INTRODUCTORY STATISTICS

3rd Edition

by OpenStax College

Adapted by
Mark Beintema
and Natalia Casper

College Lake County
Connect to Your Future

This page is purposely left blank.

Introductory Statistics

Susan Dean and Barbara Illowsky (Published 2013 by OpenStax College)

Adapted by: College of Lake County Faculty: Mark Beintema and Natalia Casper
Revised July 2018

This project was funded by a grant from the College of Lake County Foundation.

Adapted College of Lake County text can be found at:
http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf

Original text materials for Introductory Statistics by Dean and Illowski 2013 available at:
http://cnx.org/content/col11562/latest/.

Table of Contents (3rd Edition)

Table of Contents (3rd Edition)

This page is purposely left blank.

1 | SAMPLING AND DATA

Figure 1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (Credit: David Sim)

Introduction

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

- Understand basic statistical terminology
- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.

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

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

Included in this chapter are some basic ideas and terms used in Statistics. We will learn how data are gathered and how to distinguish "good" data from "bad."

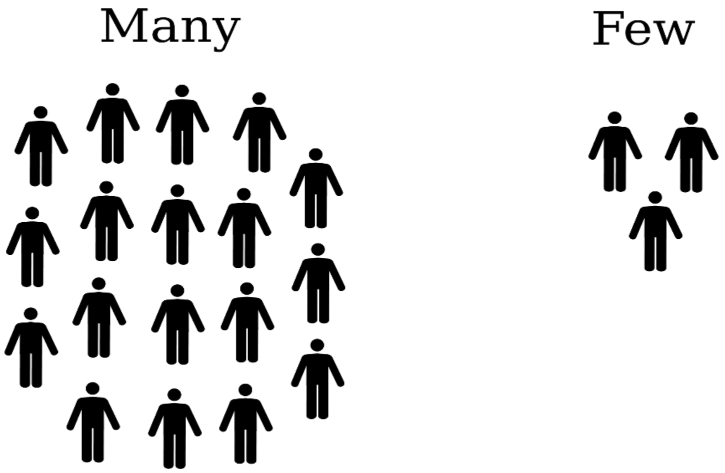
1.1 | Basic Definitions

The science of **Statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. There are two basic branches of Statistics: Descriptive and Inferential. Organizing and summarizing data is called **Descriptive Statistics**. Two general ways to summarize data are with graphs and numerical values, such as an average. Later in the text we will learn formal methods for drawing conclusions from data. These formal methods are called **Inferential Statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. We will encounter what may 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 and in your chosen career.

Key Terms

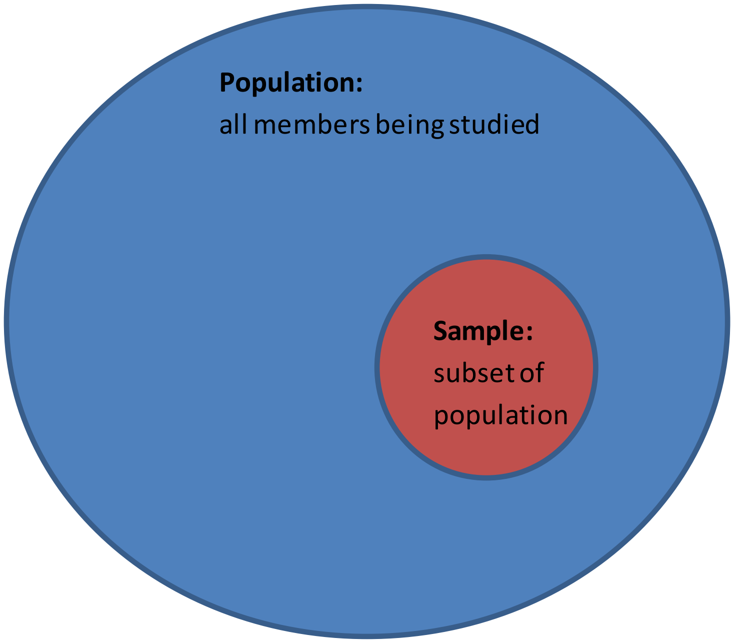
In Statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. For logistical reasons, it is usually not possible to gain access to all of the information from the entire population. So when want to study a population, we usually 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.



https://openclipart.org/detail/216524/many-and-few

Because it takes a lot of time and resources 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, samples of between 1,000 and 2,000 prospective voters are used for opinion polls. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

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



Parameter:
Measurements found using the population. μ is the average of the population. Therefore μ is a parameter.

Statistic:
Measurement found using the sample. $\bar{x}$ is the average of the sample. Therefore $\bar{x}$ is a statistic.

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**, denoted by capital letters such as *X* and *Y*, is a characteristic of interest for each person or thing in a population. Variables may be numerical or categorical. **Numerical variables** take on numerical 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 *Y* is a categorical variable, and some possible values of *Y* would be 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).

The actual values of a variable are called **data** (a single value is a **datum**); these values may be numbers, or words.

Example 1.1

Determine what the key terms refer to in 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.

Solution 1.1:

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

The **sample** is the 100 first year students surveyed at the 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. This average would be represented by μ.

The **statistic** is the average amount of money spent (excluding books) by first year college students in the sample. This average would be represented by $\bar{x}$.

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** would be the actual dollar amounts spent by the first year students. Examples of the data would be \$150, \$200, and \$225.

Try It Σ

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

Example 1.2

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

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

1. ___ Population 2. ___ Statistic 3.___ Parameter

4. ___ Sample 5.___Variable 6.___ Data

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

Solution 1.2: 1. e; 2. b; 3. f; 4. d; 5. a; 6. c

Example 1.3

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

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution 1.3

The **population** consists of all cars containing dummies in the front seat.

The **sample** is the set of 75 randomly selected cars.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries measured in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The possible **data** values would be either: yes, had head injury, or no, did not.

Example 1.4

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

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

Solution to 1.4:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors w1ho have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

1.2 | Data and Sampling

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

- Qualitative
- Quantitative

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

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

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

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in such numbers as □, □/3, 5□/6, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks would be discrete data and the weights of the backpacks would be continuous data.

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

Try It Σ

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

Try It Σ

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

Example 1.7

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

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

Solution 1.7: Answers will vary, but one possibility is:

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

Example 1.8

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.

NOTE: You may collect data as numbers and report it categorically. For example, exam scores for students are recorded throughout the term. At the end of the term, letter grades are reported as A, B, C, D, or F.

Example 1.9

Work collaboratively to determine the data is quantitative or qualitative. For quantitative data, indicate whether they are continuous or discrete.

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

Solution to 1.9:
Quantitative Discrete (a, e, l); Quantitative Continuous (d, f, j, k, n); Qualitative (b, c, g, h, i, m)

Try It Σ

1.9 Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. If quantitative, indicate whether it is continuous or discrete.

Example 1.10

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart, Figure 1.2.

What type of data does this graph show?

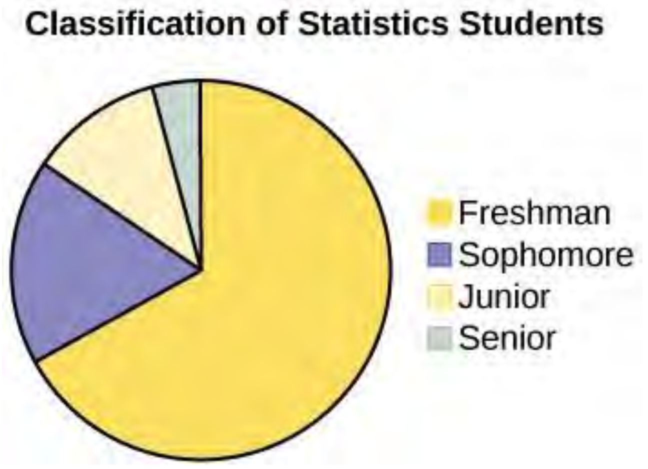


Figure 1.1

Solution 1.10 This chart shows the students in each year, which is **qualitative data**.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we usually use data from 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 sampling methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in **Table 1.2**:

ID	NAME	ID	NAME	ID	NAME
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Statcher
06	Fontecha	17	Pate;	27	Tallai
07	Hong	8	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Table 1.2

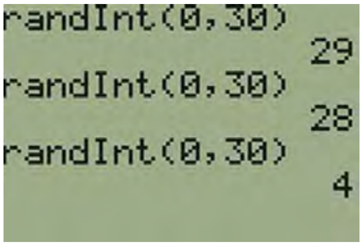
Lisa can then use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

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

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers. The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

Using the TI-83, 83+, 84, 84+ Calculator

To generate random numbers:



```
randInt(0,30)
              29
randInt(0,30)
              28
randInt(0,30)
               4
```

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

Note: We can provide third input get a specified number of random values. E.g. randInt(0, 30, 3) will generate 3 random numbers, which would yield the answer {29, 28, 4}.

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

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

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

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

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

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

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

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

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person). Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. So these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and non-sampling 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 **non-sampling errors**. A defective counting device can cause a non-sampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. In general, 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 sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example 1.11

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

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

Solution 1.11: a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Try It Σ

1.11 You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores for an elementary Statistics class.

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

Instructions: Use the Random Number Generator to pick simple random samples.

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

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

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

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

Example 1.12

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

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

Solution 1.12

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

Try It Σ

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

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

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

Example 1.13

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

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

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

Example 1.13 continued

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

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

It is unlikely that any student is in both samples.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?
b. If these samples are not representative of the entire population, would it be wise to use the results to describe the entire population?
c. Now suppose we take a third sample. We choose ten different part-time students from the disciplines of Chemistry, Math, English, Psychology, Sociology, History, Nursing, Physical Education, Art, and Early Childhood Development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

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

   Is this sample biased? Explain.

Solution 1.13

a. No. The first sample probably consists of science-oriented students. For example, in addition to the chemistry course, some of them are also Calculus or Biology courses. 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 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.
b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.
c. 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.

Try It Σ

1.13 A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Since asking all 20,000 listeners would be impossible, the station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. Of those sampled, 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music. Is this sample is representative of the entire 20,000 listener population?

Variation in Data and in Samples

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 selection methods and your accuracy.

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

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

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

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

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

Critical Evaluation

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

- Problems with samples:
  A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples give results that are inaccurate and not valid.

- Self-selected samples:
  Responses only by people who choose to respond, such as call-in surveys, are often unreliable.

- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions.
  Examples: crash testing cars or medical testing for rare conditions.

- Undue influence: collecting data or asking questions in a way that influences the response.

- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.

- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.

- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context.

- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

1.3 | Experimental Design and Ethics

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

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

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

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

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

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

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

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

Example 1.19

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

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

Solution 1.19

The *population* is men aged 50 to 84.
The *sample* is the 400 men who participated.
The *experimental units* are the individual men in the study. The *explanatory variable* is oral medication.
The *treatments* are aspirin and a placebo.
The *response variable* is whether a subject had a heart attack.

Example 1.20

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

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

Solution 1.20

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.

b. There are two treatments: a floral-scented mask and an unscented mask.

c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.

d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Example 1.21

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

Solution 1.21

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

Try It Σ

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

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

Ethics

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

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

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

[2] Yudhijit Bhattacharjee, *"The Mind of a Con Man," Magazine, New York Times,* April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

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

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

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

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

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

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

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

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

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

3. "*Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel*," Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd- bce5-4385-b9ff- 05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

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

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

Example 1.22

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected. A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.

Example 1.22 continued

b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.

c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution 1.22

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.

b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.

c. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Try It Σ

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

A study is commissioned to determine the favorite brand of fruit juice among teens in California. The survey is commissioned by the seller of a popular brand of apple juice. There are only two types of juice included in the study: apple juice and cranberry juice. Researchers allow participants to see the brand of juice as samples are poured for a taste test. Among the participants, 25% preferred Brand X, 33% preferred Brand Y and 42% had no preference between the two brands. Brand X then references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

KEY TERMS

Average a number that describes the central tendency of the data

Blinding not telling participants which treatment a subject is receiving

Categorical Variable variables that take on values that are names or labels

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

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

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

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

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

Discrete Random Variable a random variable (RV) whose outcomes are counted

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

Experimental Unit any individual or object to be measured

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

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

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

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

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

Numerical Variable variables that take on values that are indicated by numbers

Parameter a number that is used to represent a population characteristic

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

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

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

Qualitative Data See **Data**.

Quantitative Data See **Data**.

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

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

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

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

Sample a subset of the population studied

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

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

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

Sampling without Replacement A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

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

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

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

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

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

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

CHAPTER REVIEW

1.1 Basic Definitions

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

1.2 Data and Sampling

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

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

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.3 Experimental Design and Ethics

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

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

[4] Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/ published/ChanceEthics1.pdf (accessed May 1, 2013).

Exercises for Chapter 1

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

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

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

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

1. population

2. sample

3. parameter

4. statistic

5. variable

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

6. The sampling method was:
a. simple random b. systematic c. stratified d. cluster

7. "Number of times per week" is what type of data?
a. qualitative b. quantitative discrete c. quantitative continuous

8. "Duration (amount of time)" is what type of data?
a. qualitative b. quantitative discrete c. quantitative continuous

9. The colors of the houses around the park are what kind of data?
a. qualitative b. quantitative discrete c. quantitative continuous

10. What is the population for this study?

11. The table below contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

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

Use this data to answer the following questions:

a. What is the proportion of deaths between 2007 and 2012?

b. What percent of deaths occurred before 2001?

c. What is the percent of deaths that occurred in 2003 or after 2010?

d. What is the fraction of deaths that happened before 2012?

e. What kind of data is the number of deaths?

f. Earthquakes are quantified according to the Richter scale, which measures the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?

g. What contributed to the large number of deaths in 2010? In 2004? Explain.

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

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

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

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

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

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in **Table 1.31**. The second study collected the data in **Table 1.32**.

Group	Showed Improvement	No Improvement	Deterioration
Used Program	142	43	15
Did Not use Program	72	110	18

Table 1.31

Group	Showed Improvement	No Improvement	Deterioration
Used Program	105	74	19
Did Not use Program	88	99	12

Table 1.32

16. Given what you know, which study is correct?

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

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

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

20. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from question #18?

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

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

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

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

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

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

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

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

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

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

29. What is the population she is interested in?

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

30. Let X = number of days a Lake Tahoe Community College math student is absent. In this case, X is an example of a:

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

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

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

For the following exercises (32 – 40), identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

32. number of tickets sold to a concert

33. percent of body fat

34. time spent in line to buy groceries

35. number of students enrolled at Evergreen Valley College

36. most-watched television show

37. brand of toothpaste

38. distance to the closest movie theatre

39. age of executives in Fortune 500 companies

40. number of competing computer spreadsheet software packages

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

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

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

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

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

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

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

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

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.

b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.

c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.

d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

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

a. Do you consider the sample size large enough for a study of this type? Why or why not?

b. Based on your "gut feeling," do you believe the percentages accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percentage of the population is actually higher or lower than the sample statistics? Why?

Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."

c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?

d. With the additional information, comment on how accurately you think the sample statistics reflect the
population parameters.

49. The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey as qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?

b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?

c. In the last seven days, on how many days did you exercise for 30 minutes or more?

d. Do you have health insurance coverage?

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

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.

b. What effect does the low response rate have on the reliability of the sample?

c. Are these problems examples of sampling error or non-sampling error?

d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

51. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates. Which of the potential problems with samples discussed in **Section 1.2** could explain this connection?

52. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks: "Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[5] As of April 25, 11 people responded to this question. Each participant answered "NO!" Which of the potential problems with samples discussed in this module could explain this connection?

53. A scholarly article about response rates begins with the following quote: "Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[6] The Pew Research Center for People and the Press admits: "The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more." [7]

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

[5] lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).

[6] Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759 (accessed May 1, 2013).

[7] Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

REFERENCES

1.1 Basic Definitions

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

1.2 Data and Sampling

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May1, 2013).

"The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/ LiteraryDigest.html (accessed May 1, 2013).

"Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/gallup- presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

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

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013). Data from San Jose Mercury News

1.3 Experimental Design and Ethics

"Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/ vitamin-e/ (accessed May 1, 2013).

Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. "Daily Dose of Aspiring Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

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

M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

"Earthquake Information by Year," U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

"Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).
Data from www.forbes.com (accessed May 1, 2013).

"America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

"April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/ airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

2 | DESCRIPTIVE STATISTICS

Figure 2.1 When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Introduction

Chapter Objectives

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

- Create and interpret frequency distributions
- Display data graphically and interpret graphs for Qualitative Data (Bar, Pareto, Pie)
- Display data graphically and interpret graphs for Quantitative Data (Histogram, Frequency Polygon, and Time Series)
- Recognize, describe, and calculate the measures of central tendencies: mean, median, and mode.
- Recognize, describe, and calculate the measures of variation: range, variance, and standard deviation
- Recognize, describe, and calculate the measures of position: quartiles, percentiles, outliers, Chebyshev’s Theorem, and Empirical Rule.

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 how to calculate, and even more importantly, how to interpret these measurements and graphs.

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

https://openclipart.org/detail/19980/graphs

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

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

2.1 | Frequency Distributions

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. This is done by creating a table of frequencies which is known as a **frequency distribution**. There are three frequency distributions to choose from, depending on if the data is categorical or quantitative. When it is quantitative data, there are two frequency distributions to choose from. When calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

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

Ungrouped Frequency Distribution

This first type of frequency distribution is for quantitative data sets. It is best when the range of the data is less than 10 units. **Range** is the difference between the largest data value and the smallest data value. For example, twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5; 6; 3; 3; 2; 4; 8; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Since the range is 6, we will keep each data value separate and not group them together. To create an ungrouped frequency distribution is a simple task. Place the data values from smallest to the largest without skipping any values on the first column. Place the **frequency**, the count of each data value, in the corresponding row of the second column.

Table 2.1 lists the different data values in ascending order and their frequencies. Notice all the data values are listed including 7 which is not listed on the original data set.

DATA VALUES	FREQUENCY (f)
2	3
3	5
4	3
5	6
6	2
7	0
8	1

Table 2.1 Frequency distribution of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to **Table 2.1**, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample, known as **sample size** (n).

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

The following table is known as a relative frequency distribution because frequencies are translated as percent. To find the percent of the first row, take the frequency and divide by the sample size, n. 3/20 = 15%.

