dark colors or vibrant colors. The results were tabulated according to gender.

Table 14.1

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

Questions 6 – 7 refer to the following:

We are interested in the number of times a teenager must be reminded to do his/her chores each week. A survey of 40 mothers was conducted. The table below shows the results of the survey.

Table 14.2

X	$P(x)$
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

Questions 8 – 9 refer to the following:

On any given day, approximately 37.5% of the cars parked in the De Anza parking structure are parked crookedly. (Survey done by Kathy Plum.) We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

Questions 11 – 13 refer to the following:

De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1795 students enrolled in Math 1A (1st quarter calculus), 1428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

Table 14.3

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	.15
Alejandra	4.2	4.6	.25
Iris	4.5	4.9	.12

Questions 15 – 16 refer to the following:

The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog:

{$212, $292, $278, $199$280, $236}

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the average price of adult ski sweaters from Gorsuch Ltd. is at least $275.

Questions 20 - 22 refer to the following:

A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. The results of her survey are summarized in the table below.

Table 14.4

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		
6	1	.02	1.00

The next two questions refer to the following:

The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Table 14.5

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

The number of people living on American farms has declined steadily during this century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Table 14.6

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

The linear regression equation is y-hat = 1166.93 – 0.5868x

Question 34-36 refer to the following:

A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

Table 14.7 Does the data suggest that there is a relationship between the gender of students and their choice of major?

	Female	Male
Accounting	68	56
Administration	91	40
Ecomonics	5	6
Finance	61	59

14.2 Practice Final Exam 2

The next two questions refer to the following data:

Table 14.8

value	frequency
0	1
1	4
2	7
3	9
6	4

The next two questions refer to the following situation:

Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

The next two questions refer to the following situation:

Suppose that a survey yielded the following data:

Table 14.9 Favorite Pie Type

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

The next two questions refer to the following situation:

Let’s say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number that watch the news at least once per week.

The ages of campus day and evening students is known to be normally distributed. A sample of 6 campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}

The next three questions refer to the following situation:

The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

The next three questions refer to the following situation:

The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from 1 to 10 minutes.

The next three questions refer to the following situation:

At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of 5 minutes and a standard deviation of 1.5 minutes.

The next three questions refer to the following situation:

A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Table 14.10

Number of employees x	650	730	810	900	102	107	1150
Number of bathrooms y	40	50	54	61	82	110	121

14.3 Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's Log Book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

Table 14.11 Race Lap Times (in Seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Table 14.12 Practice Lap Times (in Seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading. This is historical data.

Table 14.13 IPO Offer Prices

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44
$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

Data compiled by Jay R. Ritter of Univ. of Florida using data from Securities Data Co. and Bloomberg.

14.4 Group Projects

Group Project: Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

_____ Decide what data you are going to study.

Examples

Here are two examples, but you may **NOT** use them: number of M&M's per small bag, number of pencils students have in their backpacks.

_____ Are your data discrete or continuous? How do you know?
_____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
_____ Conduct your survey. **Your data size must be at least 30.**
_____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
_____ Answer the following (rounded to 2 decimal places):

1. $\bar{x}$ =
2. s =
3. First quartile =
4. Median =
5. 70th percentile =

_____ What value is 2 standard deviations above the mean?
_____ What value is 1.5 standard deviations below the mean?
_____ Construct a histogram displaying your data.

____ In complete sentences, describe the shape of your graph.

____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula in Chapter 2 (since you have univariate data) to determine whether or not the values might be outliers.

____ Construct a box plot displaying your data.

____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.

____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

____ **Cover sheet**: name, class time, and name of your study

____ **Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.

____ **URL** for data, if your data are from the World Wide Web.

____ **Chart of data, frequency, relative frequency and cumulative relative frequency.**

____ **Page(s) of graphs:** histogram and box plot.

Group Project: Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the Central Limit Theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spends on college supplies this term or the length of a long distance telephone call.)

____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?

____ Conduct your survey. Gather **at least 150 pieces of continuous quantitative data**.

____ Define (in words) the random variable for your data. X = ________

____ Create 2 lists of your data: (1) unordered data, (2) in order of smallest to largest.

____ Find the sample mean and the sample standard deviation (rounded to 2 decimal places).

1. $\bar{x}$ =
2. s =

____ Construct a histogram of your data containing 5 - 10 intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

____ Suppose that X followed the theoretical distributions below. Set up each distribution using the appropriate information from your data.

____ Uniform: X ~ U _____________ Use the lowest and highest values as a and b.

____ Exponential: X ~ Exp _____________Use $\bar{x}$ to estimate μ .

____ Normal: X ~ N _____________ Use $\bar{x}$ to estimate for μ and s to estimate for σ.

____ **Must** your data fit one of the above distributions? Explain why or why not.

____ **Could** the data fit 2 or 3 of the above distributions (at the same time)? Explain.

____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = __________ (rounded to 2 decimal places) Note: $k = \bar{x} + \left(1.75\right) * s$

____ Determine the relative frequencies (RF) rounded to 4 decimal places.

1. RF=$\frac{\text{frequency}}{\text{total number surveyed}}$
2. RF($X < k$) =
3. RF($X > k$) =
4. RF($X = k$) =

Use a separate piece of paper for EACH distribution (uniform, exponential, normal) to respond to the following questions.

> You should have one page for the uniform, one page for the exponential, and one page for the normal

____ State the distribution: X ~ __________

____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
____ Find the following theoretical probabilities (rounded to 4 decimal places).

1. P(X < k) =
2. P(X > k)=
3. P(X = k) =

____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

______ From your original data (before ordering), use a random number generator to pick 40 samples of size 5. For each sample, calculate the average.
______ On a separate page, attached to the summary, include the 40 samples of size 5, along with the 40 sample averages.
______ List the 40 averages in order from smallest to largest.
______ Define the random variable, $\overline{X}$, in words. $\overline{X}$ =
______ State the approximate theoretical distribution of $\overline{X}$. $\overline{X}$~
______ Base this on the mean and standard deviation from your original data.
______ Construct a histogram displaying your data. Use 5 to 6 intervals of equal width. Label and scale it.

Calculate the value $\overline{k}$ (an $\overline{X}$ value) that is 1.75 standard deviations above the sample mean. $\overline{k}$= _____ (rounded to 2 decimal places)
Determine the relative frequencies (RF) rounded to 4 decimal places.

1. RF($\overline{X}$<$\overline{k}$) =
2. RF($\overline{X}$ > $\overline{k}$) =
3. RF($\overline{X}$ = $\overline{k}$) =

Find the following theoretical probabilities (rounded to 4 decimal places).

- **1.** P($\overline{X}$ < $\overline{k}$) =
- **2.** P($\overline{X}$ > $\overline{k}$) =
- **3.** P($\overline{X}$ = $\overline{k}$) =

______ Draw the graph of the theoretical distribution of *X*.
______ Answer the questions below.
______ Compare the relative frequencies to the probabilities. Are the values close?
______ Does it appear that the data of averages fit the distribution of $\overline{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
______ In 3 - 5 complete sentences for each, answer the following questions. Give thoughtful explanations.
______ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
______ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would “it” always happen? Why or why not?
______ Were the relative frequencies compared to the theoretical probabilities closer when comparing the *X* or $\overline{X}$ distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions above.
____ **URL** for data, if your data are from the World Wide Web.
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Partner Project: Hypothesis Testing - Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

____ **Find an article** in a newspaper, magazine or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.

____ **Copy or print out the article** and include a copy in your project, along with the source.

____ **State how you will collect your data.** (Convenience sampling is not acceptable.)

____ **Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.

____ **State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

____ **Make 2 copies of the appropriate solution sheet.**

____ **Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

____ **Create a graph that illustrates your data.** This may be a pie or bar chart or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.

____ **Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

1. Brief discussion of the article, including the source.

2. Statement of the claim made in the article (one of the hypotheses).

3. Detailed description of how, where, and when you collected the data, including the sampling technique. Did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As stated above, convenience sampling is not acceptable.

4. Conclusion about the article claim in light of your hypothesis test. This is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.

5. Sentence interpreting your confidence interval in the context of the situation in your project.

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

____ **Cover sheet** containing your name(s), class time, and the name of your study.

____ **Summary**, which includes all items listed on summary checklist.

____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.

____ **Graphic representation of your data**, created following the guidelines discussed above. Include only graphs which are appropriate and useful.

____ **Raw data collected AND a table summarizing the sample data** (n, xbar and s; or x, n, and p', as appropriate for your hypotheses). The raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Partner Project: Hypothesis Testing - Word Problem

Student Learning Objectives

- The student will write, edit, and solve a hypothesis testing word problem.

Instructions

Write an original hypothesis testing problem for either **ONE** population mean or **ONE** population proportion. As you complete each task, check it off. Answer all questions in your summary. Look at the homework for the Hypothesis Testing: Single Mean and Single Proportion chapter for examples (poems, two acts of a play, a work related problem). The problems with names attached to them are problems written by students in past quarters. Some other examples that are not in the homework include: a soccer hypothesis testing poster, a cartoon, a news reports, a children's story, a song.

____ Your problem must be original and creative. It also must be in proper English. If English is difficult for you, have someone edit your problem.

____ Your problem must be at least ½ page, typed and singled spaced. This **DOES NOT** include the data. Data will make the problem longer and that is fine. For this problem, the data and story may be real or fictional.

____ In the narrative of the problem, make it very clear what the null and alternative hypotheses are.

____ Your sample size must be **LARGER THAN 50** (even if it is fictional).

____ State in your problem how you will collect your data.

____ Include your data with your word problem.

____ State the statistics that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

____ Create a graph that illustrates your problem. This may be a pie or bar chart or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for your problem.

____ Make 2 copies of the appropriate solution sheet.

____ Record the hypothesis test on the solution sheet, based on your problem. Do a **DRAFT** solution first on one of the solution sheets and check it over carefully. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution

Assignment Checklist

You need to turn in the following typed (12 point) and stapled packet for your final project:

____ **Cover sheet** containing your name, the name of your problem, and the date

____ **The problem**
____ **Data for the problem**
____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
____ **Graphic representation of the data**, created following the guidelines discussed above. Include only graphs that are appropriate and useful.
____ **Sentences interpreting the results of the hypothesis test and the confidence interval** in the context of the situation in the project.

Group Project: Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

____ State the bivariate data your group is going to study.

> Examples
>
> Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____ Describe how your group is going to collect the data (for instance, collect data from the web, survey students on campus).
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____ Conduct your survey. Your number of pairs must be at least 30.
____ Print out a copy of your data.
____ On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____ State the least squares line and the correlation coefficient.
____ On your scatter plot, in a different color, construct the least squares line.
____ Is the correlation coefficient significant? Explain and show how you determined this.
____ Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____ Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____ Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, …, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____ Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
_____ Answer the following, rounded to 2 decimal places:

1. Sample mean =
2. Sample standard deviation =
3. First quartile =
4. Third quartile =
5. Median =
6. 70th percentile =
7. Value that is 2 standard deviations above the mean =
8. Value that is 1.5 standard deviations below the mean =

_____ Construct a histogram displaying your data. Group your data into 6 – 10 intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
_____ In complete sentences, describe the shape of your histogram.
_____ Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in chapter 2 (since you are now using univariate data) to determine which values might be outliers.
_____ Construct a box plot of your data.
_____ Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_____ Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: __________ (keep a copy for your records)
- Part I, Analysis: __________ (keep a copy for your records)
- Entire Project, typed and stapled: __________

____ Cover sheet: names, class time, and name of your study.

____ Part I: label the sections “Intro” and “Analysis.”

____ Part II:

____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.

____ All graphs requested in the project.

____ All calculations requested to support questions in data.

____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges.

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

14.5 Solution Sheets

Solution Sheet: Hypothesis Testing for Single Mean and Single Proportion

Class Time:

Name:

a. H_o:

b. H_a:

c. In words, **CLEARLY** state what your random variable $\bar{X}$ or P' represents.

d. State the distribution to use for the test.

e. What is the test statistic?

f. What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.

g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

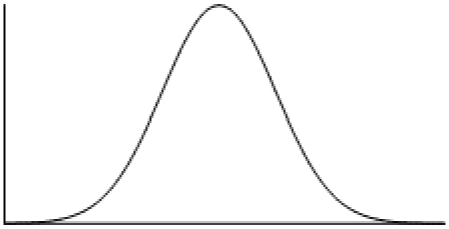


Figure 14.1

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

i. Alpha:

ii. Decision:

iii. Reason for decision:

iv. Conclusion:

i. Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.