DATA VALUE	FREQUENCY (f)	RELATIVE FREQUENCY
2	3	15%
3	5	25%
4	3	15%
5	6	30%
6	2	10%
7	0	0%
8	1	5%

Table 2.2 Frequency distribution of Student Work Hours with Relative Frequencies

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table 2.3. The third row under cumulative relative frequency has the number .55 which was found by adding the previous relative frequencies (.15 + .25 + .15). This represents that 55% of the data has the value of 4 or under.

DATA VALUE	FREQUENCY (f)	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	15%	.15
3	5	25%	.40
4	3	15%	.55
5	6	30%	.85
6	2	10%	.95
7	0	0%	.95
8	1	5%	1

Table 2.3 Frequency distribution of Student Work Hours with Relative Frequencies

Grouped Frequency Distribution

This second type of frequency distribution is also used when there is quantitative data. However, it is used when the range is large and the data values need to be grouped together. For example, 28 students were asked how many hours they worked per week. Their responses, in hours, are as follows:

15; 26; 13; 33; 22; 14; 27; 15; 32; 23; 5; 26; 25; 14; 34; 13; 15; 22; 15; 28; 10; 18; 21; 24; 20; 18; 34; 20;

Here there are too many different data values to list them separately as in the ungrouped frequency distribution. Notice the range is 29 (highest – lowest = 34 – 5). Therefore we need to construct a grouped frequency distribution (Table 2.4) to group data values into classes. A **class** is an interval where the lowest value of the interval is known as the **lower limit** and the highest value of the interval is known as the **upper limit.**

Guidelines for classes:

- There should be between 5 and 20 classes
- Classes must be mutually exclusive (no overlap of data values)
- Classes must be all inclusive and continuous
- Classes must be equal in width

Constructing a Grouped Frequency Distribution:

Find Range (highest data value – lowest data value)
Determine the number of classes (usually the minimum is 5 classes and a maximum of 20 classes)

Find Class Width = $\dfrac{\text{Range}}{\#\text{ of classes}}$ NOTE: Round up

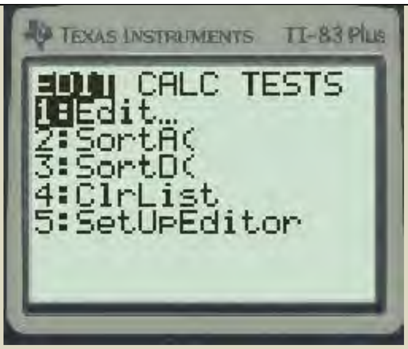
Choose first lower limit (usually the lowest data value)
Create the other lower limits of the classes by adding the class width to the previous lower limit
Create the upper limits by not overlapping the limits

Example 2.1

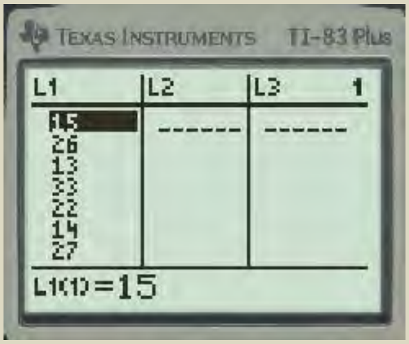
Twenty-eight students were asked how many hours they worked per week. Their responses, in hours, are as follows: 15; 26; 13; 33; 22; 14; 27; 15; 32; 23; 5; 26; 25; 14; 34; 13; 15; 22; 15; 28; 10; 18; 21; 24; 20; 18; 34; 20; construct a grouped frequency distribution using 5 classes.

The TI-83/TI-84 calculator can sort the data values.

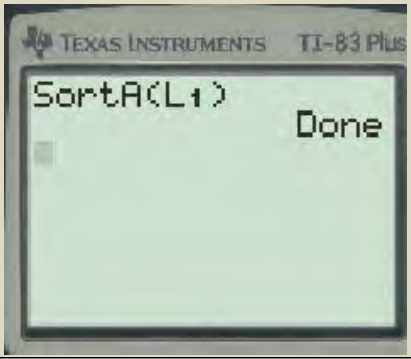
Press STAT
Choose EDIT



Once you have enter the data into L1, 2nd QUIT
Press STAT



Choose #2 SortA(
Press 2nd L1



```
SortA(L1)
                Done
```

To view the sorted list, return to STAT, EDIT

5; 10; 13; 13; 14; 14; 15; 15; 15; 15; 18; 18; 20; 20; 21; 22; 22; 23; 24; 25; 26; 26; 27; 28; 32; 33; 34; 34

1. Range = 34 – 5 = 29
2. Use 5 classes (given in **Exercise 2.1**)
3. Class Width = 29/5 = 5.8 round up to 6
4. First lower limit will be 5 which is the minimum data value
5. The other lower limits will be 11, 17, 23, 29 by adding the class width of 6 to the previous lower limit
6. The first upper limit will be 10 since the next class begins at 11. Using class width again, the other upper limits are 16, 22, 28, 34

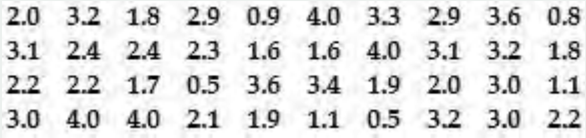
Solution to Example 2.1

CLASSES	FREQUENCY (f)
5 – 10	2
11 – 16	8
17 – 22	7
23 – 28	7
29 – 34	4

Frequency distribution of Student Work Hours.

Try It

2.1 The grade point averages for 40 students are listed below.



2.0	3.2	1.8	2.9	0.9	4.0	3.3	2.9	3.6	0.8
3.1	2.4	2.4	2.3	1.6	1.6	4.0	3.1	3.2	1.8
2.2	2.2	1.7	0.5	3.6	3.4	1.9	2.0	3.0	1.1
3.0	4.0	4.0	2.1	1.9	1.1	0.5	3.2	3.0	2.2

Construct a frequency distribution, a relative frequency distribution, and a cumulative frequency distribution using eight classes. Include the midpoints of the classes.

Solution 'Try It' 2.1:
Range = 4 - .5 = 3.5
Class width = 3.5/8 = .4375 Round up to .5 since the data values are in tenths (one decimal spot) then we round the class width to tenths.

CLASSES	FREQUENCY (f)
.5 - .9	4
1 – 1.4	2
1.5 – 1.9	7
2 – 2.4	9
2.5 – 2.9	2
3-3.4	10
3.5-3.9	2
4-4.4	4

Table 2.5 Frequency distribution of GPA

Frequency Distribution for Qualitative Data

Qualitative data is non-numeric data. Therefore, when creating a table of data values, place the category in the first column and the count of each data value, frequency, in the second column. For example, twenty students are asked their blood type. Their responses are as follows:

A; B; O; A; AB; O; O; A; O; B; A; A; A; O; O; O; B; O; AB; B

Table 2.5 lists the different data values and their frequencies.

DATA VALUE	FREQUENCY (f)
A	6
AB	2
B	4
O	8

Table 2.5 Frequency distribution of Student Blood Types

Frequency distributions (tables) are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data.

2.2 | Graphs for Qualitative Data

There are no strict rules concerning which graphs to use. However, scale of the bars or slices should be accurate. For example if a category is 32% then the slice for the category should not be smaller than a quarter of the circle or bigger than half the circle. Two graphs that are used to display qualitative data are pie charts and bar graphs.

In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category. When creating a pie chart, each slice should be labeled with the category name and the relative frequency (percent).

In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal. For vertical bars, the categories are on the x-axis and frequency or relative frequency are on the y-axis.

Example 2.2

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies).

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

Table 2.6 Fall Term 2007 (Census day)

Look at **Figure 2.1** and **Figure 2.2** and determine which graph (pie or bar) you think displays the comparisons better. It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.

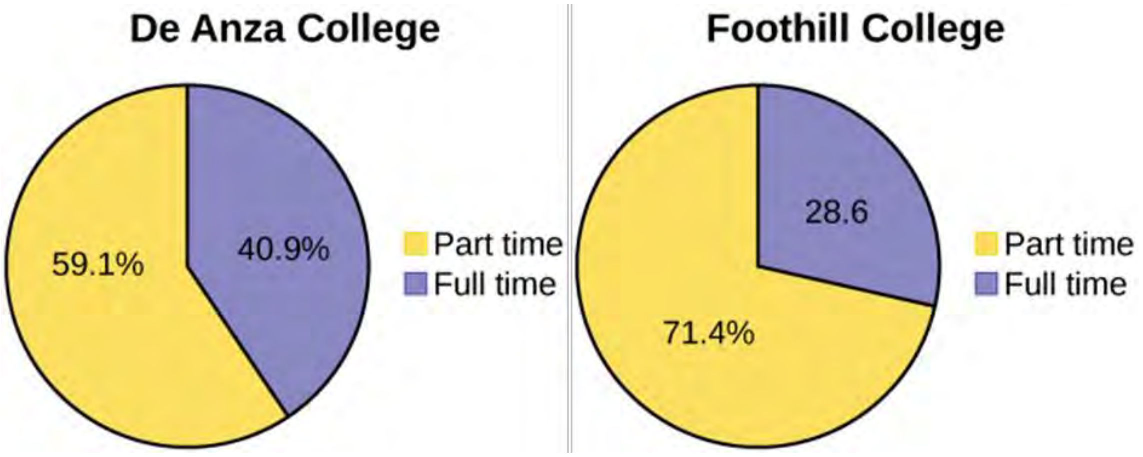


Figure 2.1

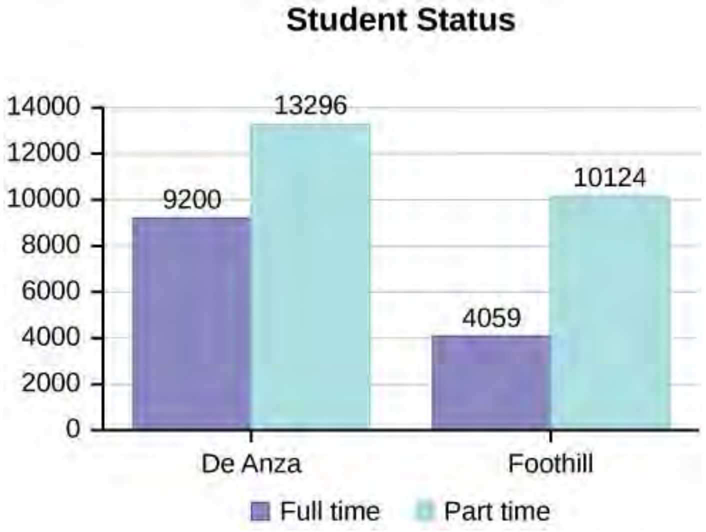


Figure 2.2

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Example 2.3

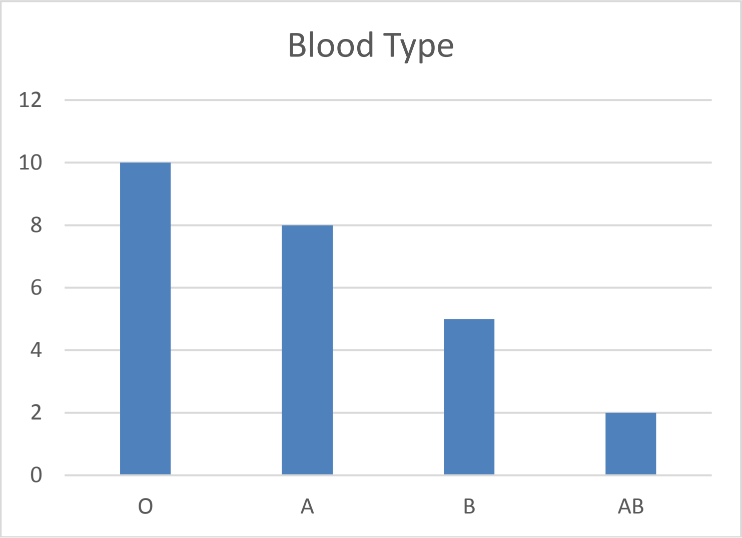
Twenty-five students are asked their blood type. Their responses are as follows:

A; B; O; A; AB; O; O; A; O; B; A; A; A; O; O; O; B; O; AB; B, O, B, O, A, A

DATA VALUE	FREQUENCY (f)
A	8
AB	2
B	5
O	10

Create a Pareto Chart.

Solution to 2.3:
Blood type O has the largest frequency therefore it should be placed first on the x-axis and blood type A follows.



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

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than100% because students can be in more than one category. A **bar graph** is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61%
Total	150.5%

Table 2.7 De Anza College Spring 2010

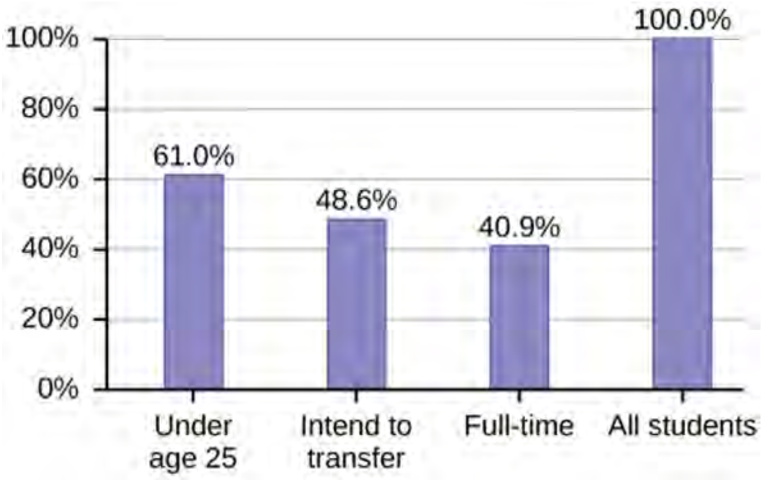


Figure 2.3

Omitting Categories/Missing Data

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

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

Table 2.8 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

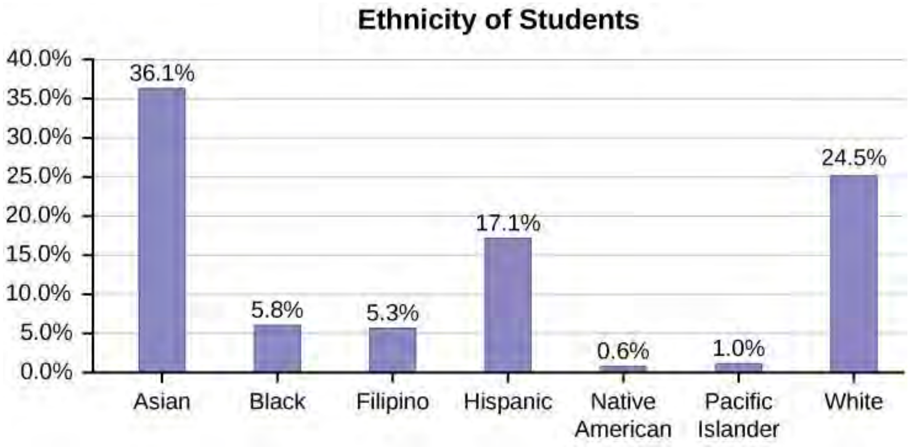


Figure 2.4

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us. This particular bar graph in **Figure 2.5** can be difficult to understand visually. Therefore a pareto chart is easier to read and interpret. The graph in **Figure 2.6** is a Pareto chart.

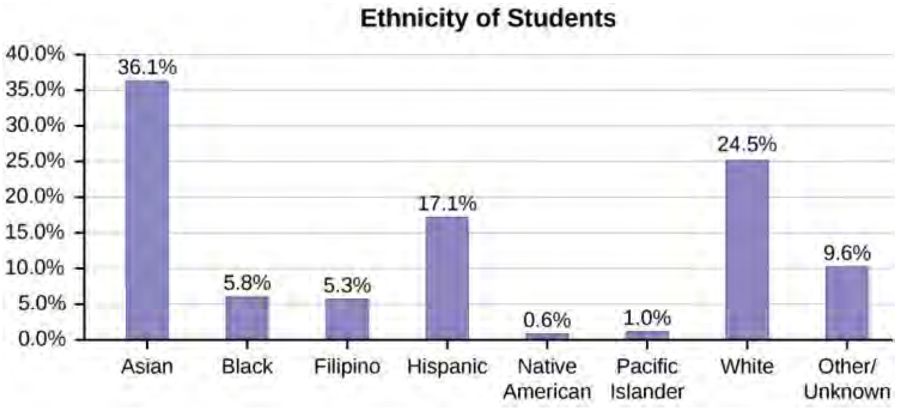


Figure 2.5 Bar Graph with Other/Unknown Category

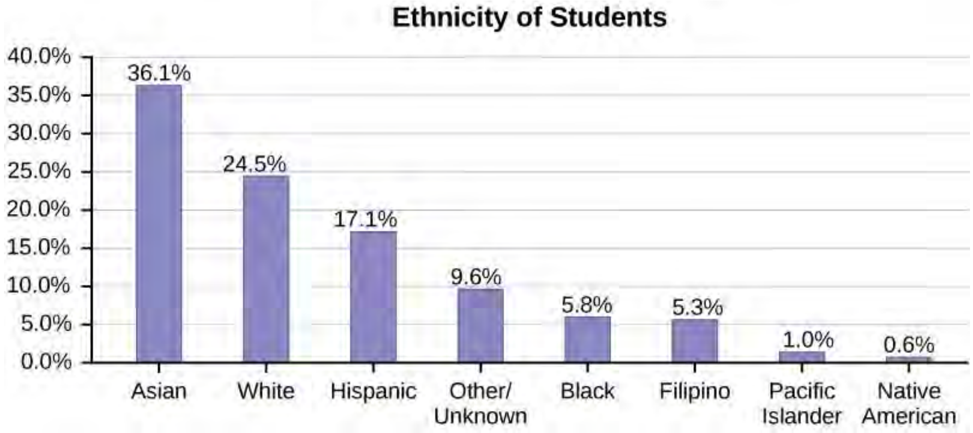


Figure 2.6 Pareto Chart (Bars sorted in descending order)

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in **Figure 2.7b** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in **Figure 2.7a**.

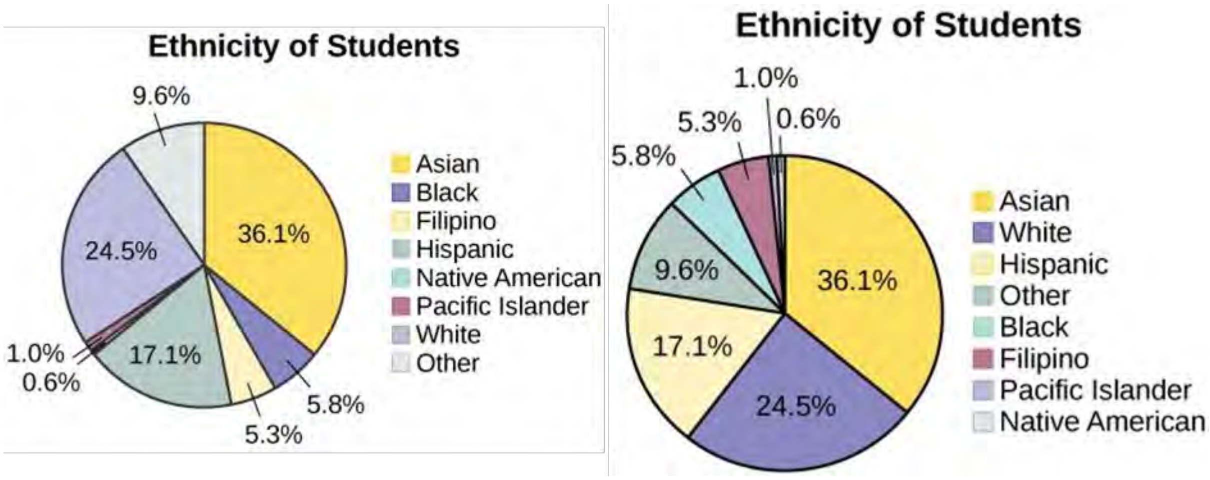


Figure 2.7 a

Figure 2.7 b

2.3 | Graphs for Quantitative Data

Histogram:

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

A histogram consists of contiguous (adjoining) bars. 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). Horizontal axis uses the class boundaries. The vertical axis is labeled either frequency or relative frequency (percent or probability). The graph will have the same shape with either label. The histogram can give you the shape of the data, the center, and the spread of the data.

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

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

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

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40,and RF = 3/40 = .075 = 7.5% of the students received 90–100%.

To construct a histogram:

1. Create class boundaries on the grouped frequency distribution. Choose a starting point for the first interval to be less than the smallest data value. A convenient starting point is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

2. Place frequency or relative frequency on the y-axis. Scale is important.

3. Draw bars as high as the frequency for each class interval within each boundary
   The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data

Example 2.4

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

Create a relative frequency histogram using the frequency distribution that follows:

CLASSES	FREQUENCY (f)
60.0 – 61.9	5
62.0 – 63.9	3
64.0 – 65.9	15
66.0 – 67.9	40
68.0 – 69.9	17
70.0 – 71.9	12
72.0 – 73.9	7
74.0 – 75.9	1

Table 2.8 Frequency distribution of Heights

Solution to Example 2.4

First find the relative frequencies and class boundaries for each class.

BOUNDARIES	RELATIVE FREQUENCY
59.95 – 61.95	.05
61.95 – 63.95	.03
63.95 – 65.95	.15
65.95 – 67.95	.4
67.95 – 69.95	.17
69.95– 71.95	.12
71.95 – 73.95	.07
73.95 – 75.95	.01

Second place the boundaries on the x-axis and the relative frequency on the y-axis.

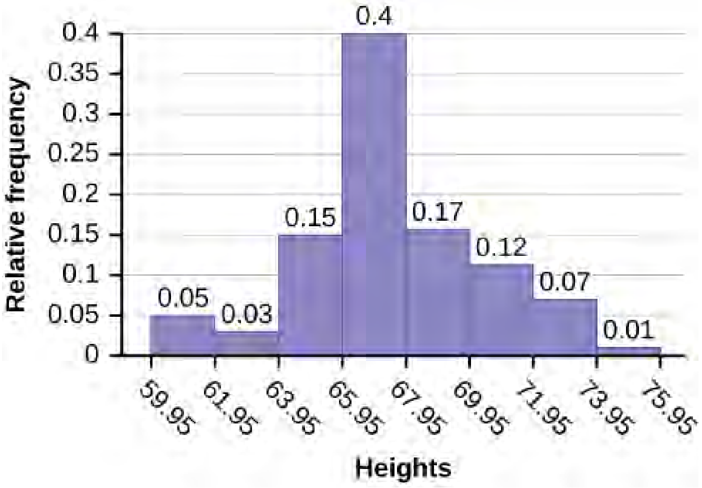


Figure 2.8

Example 2.5

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;
Construct a Histogram.

Books	FREQUENCY (f)
1	11
2	10
3	16
4	6
5	5
6	2

Table 2.9 Ungrouped Frequency distribution of Books

Solution to Example 2.5

BOUNDARIES	FREQUENCY
.5 – 1.5	11
1.5 – 2.5	10
2.5 – 3.5	16
3.5 – 4.5	6
4.5 – 5.5	5
5.5 – 6.5	2

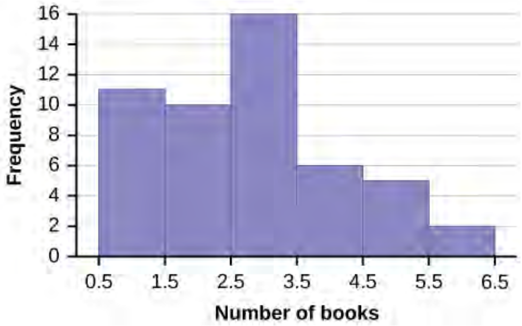


Figure 2.9

Try It Σ

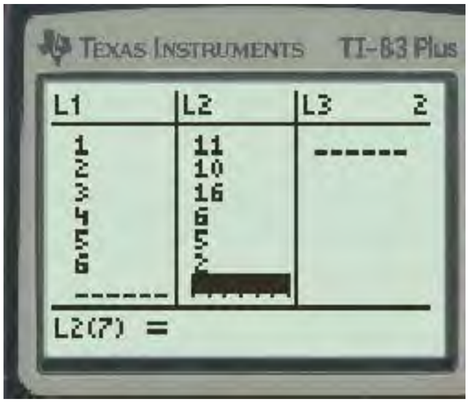
2.2 The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.
9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Using the TI-83, 83+, 84, 84+ Calculator

Go to **Appendix G**.

There are calculator instructions for entering data and for creating a customized histogram.
Create the histogram for Example 2.3.

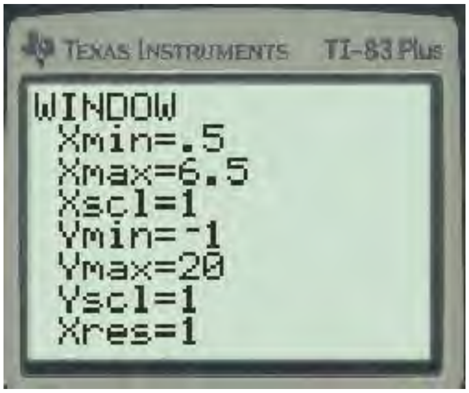
- Press **Y=**. Press CLEAR to delete any equations.
- Press **STAT** 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.



L1	L2	L3 2
1	11	------
2	10	
3	16	
4	6	
5	5	
6	2	

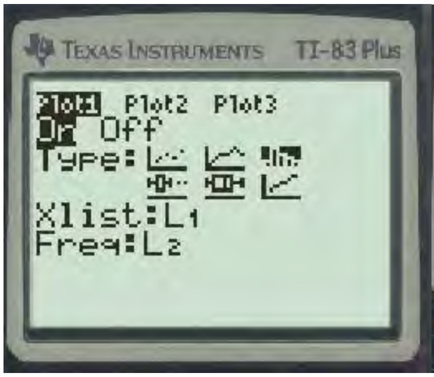
L2(7) =

- Press **WINDOW**. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.

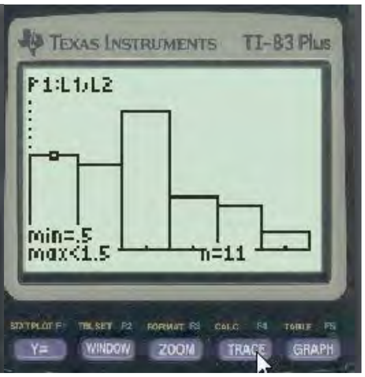


WINDOW
Xmin=.5
Xmax=6.5
Xscl=1
Ymin=-1
Ymax=20
Yscl=1
Xres=1

- Press 2nd Y=. Start by pressing 4: Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.



Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:L2



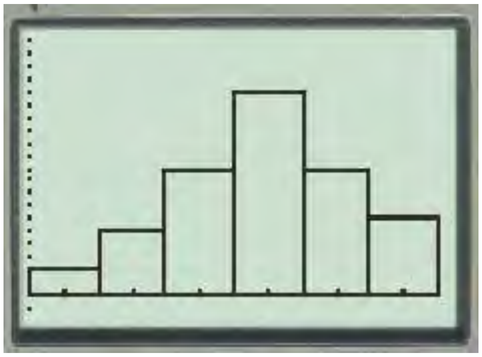
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2). Press Graph

Use TRACE key and the arrow keys to examine the histogram

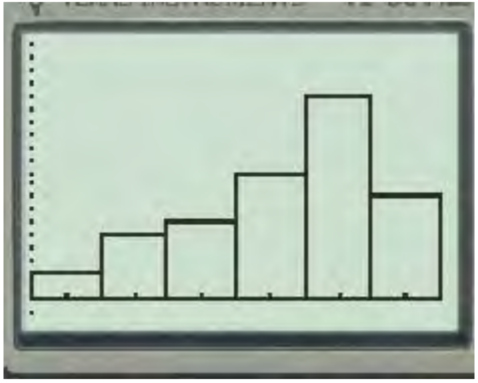
Descriptions of Histograms:

There are four descriptions of histograms. These descriptions, describe how the histogram is shaped.

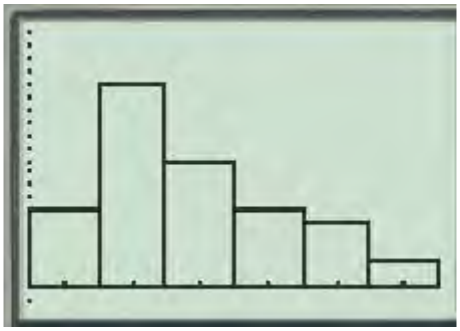
1. **Symmetrical:** the largest bar is in the center interval and the bars on each side of the center decrease about the same rate.



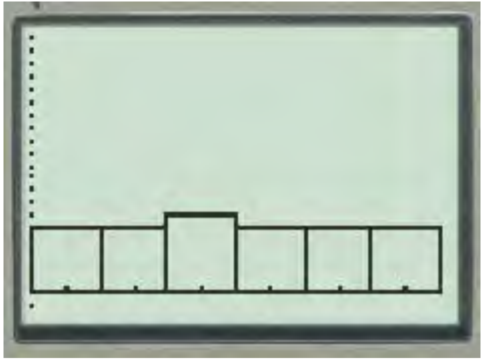
2. **Skewed Left**: the largest bar is closer to the right side and the bars start decreasing toward the left.



3. **Skewed Right:** the largest bar is closer to the left side and the bars start decreasing toward the right.



4. **Uniform**: each bar of the histogram is the same size for every interval



NOTE: Histograms will not always be exactly one of the four descriptions; therefore use adjectives to help describe the histogram. For example, "mostly symmetrical".

Stem-and-Leaf Graph:

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 **stem-and-leaf plot:**

1.) Divide each observation of data into a stem and a leaf. The leaf consists of a final significant digit.

For example:
The number 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three.

2.) Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.6

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows:
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
Create a Stem and leaf plot.

Solution to 2.6:

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

Table 2.9 Stem-and- Leaf Graph

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

Try It Σ

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

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

Construct a stem plot for the data.

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

Example 2.7

The data are the distances (in km) from a home to local supermarkets. Create a stemplot using the data:
1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3
Do the data seem to have any concentration of values? The leaves are to the right of the decimal.

Solution 2.7

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

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

Table 2.10

Try It Σ

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

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

Example 2.8

A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Table 2.11 and Table 2.12 show the ages of presidents at their inauguration and at their death. Construct a side- by-side stem-and-leaf plot using this data.

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J.Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56

Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G.W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Table 2.11 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J.Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Carter	
Tyler	71	McKinley	58	Reagan	93
Polk	53	T. Roosevelt	60	G.H.W Bush	
Taylor	65	Taft	72	Clinton	
Fillmore	50	Wilson	67	G.W. Bush	
Pierce	64	Harding	57	Obama	
Buchanan	77	Coolidge	60		

Table 2.12 Presidential Ages at Death

Solution to Example 2.8

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

Line Graphs:

Another type of graph that is useful for specific data values is a line graph. In the particular **line graph** shown in Example 2.9 is for an ungrouped frequency distributions, the x-axis (horizontal axis) consists of data values and the y-axis (vertical axis) consists of frequency points. The frequency points are connected using line segments.

Example 2.9

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table 2.13 and in Figure 2.10.

# of times	FREQUENCY (f)
0	2
1	5
2	8
3	14
4	7
5	4

Table 2.13

Solution to 2.9:

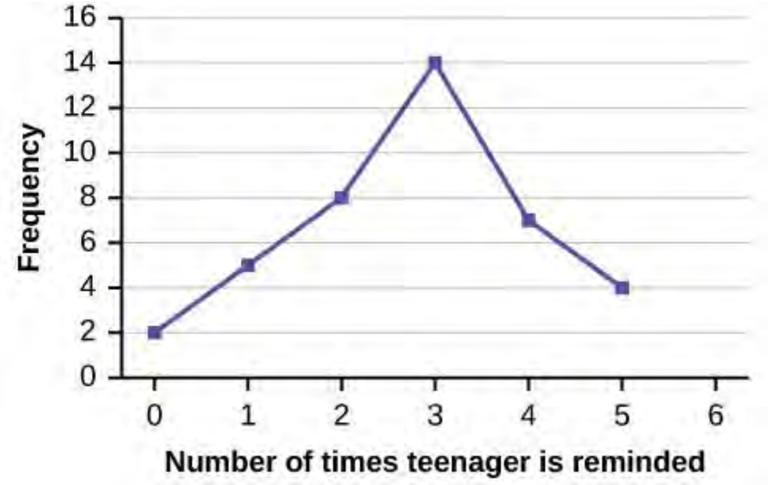


Figure 2.10

Try It Σ

2.5 In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in **Table 2.14**. Construct a line graph.

# of times	FREQUENCY (f)
0	7
1	10
2	14
3	9

Frequency Polygon

Frequency polygons are analogous to line graphs, and just as line graphs make ***continuous*** data visually easy to interpret, so too do frequency polygons.

To construct a frequency (relative frequency) polygon:

1.) Find the midpoints of each class and place them on the x-axis

2.) Place frequency or relative frequency on the y-axis

3.) Draw a line graph corresponding to each midpoint and frequency

Example 2.10

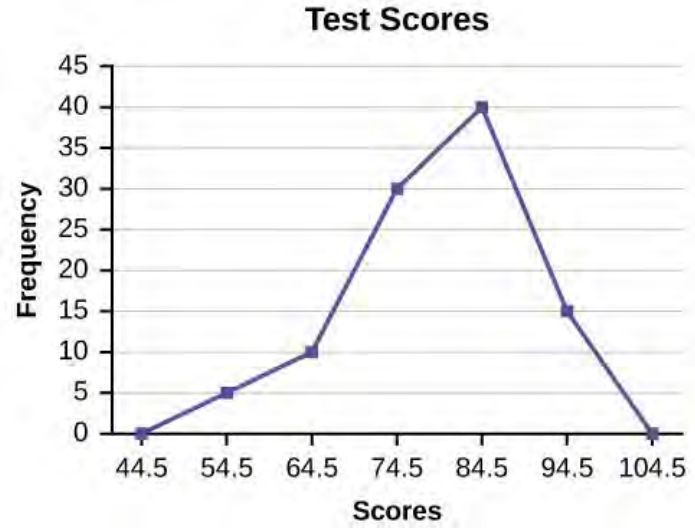
Construct a frequency polygon using the following distribution of Calculus Final Test Scores

CLASSES	FREQUENCY (f)
50 - 59	5
60 - 69	10
70 - 79	30
80 - 89	40
90 - 99	15

Solution to 2.10:

First find the midpoints of each class.

Midpoints: 54.5, 64.6, 74.5, 84.5, 94.5



Try It Σ

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

2.6 Construct a frequency polygon using the ages of President's inauguration in the following table:

Classes (Ages)	Frequency (f)
42 – 46	4
47 – 51	11
52 – 56	14
57 – 61	9
62 – 66	4
67 - 71	2

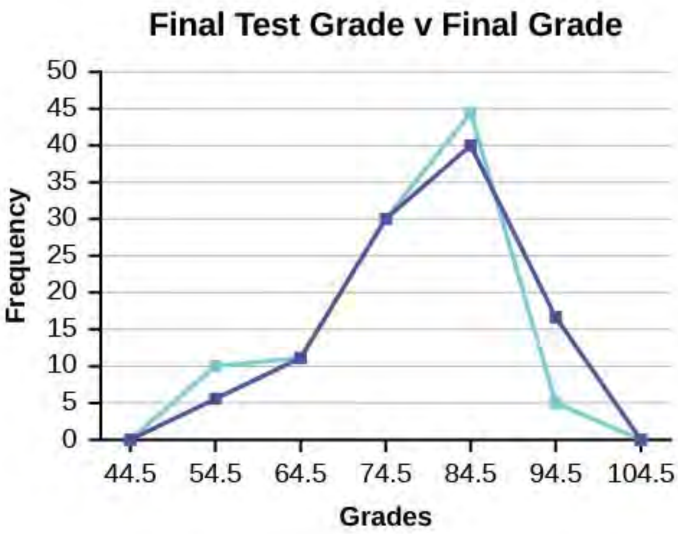
Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 2.11

We will construct an overlay frequency polygon comparing the scores from Example 2.8 with the students' final numeric grade.