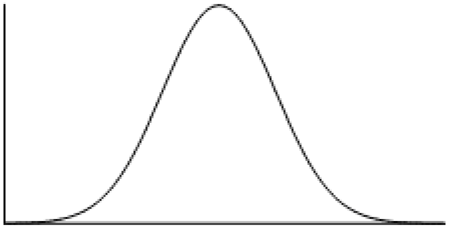


Figure 14.2

Solution Sheet: Hypothesis Testing for Two Means, Paired Data, and Two Proportions

Class Time:

Name:

a. H_o: _______

b. H_a: _______

c. In words, **clearly** state what your random variable $\overline{X}_1 - \overline{X}_2, P_1' - P_2'$- or $\overline{X}_d$ represents.

d. State the distribution to use for the test.

e. What is the test statistic?

f. What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.

g. Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

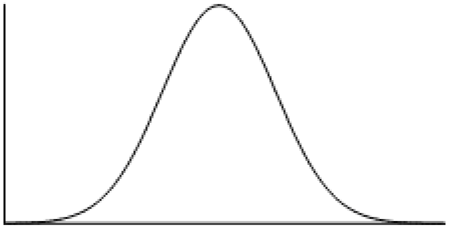


Figure 14.3

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

i. Alpha:

ii. Decision:

iii. Reason for decision:

iv. Conclusion:

i. In complete sentences, explain how you determined which distribution to use.

Solution Sheet: The Chi-Square Distribution

Class Time:

Name:

a. H_o: _______

b. H_a:

c. What are the degrees of freedom?

d. State the distribution to use for the test.

e. What is the test statistic?

f. What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.

g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

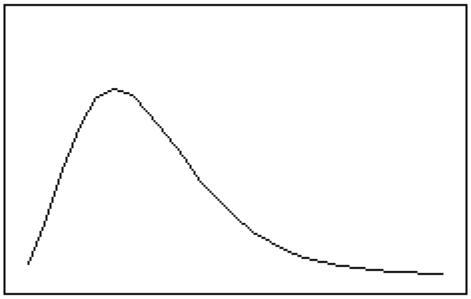


Figure 14.4

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences.**

i. Alpha:

ii. Decision:

iii. Reason for decision:

iv. Conclusion:

Solution Sheet: F Distribution and ANOVA

Class Time:

Name:

a. H_o:

b. H_a:

c. $df(n)$ =

d. $df(d)$ =

e. State the distribution to use for the test.

f. What is the test statistic?

g. What is the p-value? In 1 – 2 complete sentences, explain what the p-value means for this problem.

h. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

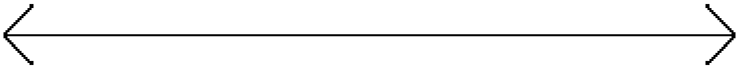


Figure 14.5

i. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.

i. Alpha:

ii. Decision:

iii. Reason for decision:

iv. Conclusion:

14.6 English Phrases Written Mathematically

English Phrases Written Mathematically

Table 14.14

When the English says:	Interpret this as:
$X$is at least 4.	$X \geq 4$
XThe minimum is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
XThe maximum is 4.	$X \leq 4$
$X$is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
Xdoes not exceed 4.	$X \leq 4$
$X$is greater than 4.	$X > 4$
XThere are more than 4.	$X > 4$
Xexceeds 4.	$X > 4$
$X$is less than 4.	$X < 4$
XThere are fewer than 4.	$X < 4$
$X$is 4.	$X = 4$
$X$is equal to 4.	$X = 4$
$X$is the same as 4.	$X = 4$
$X$is not 4.	$X \neq 4$
$X$is not equal to 4.	$X \neq 4$
$X$is not the same as 4.	$X \neq 4$
$X$is different than 4.	$X \neq 4$

14.7 Symbols and their Meanings

Table 14.15 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{}$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q1	Quartile one	the first quartile
Descriptive Statistics	Q2	Quartile two	the second quartile
Descriptive Statistics	Q3	Quartile three	the third quartile
Descriptive Statistics	IQR	inter-quartile range	Q3-Q1=IQR
Descriptive Statistics	$\bar{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	s s_x sx	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s-squared	sample variance
Descriptive Statistics	σ σ_x σx	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma-squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	{}	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A \mid B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A\text{or}B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A\text{and}B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	P(A')	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P\left(G_1\right)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	X~	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	$\neq$	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	pdf	prob. density function	same

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	Exp	exponential distribution	same
Continuous Random Variables	k	k	critical value
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	$\bar{x}$	X-bar	the random variable X-bar
The Central Limit Theorem	μ_x	mean of X	the average of X
The Central Limit Theorem	$\mu_{\bar{x}}$	mean of X-bar	the average of X-bar
The Central Limit Theorem	σ_x	standard deviation of X	same
The Central Limit Theorem	$\sigma_{\bar{x}}$	standard deviation of X-bar	same
The Central Limit Theorem	ΣX	sum of X	same
The Central Limit Theorem	Σx	sum of x	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same
Confidence Intervals	t	student-t distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student-t with a/2 area in right tail	same
Confidence Intervals	p' $\hat{p}$	p-prime; p-hat	sample proportion of success
Confidence Intervals	q' $\hat{q}$	q-prime; q-hat	sample proportion of failure
Hypothesis Testing	H_0	H-naught, H-sub 0	null hypothesis
Hypothesis Testing	H_a	H-a, H-sub a	alternate hypothesis
Hypothesis Testing	H_1	H-1, H-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	X1-bar minus X2-bar	difference in sample means
	$\mu_1 - \mu_2$	mu-1 minus mu-2	difference in population means
	$P'_1 - P'_2$	P1-prime minus P2-prime	difference in sample proportions

Chapter (1st used)	Symbol	Spoken	Meaning
	$p_1 - p_2$	p1 minus p2	difference in population proportions
Chi-Square Distribution	x^2	Ky-square	Chi-square
	O	Observed	Observed frequency
	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + \text{bx}$	y equals a plus b-x	equation of a line
	^ y	y-hat	estimated value of y
	r	correlation coefficient	same
	ε	error	same
	SSE	Sum of Squared Errors	same
	1.9s	1.9 times s	cut-off value for outliers
F-Distribution and ANOVA	F	F-ratio	F ratio