CLASSES	FREQUENCY (f)
50 - 59	10
60 - 69	10
70 - 79	30
80 - 89	45
90 - 99	5

Solution to 2.11:



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

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don‘t have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

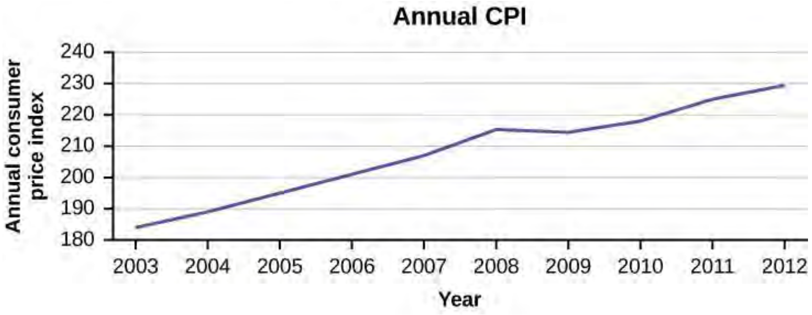
To construct a time series graph, we must look at both pieces of our paired data set. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing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

Example 2.12

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

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Annual	184	188.9	195.3	201.6	207.342	215.303	214.537	218.056	224.939	229.594

Solution to 2.12:



Try It Σ

2.7 The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO2 emissions for the United States.

CO2 Emissions			
	Ukraine	**United Kingdom**	**United States**
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,627
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Uses of a Time Series Graph

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

2.4 | Measures of Central Tendency

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

NOTE

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 symbol 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 used to be truly random.

When the values in a data set are not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. 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

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

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

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 two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49,$ so the median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5,$ so the median occurs midway between the 50th and 51st values.

In general, the values of the mean and 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 2.13

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

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

Calculate the mean and the median.

Solution to 2.13

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); so $M = \frac{24+24}{2} = 24$.

Using the TI-83, 83+, 84, 84+ Calculator

To find the mean and the median:

Go to STAT >> EDIT. Clear list L1 by pressing 4: ClrList, and then pressing 2nd 1 for list L1. Press ENTER. Enter data into the list editor. Go to STAT >> EDIT and enter the data values into list L1.

Press STAT and arrow to CALC. Select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER. At the top of the screen you will see $\bar{x}$ = 23.6. Use the arrow keys to scroll down to the second output screen, and you will see the median: Med = 24.

Try It Σ

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

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

Example 2.14

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

Solution to 2.14

The mean is $\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129,400$, whereas the median is $M = 30,000$.

The median is a better measure of the "center" than the mean. The mean is distorted because of one very large value, 5,000,000. Such a value is called an “outlier”, meaning that it is an extreme value. The 30,000 gives us a better sense of the middle of the data.

Try It Σ

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

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called **bimodal.** There also can be **no mode**. No mode exists when each data value has the same frequency.

Example 2.15

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.

Solution to 2.15

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

Try It Σ

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

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

Find the mode.

Example 2.16

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

NOTE

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

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

Try It Σ

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

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 of the sample is very likely to get closer and closer to μ. Similarly, if we take larger and larger samples, the sample proportion will approach the population proportion p. These facts are discussed in more detail later in the text. Recall that a **statistic** is a number calculated from a sample, whereas a **parameter** is a corresponding number calculated from a population. Examples of statistics include the mean, median, mode and sample proportion.

The Law of Large Numbers tells that $\bar{x}$ is a sample statistic that estimates the parameter μ (the population mean).

Similarly, the sample proportion $\hat{p}$ is a sample statistic that estimates the population proportion, p.

Calculating the Mean from Frequency Tables

Suppose that thirty randomly selected students were asked the number of movies they watched the previous week. The results are shown in the frequency table below.

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

Table 2.24

Suppose that we wanted to know the mean number of movies watched for the sample. Of course, this will be the sum of the data, divided by 30 (the size of the sample):

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

To do this, we add up the data values, multiplied by their respective frequencies and divide by 30:

$$\bar{x} = \frac{0(5)+1(15)+2(6)+3(3)+4(1)}{30} \approx 1.33$$

Note that this could also be written as:

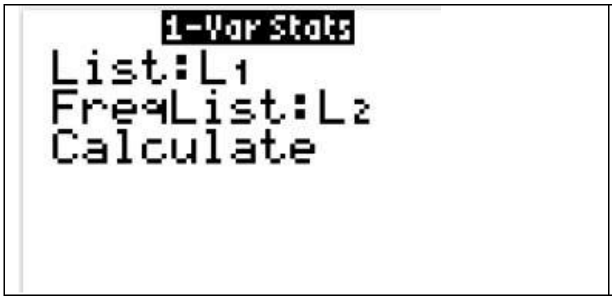
$$\bar{x} = \frac{0(5)+1(15)+2(6)+3(3)+4(1)}{30} = \frac{0\cdot 5}{30}+\frac{1\cdot 15}{30}+\frac{2\cdot 6}{30}+\frac{3\cdot 3}{30}+\frac{4\cdot 1}{30}$$
$$= 0\cdot\frac{1}{30}+1\cdot\frac{15}{30}+2\cdot\frac{6}{30}+3\cdot\frac{3}{30}+4\cdot\frac{1}{30} \approx 1.33$$

In other words, we can also calculate the mean by multiplying the data values by their respective relative frequencies and adding the resulting values. And this can be easily done using the TI-84 family of calculators:

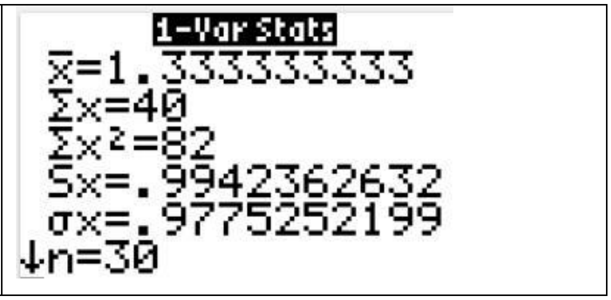
Using the TI-83, 83+, 84, 84+ Calculator

To calculate the mean of a data set from a frequency table

Go to STAT >> EDIT and clear lists L1 and L2.
Enter the data values into L1, and enter the frequencies into L2. Then go to STAT >> CALC. Select 1-VarStats, type L1, L2 and hit ENTER. The mean will appear at the top of the screen as $\bar{x}$.



1-Var Stats
List:L_1
FreqList:L_2
Calculate



1-Var Stats
$\bar{x}=1.333333333$
$\Sigma x=40$
$\Sigma x^2=82$
$Sx=.9942362632$
$\sigma x=.9775252199$
$\downarrow n=30$

When only grouped data is available, we do not know the individual data values (we only know intervals and interval frequencies); therefore, we cannot compute an exact mean for the data set. However, we can still estimate the sample mean by from the frequency table. To calculate the mean from a grouped frequency table we can apply the same method as above; but since we do not know the individual data values we instead use the midpoint of each interval. The midpoint is the average of the lower boundary and upper boundary:

$$\text{midpoint} = \frac{\text{upper boundary} + \text{lower boundary}}{2}.$$

Example 2.17

A frequency table displaying professor Blount's last statistic test is shown below:

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

Find the best estimate of the class mean.

Solution 2.15 First, find the midpoints for all intervals:

Grade Interval	Midpoint	Number of Students
50.5-56.5	53.5	1
56.5-62.5	59.5	0
62.5-68.5	65.5	4
68.5-74.5	71.5	4
74.5-80.5	77.5	2
80.5-86.5	83.5	3
86.5-92.5	89.5	4
92.5-98.5	95.5	1

Go to STAT >> EDIT, clear lists L1, L2. Enter the midpoints into L1 and the frequencies into L2.
Go to STAT >> CALC, select 1-VarStats and type L1, L2 (don't forget the comma for the TI-83!).
Hit ENTER, and the mean will appear at the top of the output screen: $\bar{x} = 76.86$.
For TI-84 (plus) type L1 into FREQ: and L2 into FREQLIST:, then CALCULATE

Try It Σ

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

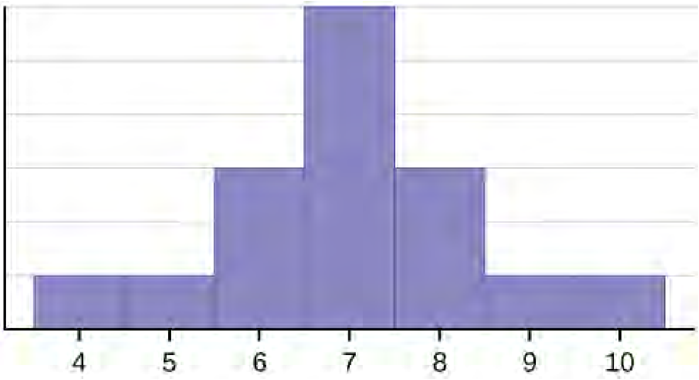
Hours Spent on Video Games	Number of teenagers
0 – 3.5	3
3.5 – 7.5	7
7.5 – 11.5	12
11.5 – 15.5	7
15.5 – 19.5	9

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

Skewness and the Mean, Median, and Mode

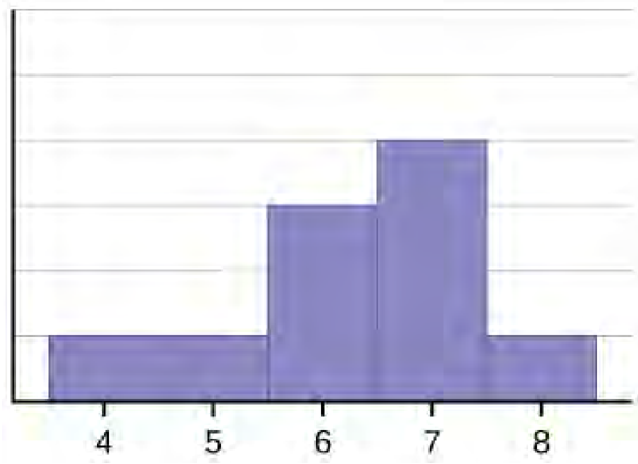
Consider the following data set: 4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

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



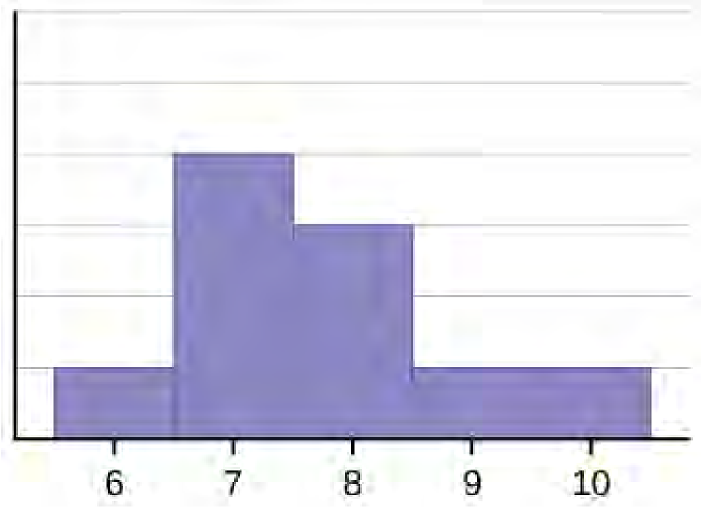
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven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

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



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

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



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

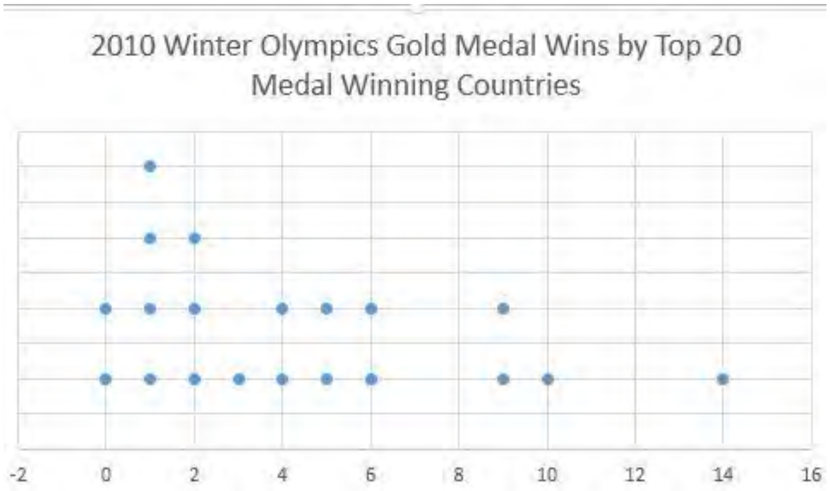
To summarize:

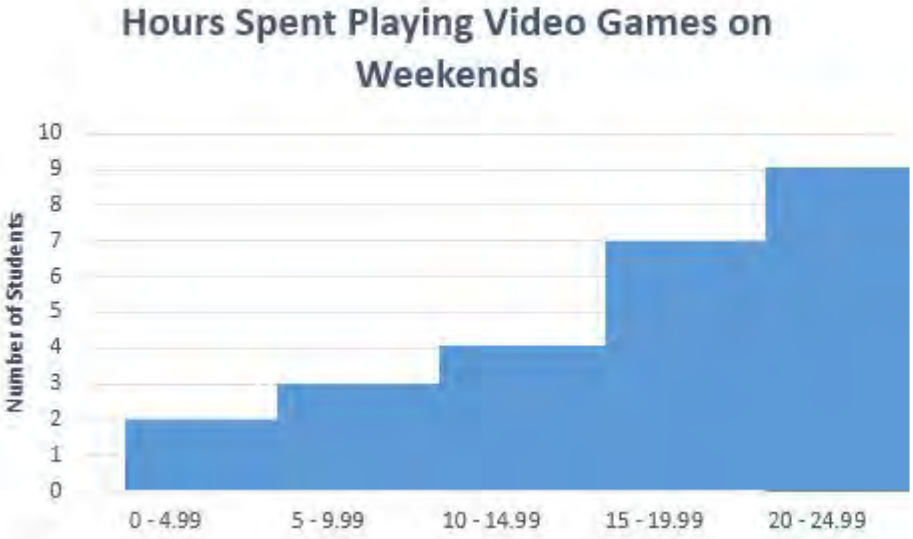
- If the distribution of data is skewed to the left, the mean is less than the median.
- If the data is symmetrical, then the median and mean are equal
- If the data is skewed to the right, then the mean is greater than the median.

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

Try It Σ

2.16 Discuss the mean, median, and mode for the following two graphs (dot plot and histogram).





2.5 | Measures of Variation

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. In this section we will discuss three measures of variation: the range, the standard deviation and the variance.

Standard deviation.

The most commonly used measure of variation, or spread, is the standard deviation. The **standard deviation** is a non-negative number that measures how far data values are from their mean. Its importance will become much clearer when we begin studying methods of Inferential Statistics. For now, we point out two key features of this important statistic.

The standard deviation provides a measure of the overall variation in a data set. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes. Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average, and hence more predictable.

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

Rosa's wait time of seven minutes is two minutes longer than the average of five minutes; that is, her wait time is **one standard deviation above the average** of five minutes.

Binh's wait time is four minutes less than the average of five minute; that is, his wait time of one minute is **two standard deviations below the average** of five minutes. A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be "unusual", or "far" from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (As we will see in the next section)

As with measures of center, we can have a standard deviation measured from sample data (a statistic) or from population data (a parameter). The lower case letter *s* represents a sample standard deviation and the Greek letter σ (sigma, lower case) represents a population standard deviation. This is similar to our notation protocol for the mean, where we used the symbol $\bar{x}$ for the sample mean and the Greek letter μ for the population mean.

Calculating the Standard Deviation

If x is a data value, then the difference between x and the mean is called its **deviation**. In a data set, there are as many deviations as there are values in the data set. The deviations are used to calculate the standard deviation; in fact, we intuitively think of the standard deviation as a sort of "average" of the deviations. If the data belong to a population, in symbols a deviation is $x - \mu$. For sample data, a deviation will be $x - \bar{x}$.

The formula for calculating the standard deviation will also be slightly different for sample data than for population data. If the sample is reasonably large and representative of the population, then s should be a good estimate of σ.

Suppose that we wanted to measure the totality of all variation in a data set. Then we might start by adding *all* of the deviations $x - \bar{x}$. However, in any data set, there are values both above and below the mean $\bar{x}$. So some deviations would be positive, and some will be negative, and there will be considerable cancellation. In fact, the sum of all the deviations would always be zero. To prevent deviations from cancelling each other out we look at the *squares* of the deviations: terms of the form

$(x - \bar{x})^2$.

The **variance is the average of the squares of the deviations.** For reasons that will become clear later, we compute these averages slightly differently for population data than we do for sample data.

The population variance is: $\sigma^2 = \sum \frac{(x-\mu)^2}{N}$, where N = the number of data values in the population.

The sample variance is: $s^2 = \sum \frac{(x-\bar{x})^2}{n-1}$, where n = the number of data values in the sample.

The symbol σ^2 represents the population variance - so **the population standard deviation σ is the square root of the population variance**. Similarly, the symbol s^2 represents the sample variance, so **the sample standard deviation *s* is the square root of the sample variance**. Taking the square root of the variance can be thought of as reversing the effect of squaring the deviations. Thus, we can think of the standard deviation as a sort of "average" of the deviations. Note also that the standard deviation will always be measured in the same units as the data values.

Formulas for the Standard Deviation:

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

Population Standard Deviation

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

Sample Standard Deviation

NOTE

☞ In practice, we will always use a calculator or computer software to calculate a standard deviation. If you are using a TI-83, 83+, 84+ calculator (or Excel) 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 it is instructive to do one step-by-step example to help you understand how the standard deviation measures variation from the mean. The calculator instructions appear at the end of this example.

Example 2.18

In a fifth grade class, 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;

The average age is 10.53 years, rounded to two places. Find the sample standard deviation, rounded to two decimal places.

Solution to 2.18

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

Data	Freq.	Deviation	Deviation^2	Freq.* Deviation^2
x	*f*	$(x - \bar{x})$	$(x - \bar{x})^2$	$f * (x - \bar{x})^2$
9	1	9 – 10.525 = –1.525	$(–1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one, which is 19.

$$s^2 = \frac{9.7375}{19} = 0.5125$$

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

$$s = \sqrt{0.5125} = 0.715891$$

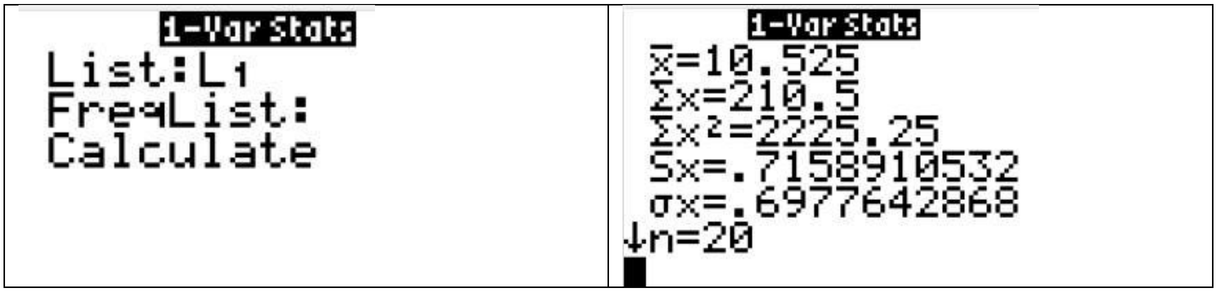
Rounded to two decimal places, ***s* = 0.72**.