14.8 Formulas

Formula

$n! = n(n-1)(n-2)...(1)$

$0! = 1$

Formula

$$\binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

Formula

$X \sim B(n, p)$

$P\left(X = x\right) = \binom{n}{x}p^x q^{n-x}$, for $x = 0, 1, 2, ..., n$

Formula

$X \sim G(p)$

$P\left(X = x\right) = q^{x-1}p$, for $x = 1, 2, 3, ...$

Formula

X~H(r, b, n)

$$P\left(X = x\right) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}}\right)$$

Formula

X~P(μ)

$$P\left(X = x\right) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula

$X \sim U(a, b)$

$f\left(X\right) = \frac{1}{b-a}$, $a < x < b$

Formula

$X \sim \text{Exp}(m)$

$f\left(x\right) = me^{-\text{mx}}$, $m > 0, x \geq 0$

Formula

$X \sim N\left(\mu, \sigma^2\right)$

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

Formula

$$\Gamma\left(z\right)=\int_0^\infty x^{z-1}e^{-x}\mathrm{dx} \; z>0$$

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$$

$\Gamma(m+1)=m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1)=\mathrm{a}\Gamma(a)$

Formula

$X \sim t_{\mathrm{df}}$

$$f\left(x\right)=\frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}}\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\mathrm{n}\pi}\Gamma\left(\frac{n}{2}\right)}$$

$$X=\frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0,1)$, $Y \sim \mathrm{X}^2_{\mathrm{df}}$,n = degrees of freedom

Formula

$X \sim \mathrm{X}^2_{\mathrm{df}}$

$$f\left(x\right)=\frac{x^{\frac{n-2}{2}}e^{\frac{-x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$$, $x>0$, n = positive integer and degrees of freedom

Formula

$X \sim F_{\mathrm{df}(n),\,\mathrm{df}(d)}$

df(n) = degrees of freedom for the numerator

df(d) = degrees of freedom for the denominator

$$f\left(x\right)=\frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)}\left(\frac{u}{v}\right)^{\frac{u}{2}}x^{\left(\frac{u}{2}-1\right)}\left[1+\left(\frac{u}{v}\right)x^{-.5(u+v)}\right]$$

$X=\frac{Y_u}{W_v}$, Y, W are chi-square

14.9 Notes for the TI-83, 83+, 84 Calculator

Quick Tips

Legend

- □ represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast

Press 2nd, then hold ▲ to increase the contrast or ▼ to decrease the contrast.

To capitalize letters and words

Press ALPHA to get one capital letter, or press 2nd, then ALPHA to set all button presses to capital letters. You can return to the top-level button values by pressing ALPHA again.

To correct a mistake

If you hit a wrong button, just hit CLEAR and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, and 84 using E notation, such that...

- 4.321 E 4 = 4.321 × 10^4
- 4.321 E -4 = 4.321 × 10^{-4}

To transfer programs or equations from one calculator to another:

Both calculators: Insert your respective end of the link cable cable and press 2nd, then `[LINK]`.

Calculator receiving information:

1. Use the arrows to navigate to and select <`RECEIVE`>
2. Press ENTER

Calculator sending information:

1. Press appropriate number or letter.
2. Use up and down arrows to access the appropriate item.
3. Press ENTER to select item to transfer.
4. Press right arrow to navigate to and select <`TRANSMIT`>.
5. Press ENTER

> ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press 2nd, then `[QUIT]` To exit when done.

Manipulating One-Variable Statistics

> These directions are for entering data with the built-in statistical program.

Table 14.16 Sample Data We are manipulating 1-variable statistics.

Data	Frequency
-2	10
-1	3
0	4
1	5
3	8

To begin:

1. Turn on the calculator.

   ON

2. Access statistics mode.

   STAT

3. Select <`4:ClrList`> to clear data from lists, if desired.

   4 , ENTER

4. Enter list `[L1]` to be cleared.

   2nd , `[L1]` , ENTER

5. Display last instruction.

   2nd , `[ENTRY]`

6. Continue clearing remaining lists in the same fashion, if desired.

   ◀ , 2nd , [L2] , ENTER
7. Access statistics mode.

   STAT
8. Select <1:Edit . . .>

   ENTER
9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1])
   - Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad)

     (−) , 9 , ENTER
   - Continue in the same manner until all data values are entered.
10. In [L2], enter the frequencies for each data value in [L1].
    - Type in a frequency and enter it. (If a data value appears only once, the frequency is "1")

      4 , ENTER
    - Continue in the same manner until all data values are entered.
11. Access statistics mode.

    STAT
12. Navigate to <CALC>
13. Access <1:1-var Stats>

    ENTER
14. Indicate that the data is in [L1]...

    2nd , [L1] , ,
15. ...and indicate that the frequencies are in [L2].

    2nd , [L2] , ENTER
16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

We will assume that the data is already entered

We will construct 2 histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

   2nd , [STAT PLOT]
2. Select <1:plot 1> To access plotting - first graph.

   ENTER
3. Use the arrows navigate go to <ON> to turn on Plot 1.

   <ON> , ENTER
4. Use the arrows to go to the histogram picture and select the histogram.
   ENTER
5. Use the arrows to navigate to <Xlist>
6. If "L1" is not selected, select it.

   2nd , [L1] , ENTER
7. Use the arrows to navigate to <Freq>.
8. Assign the frequencies to [L2].

   2nd , [L2] , ENTER
9. Go back to access other graphs.

   2nd , [STAT PLOT]

10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.

   Y=
2. Select each equal sign (=).

   ▼ ▶ ENTER
3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.

   Y=
2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

   ▼ ▶ CLEAR
3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.

   ZOOM
2. Select `<9:ZoomStat>`

   9
3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access
   WINDOW
   to set the graph parameters.
2.
   - X_{min} = −2.5
   - X_{max} = 3.5
   - X_{scl} = 1 (width of bars)
   - Y_{min} = 0
   - Y_{max} = 10
   - Y_{scl} = 1 (spacing of tick marks on y-axis)
   - X_{res} = 1
3. Access
   GRAPH
   to see the histogram.

To draw box plots:

1. Access graphing mode.

   2nd , `[STAT PLOT]`
2. Select `<1:Plot 1>` to access the first graph.

   ENTER
3. Use the arrows to select `<ON>` and turn on Plot 1.

   ENTER
4. Use the arrows to select the box plot picture and enable it.