Again, we almost never use the formulas or the procedure above to calculate a standard deviation. Instead, we will use either software or a calculator to calculate these.

Using the TI-83, 83+, 84, 84+ Calculator

To find the standard deviation:

Go to STAT >> EDIT. Clear list L1, and enter the data into the list.
Go to STAT >> CALC and select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.



```
1-Var Stats
List:L1
FreqList:
Calculate
```

```
1-Var Stats
x̄=10.525
Σx=210.5
Σx²=2225.25
Sx=.7158910532
σx=.6977642868
↓n=20
```

In the output screen you will see both Sx and σx:

Sx represents the ***sample*** **standard deviation, s**; so if the data comes from a sample, use Sx.

$\boldsymbol{\sigma_x}$ represents the ***population*** **standard deviation, σ**; so if the data comes from a population, use σ_x.

Try It Σ

2.17 On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use a calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

NOTE

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**, either *s* or σ, is non-negative; i.e. it is either zero or larger than zero. When the standard deviation is zero, there is no spread at all; 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 very large, the data values are quite spread out about the mean and conversely; so extreme values can make *s* or σ very large. For this reason the standard deviation is more useful for distributions that are reasonably symmetric. In the next section we will introduce a different statistic to measure variation that works better for skewed distributions. Thus it is often a good idea to graph the data before deciding which statistics to use.

Try It Σ

2.18 The following data show the number of different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12; Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Standard deviation of Grouped Frequency Tables

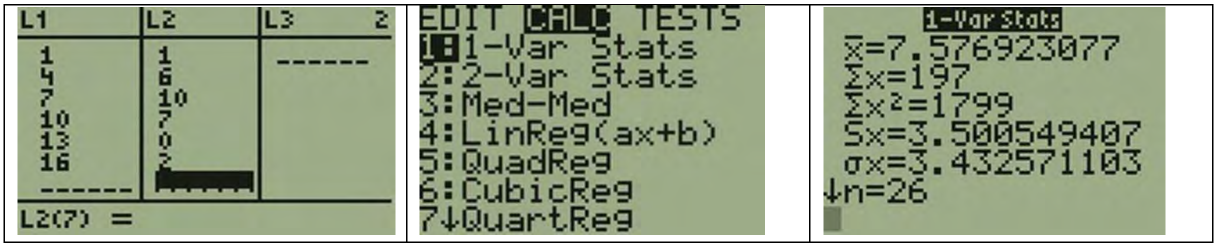
Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. Just as we could not find the exact mean from a frequency table, neither can we find the exact standard deviation. However, by again replacing the class intervals by their midpoints, we can estimate the standard deviation using the same procedure as we used for estimating the mean.

Example 2.19

Find the standard deviation for the data in the following table:

Class	Frequency
0-2	1
3-5	6
6-8	10
9-11	7
12-14	0
15-17	2

Solution: Go to STAT >> EDIT, and clear lists L1 and L2.
Enter the midpoints of the classes into L1; enter the frequencies into L2:
Next, go to STAT >> CALC and select **1-VarStats:**
Finally, type L1, L2 and hit Enter to get:



```
L1      L2      L3     2
1       1       ------
4       6
7       10
10      7
13      0
16      2
------  
L2(7) =
```

```
1-Var Stats
x̄=7.576923077
Σx=197
Σx²=1799
Sx=3.500549407
σx=3.432571103
↓n=26
```

Here we see both a population standard deviation, σ_x, and the sample standard deviation Sx displayed.

The following rules provide a little more insight into what the standard deviation tells us about the distribution of the data.

Chebyshev's Rule: 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 4.5 standard deviations of the mean.

Empirical Rule: For data having a distribution that is **bell-shaped** and **symmetric**:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- Approximately 99.7% of the data is within three standard deviations of the mean.
- It is important to note that this rule applies *only* when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying "Normal" distribution in later chapters.

Finally, we have seen how the mean and standard deviation can be affected by extreme high or extreme low values. An **outlier** is a data value that is far removed from the rest of the data. Outliers need to be examined carefully - sometimes they are the result of measurement error, and should be removed from the data and not used in the analysis at all. Other times, they hold valuable information about the population under study and should be included in the data. Now that we have introduced the standard deviation, we can present a commonly used criterion for classifying a data value as an outlier:

A data value that is more than three standard deviations from the mean is called an outlier.

Example 2.20

Using Chebychev's theorem and Emperical find the percentage of values that lie within 3 standard deviations from the mean; given the heights of 20 year old males have a symmetrical distribution with a mean of 70.2 inches and a standard deviation of 1.5 inches.

Solution to 2.20:

We are given $\mu = 70.2$ and $\sigma = 1.5$. Three standard deviations is represented by $3\cdot\sigma = 3(1.5) = 4.5$. Subtract and add 4.5 from the mean, 70.2.

$$70.2 - 4.5 = 65.7$$

$$70.2 + 4.5 = 74.7$$

Chebychev's theorem states that 89% of 20 year old male heights lie within 65.7 inches and 74.7 inches. While, Emperical Rule states that 99.7% of 20 year old male heights lie within 65.7 inches and 74.7 inches. Chebychev's theorem is more conservative since it is used for any type of distribution. Emperical Rule is specifically for symmetrical distributions.

2.6 | Measures of Relative Position

Measures of relative position are used to describe data values relative to other data values. Here we will discuss commonly used measures of relative position – in particular, *z*-scores, percentiles, and quartiles.

z-Scores

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading. But we can tell which data value is farther removed from the center by finding how many standard deviations away from its mean the value is. And we have a statistic called a "*z*-score" for this purpose. Let x be a data value, μ be the mean, and σ the standard deviation. Then the z-score corresponding to x is:

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

Solving for x in this equation, we get $x = \mu + z\sigma$. So this really does measure how many standard deviations x is from the mean. Moreover, if $z > 0$, then x is greater than the mean; and if $z < 0$, then x is less than the mean.

Example 2.21

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school.

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

Which student had the highest GPA when compared to his school?

Solution to 2.21:

For each student, we use the z-score to determine how many standard deviations his GPA is away from the average for his school:

$$\text{For John, } z = \frac{x-\mu}{\sigma} = \frac{2.85 - 3.0}{0.7} = -0.21.$$

$$\text{For Ali, } z = \frac{x-\mu}{\sigma} = \frac{77 - 80}{10} = -0.3.$$

John's GPA is 0.21 standard deviations *below* his school's mean, while Ali's GPA is 0.3 standard deviations *below* his school's mean. So John's *z*-score is higher than Ali's. For GPA, higher values are better; so John has the better GPA relative to his school.

Try It Σ

2.19 Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. The table below shows there times.

Swimmer	Time in Seconds	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Which swimmer had the fastest time when compared to her team?

Percentiles

We say that a data value *x* is at the *P*-th percentile if *P*% of all data values are less than *x*. For example, if a student taking a standardized test scores at the 85th percentile, this means that 85% of all students taking the test scored lower than the student, and so 15% of students taking the test scored better.

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

Percentiles are most often 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. We will learn more about how to calculate percentiles for large data sets in a later chapter. But sometimes we also need to calculate a percentile for small data sets. The best way to do this is using a cumulative relative frequency table.

Example 2.22

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

Hours of Sleep	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

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

b. **Find the median**. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile. Since 50% of 50 is 25, there are 25 data values less than the median. They include the two 4s, the five 5's, the seven 6's, and eleven of the 7's. The median is between the 25th and and 26th data values, both of which are 7. The median is ***M* = 7.**

c. **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 4's, 5's, 6's and 7's, you have 52% of the data. When you include all the 8's, you will have 80% of the data. **The 75th percentile must be 8**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is again an 8.

A Formula for finding the *k*th Percentile

In addition to using a frequency distribution, there are several formulas for calculating the *k*th percentile. Here is one of them.

Suppose that we want the *k*th percentile. It may or may not be part of the data. Order the data, and let *i* be the index of the *k*th percentile; i.e. this is the position that the percentile appears in the list of data. Let *n* be the total number of data values. Then we calculate $i = \frac{k}{100}(n+1)$

If *i* is a positive integer, then the *k*th percentile is the data value in the *i*th position in the ordered set of data. If *i* is not a positive integer, then round *i* up and round *i* down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.23

Listed below are the ages for 29 Academy Award winning best actors:

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

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution to 2.23

a) $k = 70$, i = the index, $n = 29$. Calculate $i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$.

This is a whole number, so we want the data value in the 21st position in the ordered data.
The 70th percentile is **64**.

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

This is *not* a whole number, so we round down to get 24, and round up to get 25.
The age in the 24th position is 71, and age in the 25th position is 72. Average these two numbers to obtain the 83rd percentile, **71.5.**

Try It Σ

2.21 Listed below are the ages for 29 Academy Award winning best actors:

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

Calculate the 20th percentile and the 55th percentile.

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

Again, order the data from smallest to largest.
Let x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
Let y = the number of data values equal to the data value for which you want to find the percentile.
Let n = the total number of data values.

Calculate $\dfrac{x+0.5y}{n}\cdot 100$ and round to the nearest integer.

Example 2.24

Listed below are the ages for 29 Academy Award winning best actors:

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

a) Find percentile for 58
b) Find percentile for 25

Solution to 2.24

a) Counting from the bottom of the list, there are 18 data values less than 58; and there is one data value equal to 58. So $x = 18$ and $y = 1$.

$$=> \frac{x+0.5y}{n}\cdot 100 = \frac{18+0.5(1)}{29}\cdot 100 = 63.8$$

Thus an age of 58 would be at the **64th percentile**.

b) Counting from the bottom of the list, there are 3 data values less than 25; and there is one data value equal to 25. So $x = 3$ and $y = 1$.

$$=> \frac{x+0.5y}{n}\cdot 100 = \frac{3+0.5(1)}{29}\cdot 100 = 12.07$$

Thus an age of 25 would be at the **12th percentile**.

Quartiles

Quartiles are special cases of percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median M is both the second quartile and the 50th percentile. Thus 25% of the data values in any data set are less than Q_1. Similarly, 75% of the data values in the data set are less Q_3, and so 25% of data values are *above* Q_3. Finally, 50% of all data values are between Q_1 and Q_3.

The quartiles separate the data into quarters. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the following data set:

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

- There are 14 data values, so the median is the average of the 7th and 8th values in the list. Thus, we have $M = (7.2 + 6.8)/2 = 7$.
- The first quartile will be the median of the lower half of the data: 1, 1, 2, 2, 4, 6, 6.8. The middle value of these seven data values is $Q_1 = 2$.
- The third quartile will be the median of the lower half of the data: 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of these seven data values is $Q_3 = 9$.

The **interquartile range** is the difference between the third quartile (Q_3) and the first quartile (Q_1):

$$IQR = Q_3 - Q_1$$

The *IQR* provides a numerical measure of the overall amount of variation in a data set; it is often used in conjunction with the median. I.e. when the median is used to describe the center, the IQR is often used to describe the spread.

The *IQR* can also be used to determine potential **outliers**. Recall that an outlier is a data value that is significantly separated from the other data. Using the *IQR*, we have the following criteria for classifying a data value as an outlier:

A data value x is considered an outlier if either $x < Q_1 - 1.5(IQR)$ or $x > Q_3 + 1.5(IQR)$.

Example 2.25

The following data represent selling prices for homes in a certain city. Calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

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

Solution to 2.25 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

There are 13 data values, so the median is the 7th value: $M = 488{,}800$

Q_1 will be the median of the lower six values: $Q_1 = (230{,}500 + 387{,}000)/2 = 308{,}750$.
Q_3 will be the median of the upper six values: $Q_3 = (639{,}000 + 659{,}000)/2 = 649{,}000$.

Thus, the inter-quartile range is $IQR = 649{,}000 - 308{,}750 = 340{,}250$

=> $Q1 - (1.5)(IQR) = 308{,}750 - (1.5)(340{,}250) = -201{,}625$

=> $Q3 + (1.5)(IQR) = 649{,}000 + (1.5)(340{,}250) = 1{,}159{,}375$

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

Try It Σ

2.23 Two classes take a test. The test Scores for Class *A* are:
69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

The test Scores for Class *B* are:
90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Find the interquartile range for the following two data sets and compare them. Which class had more variable test scores?

Five Number Summary and Boxplots

The quartiles are part of the **five number summary** - these five numbers consist of:

The minimum value
The first quartile Q_1
The median M
The third quartile Q_3
The maximum value

For example, for the real estate data in Example 2.24, the five number summary would be:

Min = 114,950
$Q_1 = 308{,}750$
$M = 488{,}800$
$Q_3 = 649{,}000$
Max = 5,500,000

For a small example like this, the quartiles can be easily calculated by hand; but for a large data set this can be tedious and time-consuming. But most statistical software packages and graphing calculators will calculate the five number summary for us. When entering the data into the calculator or software packages, the data does *not* have to be in order.

Using the TI-83, 83+, 84, 84+ Calculator

To find the Five Number Summary:

Go to STAT >> EDIT. Clear list L1, and enter the data into the list.
Go to STAT >> CALC and select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
Use the down arrow key to scroll down; the 5-number summary appears in the 2nd output screen.

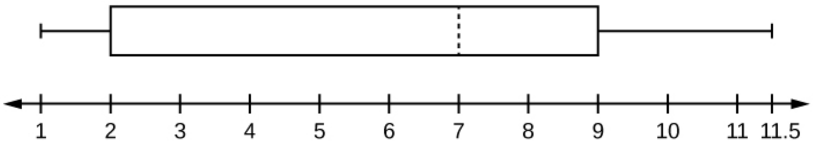
A **box plot** is a graphical representation of the five number summary. These graphs (also called **box and whisker plots** or **box-whisker plots**) allow us to see where provide an image of the concentration of the data, and also show how far the extreme values are from most of the data. So the box plot gives a good, quick picture of the data.

A box plot is constructed from the five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. To construct a box plot, use a horizontal number line and mark the five numbers along the axis. We make small vertical marks above the minimum and maximum values, and larger vertical marks above the quartiles Q_1, M, and Q_3. Finally, we "box off" the marks above the quartiles.
So the first quartile marks one end of the box and the third quartile marks the other end of the box. And the **middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values.

As an example, consider again the dataset: 1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5.
We calculated the quartiles already, and know that the five number summary is:

Min = 1; $Q_1 = 2$; $M = 7$; $Q_3 = 9$; Max = 11.5

Using this information, the boxplot is:



NOTES

1. It is important to start a box plot with a **scaled number line**. Otherwise the box plot may distort the spread of the data.

2. 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.

Example 2.26

The following data are the heights (in inches) 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

a) Find the five number summary for the data and construct a boxplot.
b) Find the *IQR*
c) Given an interval that has the middle 50% of the data.

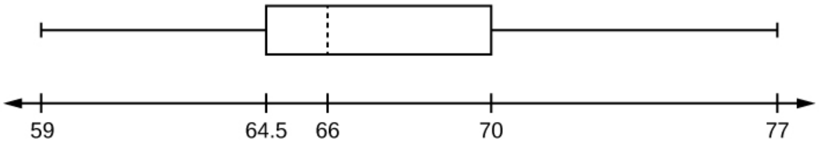
Solution to 2.26:

a.) To find the five number summary, we use the TI-84 calculator.

Go to STAT >> EDIT, and enter the data into L1. Then go to STAT >> CALC and select 1-VarStats. Press 2nd and 1 to select L1, and hit ENTER. Scroll down to the second output screen to get:

Min = 59, Q_1 = 64.5, Q_2: = M = 66, Q_3 = 70, Max = 77

Using these values, we get the following graph:



b) The interquartile range is $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$ inches.

c) About 50% of the data will be between Q_1 and Q_3. That is, about 50% of the students in the class had heights that were between 64.5 inches and 70 inches.

Finally, note that the calculator can also be used to construct the box plot.

Using the TI-83, 83+, 84, 84+ Calculator
To construct a boxplot:

Go to the STAT PLOT menu by pressing 2nd and then the Y = button.
Turn on Plot 1 (Plot 2, Plot 3 should be turned off).
Use the down arrow key to move to Type, and select the boxplot option (middle of second row).
For Xlist, type 2nd and then 1 to select L1.
Press Zoom and then press 9 to select the ZoomStat graphing option; this will automatically configure the window so that all of the data are included.

Finally, if we press TRACE, and use the arrow keys to move from left to right, we can see the quartiles.

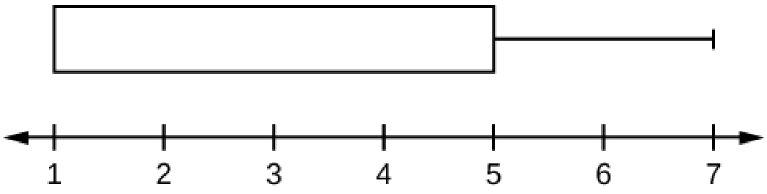
Try It Σ

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

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

It is possible that in a data set, two or more numbers in the five number summary are equal. 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.

As an example, suppose we have data set in which the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the maximum value is 7; then the box plot would look like:



In this case, at least 25% of the values are equal to 1. Twenty-five percent of the values are between 1 and 5, inclusive. At least 25% of the values are equal to 5, and the top 25% of the values fall between 5 and 7, inclusive.

Example 2.27

Test scores for a college statistics class held during the day are:
99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

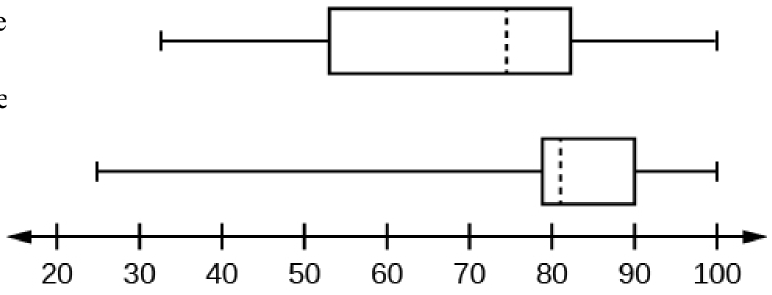
Test scores for a college statistics class held during the evening are:
98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the five number summary for the day class.
b. Find the five number summary for the night class.
c. Create a box plot for each set of data, using the same number line for both box plots.
d. Which box plot has the widest spread for the middle 50% of the data?
What does this mean for that set of data in comparison to the other set of data?