   ENTER
5. Use the arrows to navigate to `<Xlist>`
6. If "L1" is not selected, select it.

   2nd , `[L1]` , ENTER
7. Use the arrows to navigate to `<Freq>`.
8. Indicate that the frequencies are in `[L2]`.

   2nd , `[L2]` , ENTER
9. Go back to access other graphs.

2nd , `[STAT PLOT]`

10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.

GRAPH , `[STAT PLOT]`

Linear Regression

Sample Data

The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970 - 1995 was:

Table 14.17 The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

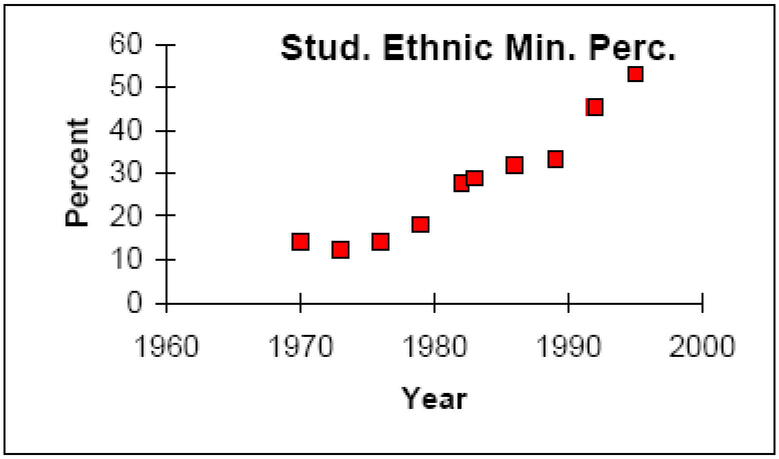


Figure 14.6 Student Ethnic Minority Percentage By hand, verify the scatterplot above.

> The TI-83 has a built-in linear regression feature, which allows the data to be edited.The x-values will be in `[L1]`; the y-values in `[L2]`.

To enter data and do linear regression:

1. ON Turns calculator on

   ON

2. Before accessing this program, be sure to turn off all plots.
   - Access graphing mode.

     2nd , `[STAT PLOT]`
   - Turn off all plots.

     4 , ENTER
3. Round to 3 decimal places. To do so:
   - Access the mode menu.

MODE , `[STAT PLOT]`

- ◦ Navigate to <`Float`> and then to the right to <3>.

  ▼ ▶

- ◦ All numbers will be rounded to 3 decimal places until changed.

  ENTER

4. Enter statistics mode and clear lists `[L1]` and `[L2]`, as describe above.

   STAT , 4

5. Enter editing mode to insert values for x and y.

   STAT , ENTER

6. Enter each value. Press
   ENTER
   to continue.

To display the correlation coefficient:

1. Access the catalog.

   2nd , `[CATALOG]`

2. Arrow down and select <`DiagnosticOn`>

   ▼... , ENTER , ENTER

3. *r* and r^2 will be displayed during regression calculations.
4. Access linear regression.

   STAT ▶

5. Select the form of *y* = *a* + *bx*

   8 , ENTER

The display will show:

LinReg

- *y* = *a* + *bx*
- *a* = −3176.909
- *b* = 1.617
- r^2 = 0.924
- *r* = 0.961

This means the Line of Best Fit (Least Squares Line) is:

- *y* = −3176.909 + 1.617*x*
- Percent = −3176.909 + 1.617(year #)

The correlation coefficient *r* = 0.961

To see the scatter plot:

1. Access graphing mode.

   2nd , `[STAT PLOT]`

2. Select <`1:plot 1`> To access plotting - first graph.

   ENTER

3. Navigate and select <`ON`> to turn on Plot 1.

   <`ON`> ENTER

4. Navigate to the first picture.
5. Select the scatter plot.

   ENTER

6. Navigate to <`Xlist`>
7. If `[L1]` is not selected, press
   2nd
   , `[L1]` to select it.
8. Confirm that the data values are in `[L1]`.

`<ON>` ENTER

9. Navigate to `<Ylist>`
10. Select that the frequencies are in `[L2]`.

    2nd , `[L2]` , ENTER
11. Go back to access other graphs.

    2nd , `[STAT PLOT]`
12. Use the arrows to turn off the remaining plots.
13. Access
    WINDOW
    to set the graph parameters.
    - X_{min} = 1970
    - X_{max} = 2000
    - X_{scl} = 10 (spacing of tick marks on x-axis)
    - Y_{min} = −0.05
    - Y_{max} = 60
    - Y_{scl} = 10 (spacing of tick marks on y-axis)
    - X_{res} = 1
14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press
    GRAPH
    to see the scatter plot.

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

   Y=
2. Access the vars menu and navigate to `<5: Statistics>`

   VARS , 5
3. Navigate to `<EQ>`.
4. `<1: RegEQ>` contains the regression equation which will be entered in Y1.

   ENTER
5. Press
   GRAPH
   . The regression line will be superimposed over scatter plot.

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

   2nd, `[LIST]`, `<RESID>`
2. Confirm twice to view the list of residuals. Use the arrows to select them.

   ENTER , ENTER
3. The critical point for an outlier is: $1.9\sqrt{\frac{SSE}{n-2}}$ where:
   - n = number of pairs of data
   - SSE = sum of the squared errors
   - $\sum\left(\text{residual}^2\right)$
4. Store the residuals in `[L3]`.

   STO► , 2nd , `[L3]` , ENTER
5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n - 2 = 8$

   2nd , `[L3]` , x^2 , ÷ , 8
6. Store this value in `[L4]`.

   STO► , 2nd , `[L4]` , ENTER
7. Calculate the critical value using the equation above.

1 , . , 9 , X , 2nd , `[V]` , 2nd , `[LIST]` ▶ , ▶ , 5 , 2nd , `[L4]` ,) ,) , ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in `[L3]` to 7.64 . If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of y for various x-values:

There are various ways to determine estimates for "y". One way is to substitute values for "x" in the equation. Another way is to use the TRACE on the graph of the regression line.

TI-83, 83+, 84 instructions for distributions and tests

Distributions

Access `DISTR` (for "Distributions").

For technical assistance, visit the Texas Instruments website at http://www.ti.com (http://www.ti.com) and enter your calculator model into the "search" box.

Binomial Distribution
- `binompdf(n,p,x)` corresponds to P(X = x)
- `binomcdf(n,p,x)` corresponds to P(X ≤ x)
- To see a list of all probabilities for x: 0, 1, . . . , n, leave off the "x" parameter.

Poisson Distribution
- `poissonpdf(λ,x)` corresponds to P(X = x)
- `poissoncdf(λ,x)` corresponds to P(X ≤ x)

Continuous Distributions (general)
- −∞ uses the value -1EE99 for left bound
- +∞ uses the value 1EE99 for right bound

Normal Distribution
- `normalpdf(x,μ,σ)` yields a probability density function value (only useful to plot the normal curve, in which case "x" is the variable)
- `normalcdf(left bound, right bound, μ,σ)` corresponds to P(left bound < X < right bound)
- `normalcdf(left bound, right bound)` corresponds to P(left bound < Z < right bound) - standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: P(X < k) = p
- `invNorm(p)` yields the critical value, k: P(Z < k) = p for the standard normal

Student-t Distribution
- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-t curve, in which case "x" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to P(left bound < t < right bound)

Chi-square Distribution
- X^2`pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "x" is the variable)
- X^2`cdf(left bound, right bound, df)` corresponds to P(left bound < X^2 < right bound)

F Distribution
- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the F curve, in which case "x" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to P(left bound < F < right bound)

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the Confidence Intervals and Hypothesis Tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals
- `ZInterval` is the confidence interval for mean when σ is known
- `TInterval` is the confidence interval for mean when σ is unknown; s estimates σ.
- `1-PropZInt` is the confidence interval for proportion

The confidence levels should be given as percents (ex. enter "`95`" for a 95% confidence level).

Hypothesis Tests
- `Z-Test` is the hypothesis test for single mean when σ is known
- `T-Test` is the hypothesis test for single mean when σ is unknown; s estimates σ.
- `2-SampZTest` is the hypothesis test for 2 independent means when both σ's are known
- `2-SampTTest` is the hypothesis test for 2 independent means when both σ's are unknown
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for 2 proportions.
- X^2`-Test` is the hypothesis test for independence.
- X^2`GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "μ∅" represents the null hypothesis. For a test of a single proportion, "p∅" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

A TABLES

When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- **Student-t table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)**
- **Normal table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm)**
- **Chi-Square table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)**
- **F-table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm)**
- All four tables can be accessed by going to **http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm (http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm)**