Solution to 2.27

a. Min = 32, Q1 = 56, *M* = 74.5, *Q*3 = 82.5, Max = 99
b. Min = 25.5, Q1 = 78, M = 81, Q3 = 89, Max = 98

c. The upper graph is for the day class.
The lower graph is for the night class.



d. The first data set has the wider spread for the middle 50% of the data; so the *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

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. 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 will depend on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies. Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINES

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

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

Example 2.28

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

Solution to 2.28:

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

Try It Σ

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

Example 2.29

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.

Solution to 2.29:

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It Σ

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

Try It Σ

2.28 During a college basketball season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

KEY TERMS

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

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

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

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

Frequency the number of times a value of the data occurs

Histogram a graphical representation in of the distribution of data in a data set; the horizontal axis represents the data and the vertical axis represents the frequencies or relative frequencies. The graph consists of contiguous rectangles.

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

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

Mean a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample is

$$\bar{x} = \text{sample mean} = \frac{\text{sum of data}}{\text{number of data values}} = \frac{\sum x}{n}$$

When the data is from a population, we denote the mean by the Greek letter μ.

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

Midpoint the number that is halfway between the endpoints of an interval

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

Outlier an observation that falls far outside the rest of the data.

Paired Data Set two data sets that have a one to one relationship so that:

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

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

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

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

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

Standard Deviation: A numerical measure of dispersion in a data set. Specifically, this is 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: The mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x-\overline{x}$ where x is a data value 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 one.

FORMULA REVIEW

$$\text{mean} = \frac{\text{sum of data}}{\text{number of data values}} = \frac{\sum x}{n}$$

Formulas for the Standard Deviation:

$$\sigma = \sqrt{\sum \frac{(x-\mu)^2}{N}} \qquad s = \sqrt{\sum \frac{(x-\bar{x})^2}{n-1}}$$

Population Standard Deviation   Sample Standard Deviation

A Formula for Finding the *k*th Percentile

Suppose that we want the *k*th percentile. It may or may not be part of the data.
Order the data, and let *i* be the index of the *k*th percentile; i.e. this is the position that the percentile appears in the list of data. Let *n* be the total number of data values.

Then we calculate $i = \frac{k}{100}(n+1)$

If *i* is a positive integer, then the *k*th percentile is the data value in the *i*th position in the ordered set of data.
If *i* is not a positive integer, then round *i* up and round *i* down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

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

Again, order the data from smallest to largest.

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

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

Let $n =$ the total number of data values.

Calculate $\frac{x+0.5y}{n} \cdot 100$ and round to the nearest integer.

EXERCISES for CHAPTER 2

For exercises #1 – 4, use the following frequency distribution

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

1. Find the boundaries for the second class.

2. Find the relative frequency for the third class.

3. What is the sample size?

4. What is the cumulative frequency for the fourth class?

5. The following data represent the number of potholes on 35 randomly selected1-mile stretches of highway around a particular city. Construct an ungrouped frequency distribution of the data.

1 4 3 1 4 7 5 1 3 6 1 2

1 1 1 7 1 6 2 7 1 6 4 4

1 1 5 3 5 2 3 2 4 1 3

For exercises #6 – 11, use the following frequency distribution

Temperature (°F)	Frequency
32 - 36	1
37 - 41	3
42 - 46	5
47 - 51	11
52 - 56	7
57 – 61	7
62 – 66	1

6. Find the class width.

7. Find the class midpoints.

8. Find the class boundaries.

9. Find the relative frequencies.

10. Find the cumulative frequencies.

11. Find the sample size.

For exercises #12 – 17, use the following frequency distribution

Time (minutes)	Frequency
70 - 78	14
79 - 87	15
88 - 96	16
97 - 105	18
106 - 114	22
115 - 123	7
124 - 132	3

12. Find the class width.

13. Find the class midpoints.

14. Find the class boundaries.

15. Find the relative frequencies.

16. Find the cumulative frequencies.

17. Find the sample size.

18. Use the given minimum and maximum data entries, and the number of classes, to find the class width, the lower class limits, and the upper class limits.

a. minimum = 12, maximum = 68, 6 classes

b. minimum = 28, maximum = 75, 6 classes

c. minimum = 5, maximum = 92, 8 classes

d. minimum = 30, maximum = 83, 8 classes

19. The following data represents the number of miles a full tank of gas allows for a sample of vehicles inspected. Create a grouped frequency distribution with 7 classes.

260	271	236	244	279	296	284	299	288
288	247	256	338	360	341	333	261	266
287	296	313	311	307	307	299	303	277
283	304	305	288	290	288	289	297	299
332	330	309	328	307	328	285	291	295
298	306	315	310	318	318	320	333	321
323	324	327						

20. The data represent the time, in minutes, spent reading a political blog in a day. Construct a frequency distribution using 5 classes. In the table, include the midpoints, relative frequencies, and cumulative frequencies. Which class has the greatest frequency and which has the least frequency?

34	18	10	44	41	47	0	33	49	45
16	10	27	24	44	46	21	25	45	23

21. Construct a frequency distribution for the given data set using 6 classes. In the table, include the midpoints, relative frequencies, and cumulative frequencies. Which class has the greatest frequency and which has the least frequency?

Amount (in dollars) spent on books for a semester

508 488 430 49 278 542 451 366 451 531 196 115

118 362 434 399 148 30 287 236 185 210 178 105

533 236 477 329

22. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The table below shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	**Number of Students**	**Percentage**
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

23. Using the data from Mrs. Ramirez's math class supplied in Exercise 22, construct a pie graph showing the percentages.

24. David County has six high schools. Each school sent students to participate in a county-wide science competition. The table below shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph and pie graph that shows the population percentage of competitors from each school.

High School	**Science Competition Percentage**	**Overall Student Population Percentage**
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

25. Use the data from the David County science competition supplied in 24.
Construct a bar graph that shows population percentage of students at each school.

For exercises #26 - 29, create a stem/leaf plot for the given data.

26. The miles per gallon rating for 30 cars are shown below (lowest to highest).

19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31,
32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

27. The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39,
40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

28. The following data are the prices of different laptops at an electronics store. Round each value to the nearest multiple of ten.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350,
350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

29. The following data are the daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71,
72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

30. In a survey, 40 people were asked how many times they visited a store before making a major purchase. Construct a frequency polygon. The results are shown in the table:

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

31. In a survey, several people were asked how many years it has been since they purchased a mattress. Construct a histogram. The results are shown below.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

32. Several children were asked how many TV shows they watch each day. Construct a relative frequency polygon. The results of the survey are shown in the table below:

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

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

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

34. Refer to the table for problem 33. Construct a histogram for the given data.

35. Construct a frequency polygon and histogram for each of the following frequency tables:

a.

Pulse Rates for Women	Frequency
60 - 69	12
70 - 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

b.

Actual Speed in a 30 mph Zone	Frequency
42 – 45	25
46 – 49	14
50 – 53	7
54 – 57	3
58 – 61	1

c.

Tar (in mg) of Nonfiltered Cigarettes	Frequency
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

36. Construct a relative frequency polygon and relative frequency histogram from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

37. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlaid frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Men	Frequency
49-55	3
56-62	3
63-69	1
70-76	1
77-83	7
84-90	5

Life Expectancy at Birth - Women	Frequency
49-55	3
56-62	3
63-69	1
70-76	3
77-83	8
84-90	2

38. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Yr	1855	1856	1857	1858	1859	1860	1861
Female	49,545	49,582	50,257	50,234	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

39. The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1972.

Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972
Police	260.4	269.8	272	273	272.5	261.3	268.9	296	319.9	341.4	356.6	376.7
Homicides	8.6	8.9	8.5	8.9	13.1	14.6	21.4	28	31.5	37.4	46.3	47.2

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

For problems #40 - 43, find the mean, median, standard deviation and five number summary for the given data set.

40. The following data shows the mileage (in mpg) for a random sample of 15 new cars.

```
30.1    32.8    29.2    32.6    29.0
28.2    31.5    32.8    32.8    30.2
31.5    32.9    31.4    30.6    31.0
```

41. The following data shows the ages (in years) of a random sample of 22 Boeing 747 airplanes.

```
17   5   5   12   8   5   8   16   14   12   22
15   5   8   17   5   4   2   22   17   20   23
```

42. The following data shows the speeds (in mph) for a random sample of 20 cars driving on I-294 at 1 pm on a single day.

```
68   65   50   79   77   60   55   61   78   75
75   67   72   58   70   62   67   72   70   74
```

43. The following data shows the number of cardiograms done each day for a random sample of 20 days
at an outpatient testing center.

```
25   31   20   32   13   14   43   2    57   23
36   32   33   32   44   32   52   44   51   45
```

44. Find the mean and standard deviation for the following frequency tables:

a.

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

b.

Daily Low Temperature	Frequency
49.5 - 59.5	53
59.5 - 69.5	52
69.5 - 79.5	15
79.5 - 89.5	1
89.5 - 99.5	0

c.

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

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

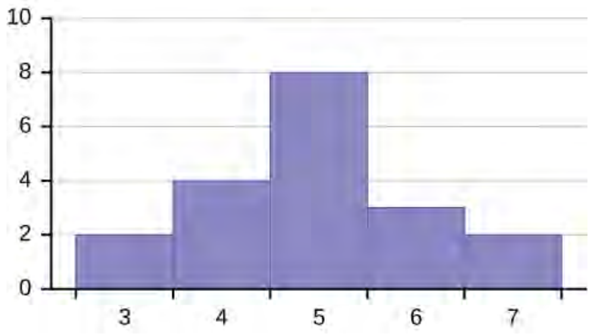
a. sample mean = ______ b) median = ______ c. mode = ______

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

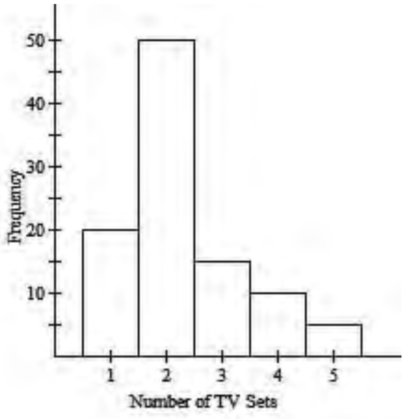
47. When the data are symmetrical, what is the typical relationship between the mean and median?

48. For each of the following histograms,

a) Describe the shape of the distribution
b) Describe the relationship between the mean and median.



(i)



(ii)

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

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

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

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

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

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

53. The following data show the distances in miles between 20 retail stores and a large distribution center:

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

a) Use a calculator to find the standard deviation and round to the nearest tenth.
b) Find the value that is one standard deviation below the mean.

For the next four exercises, calculate the mean, median and standard deviation:

54. The miles per gallon rating for 30 cars are shown:
19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31,
32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

55. The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39,
40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

56. The following data are the prices of different laptops at an electronics store.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350,
350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

57. The following data are the daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71,
72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

58. Fredo and Karl, two baseball players on different teams, wanted to find out who had the higher batting
average when compared to his team.

Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Which baseball player had the higher batting average when compared to his team?

59. The ages for 29 Academy Award winning best actors are shown below, in order from smallest to largest.

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

a. Find the 40th percentile.
b. Find the 78th percentile.
c. Find the percentile for age 37.
d. Find the percentile for age 72.

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

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

62. 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.

63. 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.

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

65. 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?

66. The following data show the lengths of boats moored in a marina. Use this data to construct a boxplot.

16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

67. Find the five number summary and construct a boxplot for the given data set.

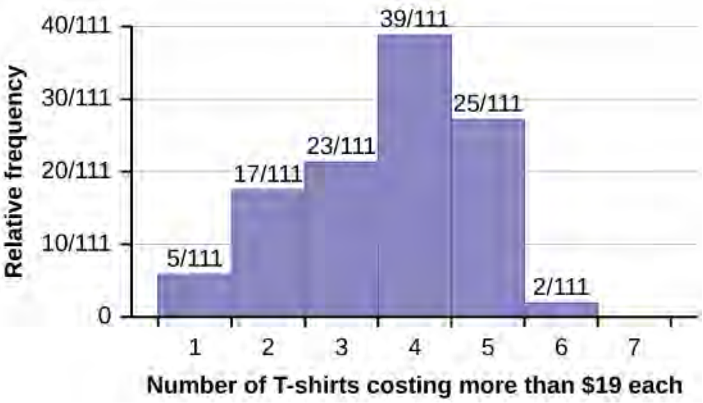
Amount (in dollars) spent on books for a semester

508 488 430 49 278 542 451 366 451 531 196 115

118 362 434 399 148 30 287 236 185 210 178 105

533 236 477 329

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each. The results are summarized in the histogram:

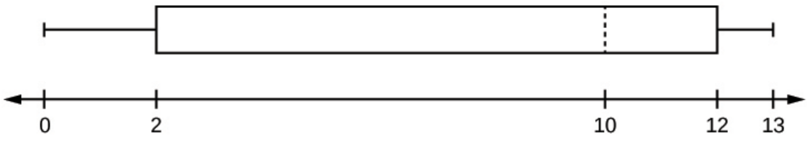


68. The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21% b. 59% c. 41% d. Cannot be determined

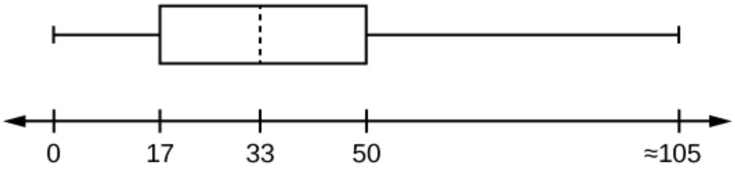
69. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is: a. cluster b. simple random c. stratified d. convenience

70. Given the following box plot:



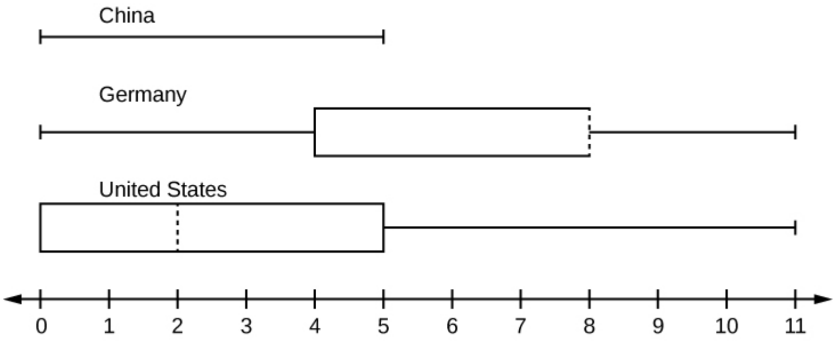
a. Which quarter has the smallest spread of data? What is that spread?
b. Which quarter has the largest spread of data? What is that spread?
c. Find the interquartile range (*IQR*).
d. Are there more data in the interval [5, 10] or in the interval [10, 13]? Explain

71. The following box plot shows ages of the U.S. population for 1990, the latest available year.



a. Are there fewer or more children (age 17 & under) than senior citizens (age 65 & over)?Explain.
b. Given that about 12.6% are age 65 and over, approximately what percentage of the population are working age adults (above age17 to age 65)?

72. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.

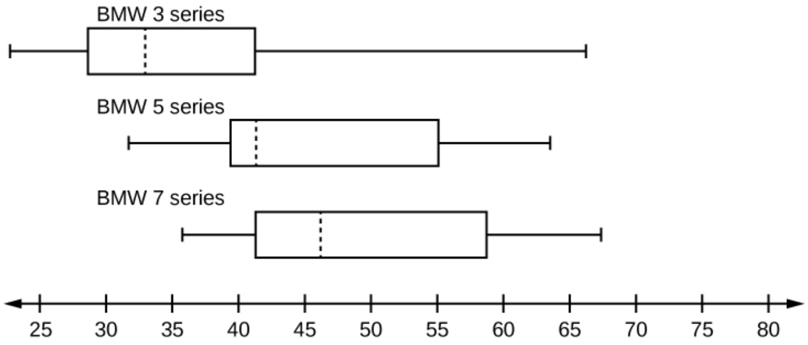


a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.

b. Of the U.S. and Germany, which country has the greater percentage of 20-year olds that have visited more than eight foreign countries?

c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

73. A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In the survey, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.

b. Which group is most likely to have an outlier? Explain how you determined that.

c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?

d. For the BMW 5 series, which quarter has the smallest spread of data? What is the spread?

e. For the BMW 5 series, which quarter has the largest spread of data? What is the spread?

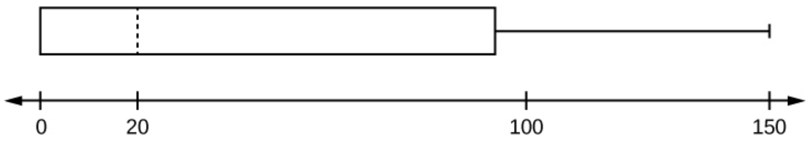
f. Estimate the interquartile range (IQR) for the BMW 5 series.

g. For the BMW 5 series, are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?

h. For the BMW 5 series, which interval has the fewest data in it? How do you know this?

i. 31 to 35   ii. 38 to 41   iii. 41 to 64

74. Given the following box plot:



What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

75. The most obese countries in the world have obesity rates that range from 11.5% to 83.4%. This data is summarized in the following table:

Percent of Population Obese	Number of Countries
11.45–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. What is the standard deviation for the listed obesity rates?
d. How does the United States compare to other countries? Would it be considered "unusual"?

76. The following table gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

a. What is the best estimate for the mean percentage of underweight children?
b. What is the standard deviation?
c. Which interval(s) could be considered unusual? Explain.

77. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

78. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	**GPA**	**School Average GPA**	**School Std. Deviation**
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

79. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain.

80. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

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

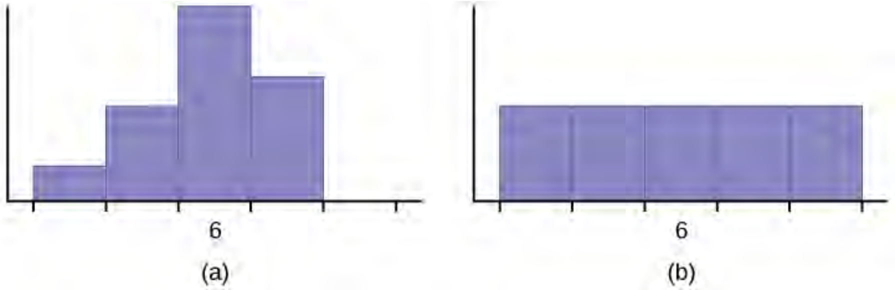
Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

a. What is the *IQR*?
b. What is the mode?

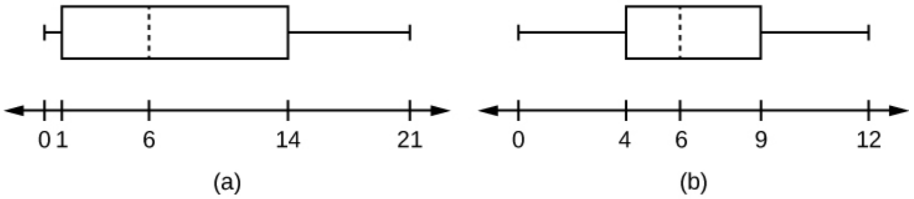
82. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

	Javier	Ercilia
Sample mean $\bar{x}$	6.0 miles	6.0 miles
Sample SD, s	4.0 miles	7.0 miles

a. How can you determine which survey was more accurate?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?



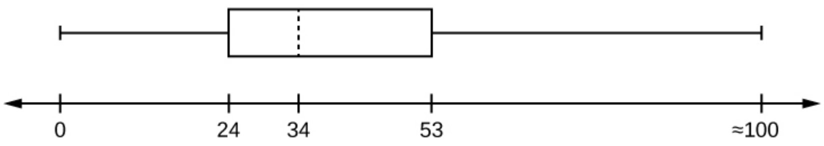
83. Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

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

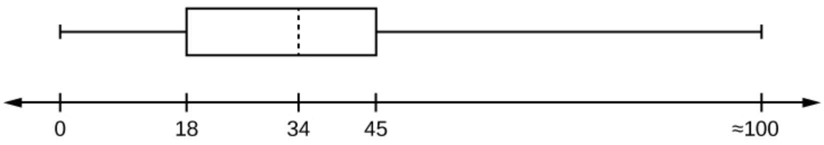
a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?

b. What percentage of the community is under age 35?

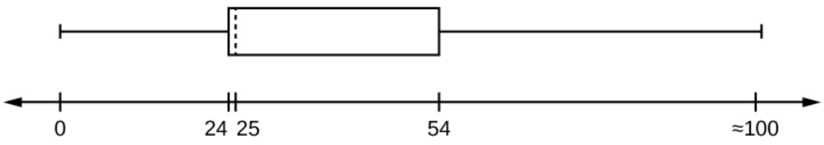
c. Which box plot most resembles the information above?



(a)



(b)



(c)

84. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X = # pairs of sneakers	Frequency	Relative Frequency
1	2	
2	5	
3	8	
4	12	
5	12	
6	0	
7	1	

a. Find the sample mean $\overline{x}$

b. Find the sample standard deviation, *s*.

c. Construct a histogram of the data.

d. Complete the third column of the chart.

e. Find the first quartile.

f. Find the median.

g. Find the third quartile.

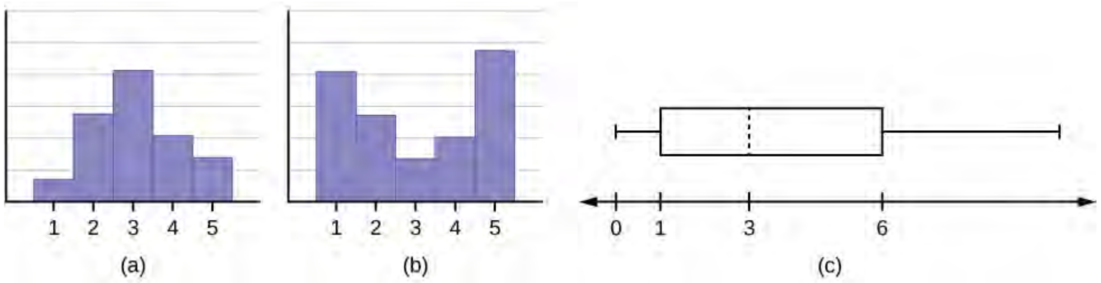
h. Construct a boxplot for the data.

i. What percent of the students owned at least five pairs?

j. Find the 40th percentile.

k. Find the 90th percentile.

85. Refer to the graphs below determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

86. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers
from a given year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212;
184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260;
245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280;
285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302;
265; 290; 276; 228; 265

a. Find the five-number summary for the data.
b. Construct a box plot of the data.
c. The middle 50% of the weights are from ___ to ____.
d. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
e. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
f. Assume the population was this particular year's San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, σ.
   iii. the weight that is two standard deviations below the mean.
   iv. When quarterback Steve Young played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

87. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a relative frequency table.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. The middle 50% of the conferences last from ______days to _______days.
e. Construct a box plot of the data.
f. Find the 10th percentile.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean, median, or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

88. A survey of enrollment at 35 community colleges across the United States yielded the following data:

6414; 1550; 2109; 9350; 21,828; 4300; 5944; 5722; 2825;
2044; 5481; 5200; 5853; 2750; 10,012; 6357; 27,000; 9414;
7681; 3200; 17,500; 9200; 7380; 18,314; 6557; 13,713; 17,768;
7493; 2771; 2861; 1263; 7285; 28,165; 5080; 11,622

a. Make a frequency table for the data using five intervals of equal width.
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

89. The average height of US men in 2010 is 69.3 inches according to Medical News Today. Let's say the standard deviation is 2.8 inches.

a. Using Chebychev's theorem, at least 75% of heights will be between what two heights?

b. Using Empirical rule, 68% of heights will be between what two heights?

c. If a randomly chosen male has a height of 78 inches, what is his z-score?

d. Using Empirical rule, what percentage of heights are above 74.9 inches?

90. The average ACT score in 2010 was 21 with a standard deviation of 5.2 according to Digest of Education Statistics.

a. Using Chebychev's theorem, at least 89% of scores will be between what two scores?

b. Using Empirical rule, 95% of scores will be between what two scores?

c. If a randomly chosen student has a score of 18, what is this student's z-score?

d. Using Empirical rule, what percentage of scores are below 26.2?

91. Let's say that the mean return of mutual funds in the last quarter is 2.8% with a standard deviation of 4.5%. Assume the returns are symmetrically distributed. Find the return value(s) that would separate the

a. top 2.5%

b. middle 99.7%

c. bottom 84%

92. Let's say that the mean return of mutual funds in the last quarter is 2.8% with a standard deviation of 4.5%. Assume the returns are symmetrically distributed.

a. Find the percent of returns that are above 7.3%.

b. Find the percent of returns below 11.8%.

93. Let's say that the mean return of mutual funds in the last quarter is 5.8% with a standard deviation of 2.1%. Assume the returns are skewed right.

a. Find the percent of returns that are between 1.6% and 10%.

b. At least 89% of returns are between what two values?

REFERENCES

2.2 Graphs for Qualitative Data

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics,* 2011. Available online at http://www.kenburbary.com/
2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and- findings/promoting-equity (accessed September 13, 2013).

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

2.3 Graphs for Quantitative Data

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at
http://www.factmonster.com/ipka/A0194030.html
(accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at
http://databank.worldbank.org/data/home.aspx (accessed
April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

2.4 Measures of the Center of the Data

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/
r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

2.5 Measures of Variation

Data from Microsoft Bookshelf.

King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Villines, Z. "What is the average height for men?" Medical News Today. Available online at https://www.medicalnewstoday.com/articles/318155.php (accessed May 20, 2018).

ACT Scores data from https://nces.ed.gov/programs/digest/d10/tables/dt10_155.asp

2.6 Measures of Relative Position

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Data from *West Magazine*.

3 | PROBABILITY TOPICS

Figure 3.1 Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Introduction

Chapter Objectives

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

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

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

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

3.1 | Terminology

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

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. We will discuss three ways to represent a sample space are: list the possible outcomes, create a tree diagram, or using 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. Sample spaces can often be written in multiple ways. For example, if we toss a coin three times in succession, we could describe the sample space in terms of the number of heads as $S = \{0, 1, 2, 3\}$ or we can list the possible sequences of heads and tails as $S = \{\text{HHH, THH, HTH, HHT, HTT, THT, TTH, TTT}\}$. We will see in a moment that the second way is more useful, since it has **equally likely outcomes.** That is, each of the eight outcomes is equally likely to occur. It should be intuitively clear that this is not true of the first set; i.e. there is only one way for all three tosses to land as heads, but there is more than one way to get 1 head in three tosses.

An **event** is any collection 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 as P(*A*).

There are two basic ways to calculate probabilities; Theoretical Probability and Empirical Probability.

Theoretical Probability: Suppose that all outcomes in the sample space are equally likely. To calculate the probability of an event *A*, count the number of outcomes for event *A* and divide by the total number of outcomes in the sample space. That is,

$$\mathrm{P}(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S}.$$

For example, suppose we toss a coin three times in succession. As we saw above, there is a total of 8 outcomes in the sample space, and three of these outcomes have exactly two heads. Thus, P(two heads) = 3/8.

Empirical Probability: The probability of any event is the **long-term relative frequency** of that event. That is, if we were to do a very large number of trials (repetitions of our experiment) then the probability of *A* is:

$$\mathrm{P}(A) = \frac{\text{number of times } A \text{ occurs}}{\text{total number of trials}}$$

For example, suppose that an insurance company wants to find the probability that a suburban male driver, whose age is between 20 and 25 years, will have an accident in the coming year. Then they would use data from thousands of such drivers and calculate the proportion of those drivers that had an accident in the past year.

The **Law of Large Numbers** states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment will become closer and closer to the theoretical probability. For example, suppose you roll one fair six-sided die, with the numbers 1, 2, 3, 4, 5, 6 on its faces. Let *E* be the event that we roll a number that is at least five. There are two outcomes (5, 6) in *E*, so P(*E*) = 1/3. 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. However, if we were to roll the die a very large number of times, the ratio of times *E* occurred would get closer and closer to 1/3.

While a prisoner of war during World War II, the English mathematician John Kerrich (1903–1985) performed a number of probability experiments to demonstrate the Law of Large Numbers. The most famous of these consisted of tossing a coin 10,000 times and recording whether it landed heads or tails. He obtained 5067 heads, so the empirical probability of heads was 5067/10,000 or 50.67%.

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

Using either definition of probability, it is easy to obtain the following three general properties:

Properties of Probability

1) For any event *A*, $\mathbf{0 \leq P(\mathit{A}) \leq 1}$. That is, the probability of any event is always a number between 0 and 1.

2) The probability of any event *A* is equal to the sum of the probabilities of the individual outcomes in *A*.

3) The sum of the probabilities of all outcomes in *S* must equal 1. This is true whether or not *S* consists of equally likely outcomes.

Note that we can write probabilities as a fraction, a decimal or a percent. Moreover, if P(*A*) = 0 this means the event *A* can never happen, whereas P(*A*) = 1 means the event *A* must always occur (*A* is a certainty).

Often we deal with events that are described in terms of other events; in particular, events involving the connectives "and" , "or" and "not". These are sometimes called *compound events*.

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}. Then the event *A* or *B* = {1, 2, 3, 4, 5, 6, 7, 8}.

Notice that 4 and 5 are NOT listed twice.

The event *A* **and** *B* is the set of outcomes that are common to both *A* and *B*. That is, an outcome is in *A* and *B* if it 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* consists of all outcomes that are **not** in *A*. The complement of *A* is denoted as *A′* (read "*A* prime"). Notice that P(*A*) + P(*A′*) = 1, since every outcome is either in *A* or is in its complement. This observation gives us a valuable rule for calculating probabilities:

Complement Rule

The probability that *A* does not occur is $\mathrm{P}(A') = 1 - \mathrm{P}(A)$.

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. The additional information that event *B* occurs changes the sample space, since now we are interested *only* in those outcomes that are in *B*. And since *A* must also occur, we see that $\mathrm{P}(A \mid B) = \dfrac{\text{number of outcomes in } (A \text{ and } B)}{\text{number of outcomes in } B}$. If we divide the numerator and denominator by the number of outcomes in *S*, we get the equivalent formula:

Conditional Probability

The probability that event *A* occurs, *given* that event *B* also occurs is:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(B)}.$$

For example, suppose we toss one fair, six-sided die. The sample space *S* = {1, 2, 3, 4, 5, 6}. Let A = {2, 3}, and *B* = {2, 4, 6}. To calculate *P*(*A* | *B*), first count the number of outcomes common to the two events; the outcome 2 is in both. Using the first formula, we have $P(A \mid B) = \dfrac{\text{number of outcomes in } (A \text{ and } B)}{\text{number of outcomes in } B} = \dfrac{1}{3}$.

We get the same result by using the second formula: $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{1/6}{3/6} = \dfrac{1}{6} \cdot \dfrac{6}{3} = \dfrac{1}{3}$.

Understanding Terminology and Symbols

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

Example 3.1

The sample space S is the set of the first 20 whole numbers.
Let event A = the even numbers and event B = numbers greater than 13. Fill in the blanks:

a. A = ______________________, B = ______________________

b. P(A) = ______, P(B) = __________

c. A and B = _____________, A or B = ___________________________

d. P(A and B) = _________, P(A or B) = ___________

e. A' = _____________, P(A') = ___________

f. P(A) + P(A') = ____

g. P(A | B) = ______, P(B | A) = __________. Are the probabilities equal?

Solution 3.1

a. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$, $B = \{14, 15, 16, 17, 18, 19, 20\}$

b. $P(A) = \frac{10}{20} = \frac{1}{2}$; $P(B) = \frac{7}{20}$

c. A and $B = \{14, 16, 18, 20\}$; A or $B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19, 20\}$

d. $P(A \text{ and } B) = \frac{4}{20}$; $P(A \text{ or } B) = \frac{13}{20}$

e. $A' = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$; $P(A') = \frac{10}{20} = \frac{1}{2}$

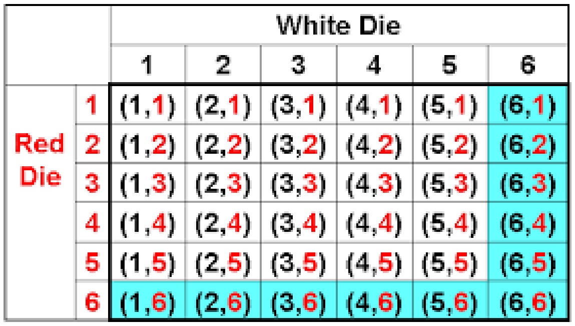
f. $P(A) + P(A') = \frac{10}{20} + \frac{10}{20} = \frac{20}{20} = 1.$

g. $P(A \mid B) = \frac{\text{number of outcomes in } (A \text{ and } B)}{\text{number of outcomes in } B} = \frac{4}{7}$;

$P(B \mid A) = \frac{\text{number of outcomes in } (A \text{ and } B)}{\text{number of outcomes in } A} = \frac{4}{10}$. The probabilities are not equal.

Example 3.2

Suppose that we roll two fair dice, one red and one white. Then there are 36 possible outcomes in the sample space, as shown in the table below:



		White Die					
		1	2	3	4	5	6
Red Die	1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
	2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
	3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
	4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
	5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
	6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Suppose we write the sample space in terms of the number of dots facing up on the two dice. That is, write $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Use the table to answer each of the following:

a. Are the outcomes in *S* equally likely?
b. Find P(4).
c. Find P(6).
d. Let *E* be the event "number of dots is even". Find P(*E*).
e. Let *G* be the event "number of dots is greater than 9". Find P(*G*).
f. Suppose that we rolled the dice 7200 times. How many times would we expect to roll a 7?

Solution 3.2

a) The events are not equally likely – for example the probability of rolling "snake eyes" is P(2) = 1/36 whereas the probability of rolling a "3" is P(3) = 2/36.

b) There are 3 outcomes in which the dots add to 4, so P(4) = 3/36.

c) There are 5 outcomes in which the dots add to 6, so P(6) = 5/36.

d) The event $E = \{2, 4, 6, 8, 10, 12\}$. So we add the probabilities of the outcomes:

$$\mathrm{P}(E) = \mathrm{P}(2) + \mathrm{P}(4) + \mathrm{P}(6) + \mathrm{P}(8) + \mathrm{P}(10) + \mathrm{P}(12) = \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36}$$

e) The event $G = \{10, 11, 12\}$. So we add the probabilities of the outcomes:

$$\mathrm{P}(G) = \mathrm{P}(10) + \mathrm{P}(11) + \mathrm{P}(12) = \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{6}{36}$$

f) The theoretical probability of rolling a 7 is $\mathrm{P}(7) = \frac{6}{36} = \frac{1}{6}$.

By the Law of Large Numbers, we expect the proportion of rolls that equal 7 to be about 1/6. So we would expect the number of 7's to be about $\frac{7200}{6}$ = **1200**.

Example 3.3

The table below describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	**Left-handed**
Males	43	9
Females	44	4

Let's denote the events as M = the subject is male, F = the subject is female, R = the subject is right-handed, and L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$ b. $P(F)$ c. $P(R)$ d. $P(L)$

e. $P(M \text{ and } R)$ f. $P(F \text{ and } L)$ g. $P(M \text{ or } F)$ h. $P(M \text{ or } R)$

i. $P(F \text{ or } L)$ j. $P(M')$ k. $P(R \mid M)$ l. $P(L \mid F)$

Solution 3.3

a. $P(M) = 0.52$

b. $P(F) = 0.48$

c. $P(R) = 0.87$

d. $P(L) = 0.13$

e. $P(M \text{ and } R) = 0.43$

f. $P(F \text{ and } L) = 0.04$

g. $P(M \text{ or } F) = 1$

h. $P(M \text{ or } R) = \dfrac{43+9+44}{100} = 0.96$

i. $P(F \text{ or } L) = \dfrac{44+4+9}{100} = 0.57$

j. $P(M') = P(F) = 0.48$

k. $P(R \mid M) = \dfrac{\text{number of outcomes in } R \text{ and } M}{\text{number of outcomes in } M} = \dfrac{43}{52} = 0.8269$

l. $P(L \mid F) = \dfrac{\text{number of outcomes in } L \text{ and } F}{\text{number of outcomes in } F} = \dfrac{4}{48} = 0.0833$

3.2 | Independent and Mutually Exclusive Events

When applying the rules of probability, we will sometimes need to consider the relationship between two events; in particular, we will need to consider whether two events are *independent* or whether they are *mutually exclusive*. Although students often confuse the two, they do **not** mean the same thing.