95% Critical Values of the Sample Correlation Coefficient Table

- **95% Critical Values of the Sample Correlation Coefficient**

  The url for this table is http://cnx.org/content/m17098/latest/

Index

ATTRIBUTIONS

Collection: **Collaborative Statistics**
Edited by: Barbara Illowsky and Susan Dean
URL: **http://cnx.org/content/col10522/1.39/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Preface to "Collaborative Statistics"**
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16026/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Additional Resources**
By: Barbara Illowsky and Susan Dean
URL: **http://cnx.org/content/m18746/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Collaborative Statistics: Author Acknowledgements**
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16308/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Collaborative Statistics: Student Welcome Letter**
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16305/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Sampling and Data: Introduction**
Used here as: Sampling and Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16008/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Statistics**
Used here as: Statistics
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16020/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Probability**
Used here as: Probability
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16015/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Key Terms**
Used here as: Key Terms
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16007/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Data**
Used here as: Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16005/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Sampling**
Used here as: Sampling
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16014/1.17/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Variation and Critical Evaluation**
Used here as: Variation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16021/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Answers and Rounding Off**
Used here as: Answers and Rounding Off
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16006/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Sampling and Data: Frequency, Relative Frequency, and Cumulative Frequency**
Used here as: Frequency
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16012/1.19/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Summary**
Used here as: Summary
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16023/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Practice 1**
Used here as: Practice: Sampling and Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16016/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16010/1.18/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Sampling and Data: Data Collection Lab I**
Used here as: Lab 1: Data Collection
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16004/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Sampling and Data: Sampling Experiment Lab II**
Used here as: Lab 2: Sampling Experiment
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16013/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Introduction**
Used here as: Descriptive Statistics
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16300/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Displaying Data**
Used here as: Displaying Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16297/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs**
Used here as: Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16849/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Histogram**
Used here as: Histograms
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16298/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Box Plot**
Used here as: Box Plots
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16296/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Measuring the Location of the Data**
Used here as: Measures of the Location of the Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16314/1.17/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Measuring the Center of the Data**
Used here as: Measures of the Center of the Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17102/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Skewness and the Mean, Median, and Mode**
Used here as: Skewness and the Mean, Median, and Mode
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17104/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Measuring the Spread of the Data**
Used here as: Measures of the Spread of the Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17103/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16310/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Practice 1**
Used here as: Practice 1: Center of the Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16312/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Descriptive Statistics: Practice 2**
Used here as: Practice 2: Spread of the Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17105/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16801/1.24/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Descriptive Statistics: Descriptive Statistics Lab**
Used here as: Lab: Descriptive Statistics
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16299/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Introduction**
Used here as: Probability Topics
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16838/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Terminology**
Used here as: Terminology
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16845/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Independent & Mutually Exclusive Events**
Used here as: Independent and Mutually Exclusive Events
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16837/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Two Basic Rules of Probability**
Used here as: Two Basic Rules of Probability
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16847/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Contingency Tables**
Used here as: Contingency Tables
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16835/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Venn Diagrams (optional)**
Used here as: Venn Diagrams (optional)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16848/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Tree Diagrams (optional)**
Used here as: Tree Diagrams (optional)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16846/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Probability Topics: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16843/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Practice**
Used here as: Practice 1: Contingency Tables
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16839/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Practice II**
Used here as: Practice 2: Calculating Probabilities
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16840/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16836/1.20/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Probability Topics: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16842/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Probability Topics: Probability Lab**
Used here as: Lab: Probability Topics
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16841/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Discrete Random Variables: Introduction**
Used here as: Discrete Random Variables
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16825/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Probability Distribution Function (PDF) for a Discrete Random Variable**
Used here as: Probability Distribution Function (PDF) for a Discrete Random Variable
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16831/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Mean or Expected Value and Standard Deviation**
Used here as: Mean or Expected Value and Standard Deviation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16828/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Common Discrete Probability Distribution Functions**
Used here as: Common Discrete Probability Distribution Functions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16821/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Binomial**
Used here as: Binomial
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16820/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Geometric (optional)**
Used here as: Geometric (optional)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16822/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Hypergeometric (optional)**
Used here as: Hypergeometric (optional)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16824/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Poisson (optional)**
Used here as: Poisson
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16829/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Summary of the Discrete Probability Functions**
Used here as: Summary of Functions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16833/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Practice 1: Discrete Distributions**
Used here as: Practice 1: Discrete Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16830/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Practice 2: Binomial Distribution**
Used here as: Practice 2: Binomial Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17107/1.18/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Practice 3: Poisson Distribution**
Used here as: Practice 3: Poisson Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17109/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Practice 4: Geometric Distribution**
Used here as: Practice 4: Geometric Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17108/1.17/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Practice 5: Hypergeometric Distribution**
Used here as: Practice 5: Hypergeometric Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17106/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16823/1.20/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16832/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Lab I**
Used here as: Lab 1: Discrete Distribution (Playing Card Experiment)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16827/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Discrete Random Variables: Lab II**
Used here as: Lab 2: Discrete Distribution (Lucky Dice Experiment)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16826/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Introduction**
Used here as: Continuous Random Variables
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16808/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Continuous Probability Functions**
Used here as: Continuous Probability Functions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16805/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: The Uniform Distribution**
Used here as: The Uniform Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16819/1.17/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: The Exponential Distribution**
Used here as: The Exponential Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16816/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Summary of The Uniform and Exponential Probability Distributions**
Used here as: Summary of the Uniform and Exponential Probability Distributions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16813/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Continuous Random Variables: Practice 1**
Used here as: Practice 1: Uniform Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16812/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Practice 2**
Used here as: Practice 2: Exponential Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16811/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16807/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16810/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Continuous Random Variables: Lab I**
Used here as: Lab: Continuous Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16803/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Normal Distribution: Introduction**
Used here as: The Normal Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16979/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Standard Normal Distribution**
Used here as: The Standard Normal Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16986/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Normal Distribution: Z-scores**
Used here as: Z-scores
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16991/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Areas to the Left and Right of x**
Used here as: Areas to the Left and Right of x
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16976/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Normal Distribution: Calculations of Probabilities**
Used here as: Calculations of Probabilities
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16977/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16987/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Practice**
Used here as: Practice: The Normal Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16983/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16978/1.20/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Normal Distribution: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16985/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Normal Distribution: Normal Distribution Lab I**
Used here as: Lab 1: Normal Distribution (Lap Times)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16981/1.18/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Normal Distribution: Normal Distribution Lab II**
Used here as: Lab 2: Normal Distribution (Pinkie Length)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16980/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Introduction**
Used here as: The Central Limit Theorem
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16953/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Central Limit Theorem for Sample Means (Averages)**
Used here as: The Central Limit Theorem for Sample Means (Averages)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16947/1.21/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Central Limit Theorem for Sums**
Used here as: The Central Limit Theorem for Sums
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16948/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Using the Central Limit Theorem**
Used here as: Using the Central Limit Theorem
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16958/1.20/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16956/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Practice**
Used here as: Practice: The Central Limit Theorem
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16954/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16952/1.23/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16955/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Central Limit Theorem Lab I**
Used here as: Lab 1: Central Limit Theorem (Pocket Change)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16950/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Central Limit Theorem: Central Limit Theorem Lab II**
Used here as: Lab 2: Central Limit Theorem (Cookie Recipes)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16945/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Confidence Intervals: Introduction**
Used here as: Confidence Intervals
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16967/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Confidence Intervals: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal**
Used here as: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16962/1.19/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Confidence Intervals: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T**
Used here as: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16959/1.21/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Confidence Intervals: Confidence Interval for a Population Proportion**
Used here as: Confidence Interval for a Population Proportion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16963/1.19/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Confidence Intervals: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16973/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Practice 1**
Used here as: Practice 1: Confidence Intervals for Averages, Known Population Standard Deviation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16970/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Practice 2**
Used here as: Practice 2: Confidence Intervals for Averages, Unknown Population Standard Deviation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16971/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Practice 3**
Used here as: Practice 3: Confidence Intervals for Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16968/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16966/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16972/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Confidence Interval Lab I**
Used here as: Lab 1: Confidence Interval (Home Costs)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16960/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Confidence Interval Lab II**
Used here as: Lab 2: Confidence Interval (Place of Birth)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16961/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Confidence Intervals: Confidence Interval Lab III**
Used here as: Lab 3: Confidence Interval (Womens' Heights)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16964/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Introduction**
Used here as: Hypothesis Testing: Single Mean and Single Proportion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16997/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Null and Alternate Hypotheses**
Used here as: Null and Alternate Hypotheses
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16998/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Outcomes and the Type I and Type II Errors**
Used here as: Outcomes and the Type I and Type II Errors
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17006/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Distribution Needed for Hypothesis Testing**
Used here as: Distribution Needed for Hypothesis Testing
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17017/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Assumptions**
Used here as: Assumption
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17002/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Rare Events**
Used here as: Rare Events
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16994/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Using the Sample to Support One of the Hypotheses**
Used here as: Using the Sample to Support One of the Hypotheses
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16995/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Decision and Conclusion**
Used here as: Decision and Conclusion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16992/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Additional Information**
Used here as: Additional Information
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16999/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Summary of the Hypothesis Test**
Used here as: Summary of the Hypothesis Test
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16993/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Examples**
Used here as: Examples
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17005/1.21/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16996/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Practice 1**
Used here as: Practice 1: Single Mean, Known Population Standard Deviation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17004/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Practice 2**
Used here as: Practice 2: Single Mean, Unknown Population Standard Deviation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17016/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Practice 3**
Used here as: Practice 3: Single Proportion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17003/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17001/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17013/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Single Mean and Single Proportion: Lab**
Used here as: Lab: Hypothesis Testing of a Single Mean and Single Proportion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17007/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Introduction**
Used here as: Hypothesis Testing: Two Population Means and Two Population Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17029/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Means with Unknown Population Standard Deviations**
Used here as: Comparing Two Independent Population Means with Unknown Population Standard Deviations
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17025/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Means with Known Population Standard Deviations**
Used here as: Comparing Two Independent Population Means with Known Population Standard Deviations
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17042/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Proportions**
Used here as: Comparing Two Independent Population Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17043/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Matched or Paired Samples**
Used here as: Matched or Paired Samples
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17033/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Summary of Types of Hypothesis Tests**
Used here as: Summary of Types of Hypothesis Tests
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17044/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Practice 1**
Used here as: Practice 1: Hypothesis Testing for Two Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17027/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing: Two Population Means and Two Population Proportions: Practice 2**
Used here as: Practice 2: Hypothesis Testing for Two Averages
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17039/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Two Means and Two Proportions: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17023/1.16/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Hypothesis Testing of Two Means and Two Proportions: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17021/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Hypothesis Testing of Two Means and Two Proportions: Lab I**
Used here as: Lab: Hypothesis Testing for Two Means and Two Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17022/1.12/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Introduction**
Used here as: The Chi-Square Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17048/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Notation**
Used here as: Notation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17052/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Facts About The Chi-Square Distribution**
Used here as: Facts About the Chi-Square Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17045/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Goodness-of-Fit Test**
Used here as: Goodness-of-Fit Test
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17192/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Test of Independence**
Used here as: Test of Independence
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17191/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Test of a Single Variance**
Used here as: Test of a Single Variance (Optional)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17059/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Summary of Formulas**
Used here as: Summary of Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17058/1.5/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Practice 1**
Used here as: Practice 1: Goodness-of-Fit Test
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17054/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Practice 2**
Used here as: Practice 2: Contingency Tables
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17056/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Practice 3**
Used here as: Practice 3: Test of a Single Variance
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17053/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17028/1.18/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17057/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Lab I**
Used here as: Lab 1: Chi-Square Goodness-of-Fit
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17049/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **The Chi-Square Distribution: Lab II**
Used here as: Lab 2: Chi-Square Test for Independence
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17050/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Introduction**
Used here as: Linear Regression and Correlation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17089/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Linear Equations**
Used here as: Linear Equations
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17086/1.4/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation**
Used here as: Slope and Y-Intercept of a Linear Equation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17083/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Scatter Plots**
Used here as: Scatter Plots
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17082/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: The Regression Equation**
Used here as: The Regression Equation
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17090/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Correlation Coefficient and Coefficient of Determination**
Used here as: Correlation Coefficient and Coefficient of Determination
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17092/1.11/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient**
Used here as: Testing the Significance of the Correlation Coefficient
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17077/1.14/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Prediction**
Used here as: Prediction
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17095/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Outliers**
Used here as: Outliers
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17094/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: 95% Critical Values of the Sample Correlation Coefficient Table**
Used here as: 95% Critical Values of the Sample Correlation Coefficient Table
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17098/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Summary**
Used here as: Summary
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17081/1.4/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Practice**
Used here as: Practice: Linear Regression
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17088/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17085/1.13/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Linear Regression and Correlation: Regression Lab I**
Used here as: Lab 1: Regression (Distance from School)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17080/1.10/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Regression Lab II**
Used here as: Lab 2: Regression (Textbook Cost)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17087/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Linear Regression and Correlation: Regression Lab III**
Used here as: Lab 3: Regression (Fuel Efficiency)
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17079/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Introduction**
Used here as: F Distribution and ANOVA
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17065/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Purpose and Basic Assumption of ANOVA**
Used here as: ANOVA
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17068/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **F Distribution and ANOVA: The F Distribution And The F Ratio**
Used here as: The F Distribution and the F Ratio
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17076/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **F Distribution and ANOVA: Facts About the F Distribution**
Used here as: Facts About the F Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17062/1.11/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **F Distribution and ANOVA: Test of Two Variances**
Used here as: Test of Two Variances
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17075/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Summary**
Used here as: Summary
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17072/1.3/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Practice**
Used here as: Practice: ANOVA
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17067/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Homework**
Used here as: Homework
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17063/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: Review**
Used here as: Review
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17070/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **F Distribution and ANOVA: ANOVA Lab**
Used here as: Lab: ANOVA
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17061/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Practice Final Exam 1**
Used here as: Practice Final Exam 1
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16304/1.16/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Practice Final Exam 2**
Used here as: Practice Final Exam 2
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16303/1.15/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Data Sets**
Used here as: Data Sets
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17132/1.5/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Collaborative Statistics: Projects: Univariate Data**
Used here as: Group Project: Univariate Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17142/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Projects: Continuous Distributions & Central Limit Theorem**
Used here as: Group Project: Continuous Distributions and Central Limit Theorem
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17141/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Projects: Hypothesis Testing Article**
Used here as: Partner Project: Hypothesis Testing - Article
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17140/1.8/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Collaborative Statistics: Projects: Hypothesis Testing Word Problem**
Used here as: Partner Project: Hypothesis Testing - Word Problem
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17144/1.7/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Projects: Bivariate Data, Linear Regression and Univariate Data**
Used here as: Group Project: Bivariate Data, Linear Regression, and Univariate Data
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17143/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Solution Sheets: Hypothesis Testing: Single Mean and Single Proportion**
Used here as: Solution Sheet: Hypothesis Testing for Single Mean and Single Proportion
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17134/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Solution Sheets: Hypothesis Testing: Two Means, Paired Data, Two Proportions**
Used here as: Solution Sheet: Hypothesis Testing for Two Means, Paired Data, and Two Proportions
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17133/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Solution Sheets: The Chi-Square Distribution**
Used here as: Solution Sheet: The Chi-Square Distribution
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17136/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Solution Sheets: F Distribution and ANOVA**
Used here as: Solution Sheet: F Distribution and ANOVA
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m17135/1.5/**

Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: English Phrases Written Mathematically**
Used here as: English Phrases Written Mathematically
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16307/1.5/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Symbols and their Meanings**
Used here as: Symbols and their Meanings
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16302/1.9/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Formulas**
Used here as: Formulas
By: Susan Dean and Barbara Illowsky
URL: **http://cnx.org/content/m16301/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/2.0/**

Module: **Collaborative Statistics: Notes for the TI-83, 83+, 84 Calculator**
Used here as: Notes for the TI-83, 83+, 84 Calculator
By: Barbara Illowsky and Susan Dean
URL: **http://cnx.org/content/m19710/1.6/**
Copyright: Maxfield Foundation
License: **http://creativecommons.org/licenses/by/3.0/**

Module: **Tables**
By: Susan Dean
URL: **http://cnx.org/content/m19138/1.3/**
Copyright: Susan Dean
License: **http://creativecommons.org/licenses/by/2.0/**

ABOUT CONNEXIONS