Independent Events

Two events A and B are **independent** if the knowledge that one occurred does not affect the probability that 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**. If it is not known whether A and B are independent or dependent, you should assume they are dependent until you can show otherwise.

To show that two events are independent, we show that any of following statements is true:

- $\text{P}(A \mid B) = \text{P}(A)$
- $\text{P}(B \mid A) = \text{P}(B)$
- $\text{P}(A \text{ and } B) = \text{P}(A)\text{P}(B)$

The first two are just restatements of the definition. The third statement follows from first two statements and the formula for conditional probability. If we multiply both sides of the equation

$$\text{P}(A \mid B) = \frac{\text{P}(A \text{ and } B)}{\text{P}(B)}$$

by $\text{P}(B)$, we get $\text{P}(A \text{ and } B) = \text{P}(B)\text{P}(A \mid B)$. If the events are independent, we substitute $\text{P}(A)$ in place of $\text{P}(A \mid B)$ to get $\text{P}(A \text{ and } B) = \text{P}(A)\text{P}(B)$.

There are some situations where we will have information about the probabilities and/or conditional probabilities, and we will use the rules above to determine whether the events are independent. There are other situations where the events are known to be independent, and then we will use this fact to help compute the probability of $\text{P}(A \text{ and } B)$.

Recall that sampling may be done **with replacement** or **without replacement**.

- **Sampling with replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then the sample space will not change from one selection to the other; meaning the result of the first pick will not change the probabilities for the second pick. Thus events will be independent from one selection to the next.

- **Sampling without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

Example 3.4

Suppose we have a standard, well-shuffled deck of 52 cards. It consists of four suits: clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q and K. Suppose that we select two cards. The first card is a Queen of Diamonds.

a. Find the probability of selecting a queen on the second draw if the first card is replaced before the second is drawn.

b. Find the probability of selecting a queen on the second draw if the first card is not replaced before the second is drawn.

Solution to 3.4:

a. If the first card is replaced before the second is drawn, then the second draw is independent of the first. So the probability of selecting a queen on the second draw is $\frac{4}{52}$.

b. If the first card is *not* replaced before the second is drawn, then the second draw is dependent on the first draw. There are now 51 cards left, of which 3 are queens. So the probability of selecting a queen on the second draw is $\frac{3}{51}$.

Try It Σ

3.4 Suppose we have a standard, well-shuffled deck of 52 cards. It consists of four suits: clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q and K. Suppose that we select two cards. The first card is a Queen of Diamonds.

a) Find the probability of selecting a diamond on the second draw if the first card is replaced before the second is drawn.

b. Find the probability of selecting a diamond on the second draw if the first card is not replaced before the second is drawn.

In Chapter 1 we saw that if we are selecting a sample from a very large population, the probability of getting the same individual twice will be very small. As a result, there is virtually no difference between sampling with replacement and treating without replacement. Thus, when sampling from a very large population, we can think of the individual selections as being independent of one another. This is useful for the next example:

Example 3.5

In a large metropolitan area, it is known that 45% of all voters are registered as Democrats. Suppose we select two voters at random.

a. What is the probability that both are registered as Democrats?
b. What is the probability that neither are registered as Democrats?

Solution 3.5

Let event D_1 = Democrat is selected on the first choice, and D_2 = Democrat is selected on the second**.** The key observation here is that the draws are independent, so we can use the rule P(D_1 and D_2) = P(D_1)P(D_2).

a. The probability that both are registered as Democrats is:

P(D_1 and D_2) = P(D_1) P(D_2) = 0.45 x 0.45 = **0.2025.**

b. If 45% are registered as Democrats, then the remaining 55% are *not* registered as Democrats.

P(D_1' and D_2') = P(D_1') P(D_2') = 0.55 x 0.55 = **0.3025.**

Try It Σ

3.5 Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Example 3.6

Let G = event that a student is taking a math class. Let H = event that a student is taking a science class. Then, G and H = the event that a student is taking *both* a math class and a science class. Suppose that P(G) = 0.6, P(H) = 0.5, and P(G and H) = 0.3. Based on this information, are G and H independent?

Solution 3.6

To show that G and H are independent, we 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 option we choose depends on the information given in the problem. We could choose any of the methods here because we have the necessary information.

For example, we can show that P(G | H) = P(G):

$$\mathrm{P}(G \mid H) = \frac{\mathrm{P}(G \text{ and } H)}{\mathrm{P}(H)} = \frac{0.3}{0.5} = 0.6 = \mathrm{P}(G)$$

Or, we can show P(G and H) = P(G)P(H):

$$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ and } H)$$

Since G and H are independent, knowing that a person is taking a science class does not affect the probability that he or she is taking a math class.

Try It Σ

3.6 In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, 4. One marble is selected. Let R = event that a red marble is drawn; let G = event that a green marble is drawn; let O = event that an odd-numbered marble is selected.

a) Calculate P(G and O).

b) Are P and O independent? Explain.

Example 3.7

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. One student is picked randomly. Let F be the event that a student is female, and let L be the event that a student has long hair. Are the events F and L independent?

Solution 3.7

There are three conditions we can check to determine independence:

$$P(F) = P(F \mid L)$$
$$P(L) = P(L \mid F)$$
$$\text{or} \quad P(F \text{ and } L) = P(F)P(L)$$

The one we use will depend on the information given in the problem. So we first write down the probabilities that are given in the problem:

$$P(F) = 0.60; \quad P(L) = 0.50; \quad P(F \text{ AND } L) = 0.45; \quad P(L \mid F) = 0.75.$$

Based on this information we could use either the second condition or the third.
(We do not know $P(F \mid L)$ yet, so we cannot use the first condition.)

Using the second rule, we check whether P(L | F) equals P(L):
We are given that P(L |F) = 0.75, whereas P(L) = 0.50. Since these are not equal, the events are *not* independent. This shows that a female student is more likely to have long hair than is a male student.

Using the third rule, we check whether P(F and L) = P(F)P(L):
We are given that P(F and L) = 0.45, whereas P(F)P(L) = (0.60)(0.50) = 0.30. Since these are not equal, the events F and L are not independent.

Mutually Exclusive Events

We say that *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 so it follows that P(*A* and *B*) = 0.

For example, suppose that *S* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, *A* = {1, 2, 3, 4, 5}, *B* = {4, 5, 6, 7, 8}, and *C* = {7, 9}. Then *A* and *B* = {4, 5}; so the events *A, B* are not mutually exclusive. On the other hand, the events *B, C* are mutually exclusive, since they do not have any outcomes in common.

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 3.8

Suppose that we toss two fair coins. The sample space is {*HH*, *HT*, *TH*, *TT*} where *T* = tails and *H* = heads. The possible outcomes are *HH*, *HT*, *TH*, and *TT*. Note that the outcomes *HT* and *TH* are different. The outcome *HT* means that the first coin showed heads and the second coin showed tails. The outcome *TH* means that the first coin showed tails and the second coin showed heads.

- Let *A* = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then *A* can be written as {*HH*, *HT*, *TH*}. The outcome *HH* shows zero tails. *HT* and *TH* each show one tail.
- Let *B* = the event of getting all tails. *B* can be written as {*TT*}. Note that *A,B* are mutually exclusive. In fact, *B* is the **complement** of *A*, so P(*A*) + P(*B*) = P(*A*) + P(*A*′) = 1.
- P(*A*) = 3/4, and P(*B*) = 1/4.
- Let *C* = the event of getting all heads; then *B* and *C* are mutually exclusive. Clearly *B*, *C* have no outcomes in common because it is not possible to have all tails and all heads at the same time.
- Let *D* = event of getting **more than one** tail. *D* = {*TT*}, so P(*D*) = 1/4.
- Let *E* = event of getting a head on the first toss. This means that we can get either a head or tail on the second toss, so *E* = {*HT*, *HH*}. Thus, P(*E*) = 2/4.
- Find the probability of getting **at least one** (one or two) tail in two flips. Let *F* = event of getting at least one tail in two flips. Then *F* = {*HT*, *TH*, *TT*} and so P(*F*) = 3/4.

Example 3.9

Suppose that we toss two fair coins. Find the probabilities of the events.

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

Solution to 3.9:

Refer to the sample space in **Example 3.7**.

a. Zero (0) or one (1) tails occur when the outcomes *HH*, *TH*, *HT* show up. P(*F*) = 3/4.
b. Two faces are the same if *HH* or *TT* show up. P(*G*) = 2/4 = 1/2.
c. A head on the first flip followed by a head or tail on the second flip occurs when *HH* or *HT* show up. So P(*H*) = 2/4 = 1/2.
d. *F* and *G* share the outcome *HH* so *F* and *G* are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins (*TT*). *H*'s outcomes are *HH* and *HT*. *J* and *H* have no outcomes in common, *J* and *H* are mutually exclusive.

Try It Σ

3.8 A box has two balls, one white and one red. We select one ball, put it back in the box, and then select a second ball. Find the probability of the following events:

a. *F* = the event of getting the white ball twice.
b. *G* = the event of getting two balls of different colors.
c. *H* = the event of getting white on the first pick.
d. Are *F* and *G* mutually exclusive?
e. Are *G* and *H* mutually exclusive?

Example 3.10

Let event *C* = taking an English class. Let event *D* = taking a speech class. Suppose P(*C*) = 0.75, P(*D*) = 0.3, P(*C* | *D*) = 0.75 and P(*C* and *D*) = 0.225. Use this information to answer the following:

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

Solution to 3.10

a. Yes, because P(*C* | *D*) = P(*C*).
b. No, because P(*C* and *D*) ≠ 0.
c. $P(D|C) = \frac{P(C \text{ and } D)}{P(C)} = \frac{0.225}{0.3} = 0.75$

3.3 | Two Basic Rules of Probability

Recall that in classical probability, the probability of an event is a fraction where the numerator is the number of outcomes in the event and the denominator is the number of outcomes in the sample space. When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

Addition Rule

If Events A and B are defined on the same sample space, then the probability of event A or event B is written as

If A and B are two events defined on the same sample space, then
P(A **or** B) = P(A) + P(B) - P(A and B).

If A and B are mutually exclusive which means that event A and event B can't happen at the same time; that is, P(A and B) = 0. Therefore the addition rule becomes:

If A and B are mutually exclusive, then:
P(A **or** B) = P(A) + P(B)

The Multiplication Rule

Recall that the probability of A given B equals the probability of A and B divided by the probability of B. That is, $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(B)}$. Multiplying both sides by P(B), we have:

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

If A and B are independent, then P(A|B) = P(A). So for independent events, the multiplication rule simplifies to:

If A and B independent, then:
P(A **and** B) = P(A)·P(B)

Example 3.11

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

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) = 0.6 + .35 - 0 = .95$$

Note that the probability that he does not choose to go anywhere on vacation must be 0.05

Example 3.12

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. Let A = the event Carlos is successful on his first attempt, so P(A) = 0.65. Let 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.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are A and B independent?
d. Are A and B mutually exclusive?

Solution to 3.12:

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

$$\begin{aligned} \mathrm{P}(B \text{ AND } A) &= \mathrm{P}(B|\mathrm{A})\mathrm{P}(A) \\ &= (0.90)(0.65) \\ &= 0.585 \end{aligned}$$

Carlos makes the first and second goals with probability 0.585.

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

$$\begin{aligned} \mathrm{P}(A \text{ OR } B) &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ AND } B) \\ &= 0.65 + 0.65 - 0.585 \\ &= 0.715 \end{aligned}$$

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

c. No, they are not independent events, because P(B AND A) = 0.585.
$\mathrm{P}(B)\cdot\mathrm{P}(A) = (0.65)(0.65) = 0.423$
$0.423 \neq 0.585 = \mathrm{P}(B \text{ AND } A)$ So, P(B AND A) is not equal to P(B)·P(A).

d. No, they are not mutually exclusive because P(A and B) = 0.585.
To be mutually exclusive P(A and B) = 0.

Try It Σ

3.11 Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. P(C) = 0.75. D = the event Helen makes the second shot. P(D) = 0.75. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Example 3.13

A community swim team has 150 members. Seventy-five of the members are advanced swimmers. Forty- seven of the members are intermediate swimmers. The rest are novice swimmers. Forty of the advanced swimmers practice four times a week. Thirty of the intermediate swimmers practice four times a week. Ten of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution to 3.13

a. 28/150
b. 80/150
c. 40/150
d. 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.
e. No, these are not independent events.
P(novice and practices four times per week) = 0.0667
P(novice)P(practices four times per week) = 0.0996
Since $0.0667 \neq 0.0996$, the events are not independent.

Try It Σ

3.12 A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder will be taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Example 3.14

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class given that she enrolls in speech class is 0.25. Let: M = math class, S = speech class.

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

Solution to 3.14

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

Try It Σ

3.13 A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that P(B) = 0.40, P(D) = 0.30 and P($D|B$) = 0.5.

a. Find P(B and D).
b. Find P(B or D)

Example 3.15

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

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative? What is the probability that the woman has breast cancer or tests negative? Are having breast cancer and testing negative independent events? Are having breast cancer and testing negative mutually exclusive?

Solution to 3.15

a. P(B) = 0.143; P(N) = 0.85
b. P($N|B$) = 0.02
c. P(B and N) = P(B)·P($N|B$) = (0.143)(0.02) = 0.0029
d. P(B or N) = P(B) + P(N) - P(B and N) = 0.143 + 0.85 - 0.0029 = 0.9901

e. No. P(*N*) = 0.85; P(*N*|*B*) = 0.02. So, P(*N*|*B*) does not equal P(*N*).
f. No. P(*B* and *N*) = 0.0029. For *B*, *N* to be mutually exclusive, P(*B* and *N*) must equal 0.

Try It Σ

3.14 A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder is taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Example 3.16

Refer to the information in **Example 3.14,** and let *P* = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive? Find P(*P*|*B*) = 1 - P(*N*|*B*).
b. What is the probability that a woman develops breast cancer and tests positive. Find P(*B* and *P*) = P(*P*|*B*)·P(*B*).
c. What is the probability that a woman does not develop breast cancer? Find P(*B*′).
d. What is the probability that a woman tests positive for breast cancer. Find P(*P*) = 1 - P(*N*).

Solution to 3.16

a. 0.98;
b. 0.1401;
c. 0.857;
d. 0.15

Try It Σ

3.15 A student goes to the library. Let events *B* = the student checks out a book and *D* = the student checks out a DVD. Suppose that P(*B*) = 0.40, P(*D*) = 0.30 and P(*D*|*B*) = 0.5.

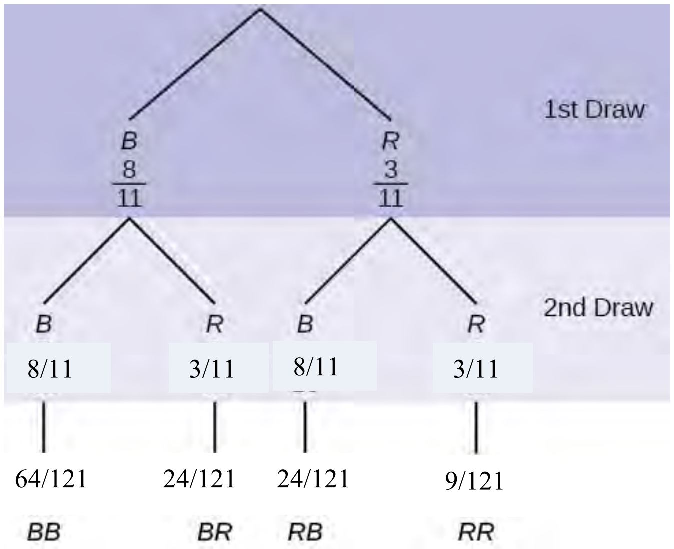
a. Find P(*B*′).
b. Find P(*D* and *B*).
c. Find P(*B*|*D*).
d. Find P(*D* AND *B*′).
e. Find P(*D*|*B*′).

Tree Diagrams and Venn Diagrams

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

Example 3.17

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, with replacement. "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.



The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8. Then the nine RR outcomes can be written as:
R1R1; R1R2; R1R3; R2R1; R2R2; R2R3; R3R1; R3R2; R3R3.
The other outcomes are similar.

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

a. List the 24 BR outcomes: B1R1, B1R2, B1R3, ...
b. Using the tree diagram, calculate P(RR)
c. Using the tree diagram, calculate *P*(RB or BR)..
d. Using the tree diagram, calculate P(R on 1st draw and B on 2nd draw).
e. Using the tree diagram, calculate P(R on 2nd draw given B on 1st draw)
f. Using the tree diagram calculate P(BB)

Solution to 3.17

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

b. P(RR) = P(1st Draw Red and 2nd Draw Red) = (3/11)·(3/11) = (9/121)
Note that we get this probability by "multiplying along the branches"

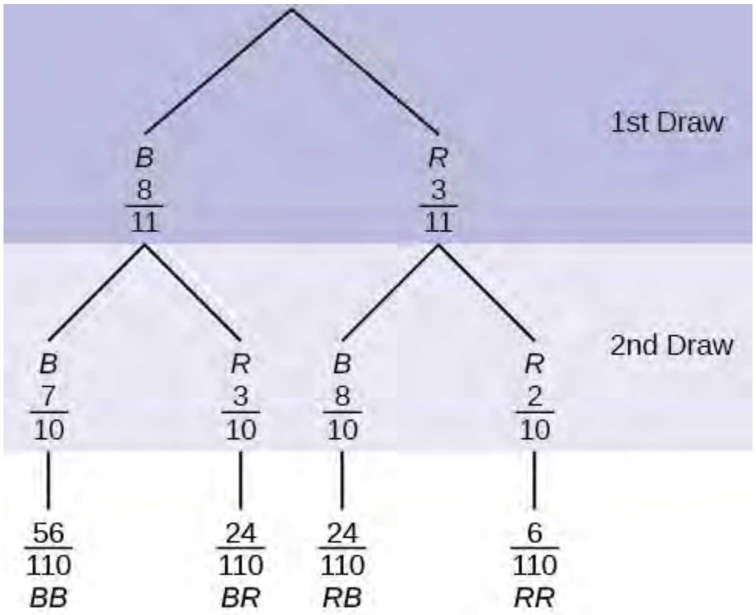
c. *P*(RB or BR) = (3/11)·(8/11) + (8/11)·(3/11) = (48/121)

d. P(R on 1st draw and B on 2nd draw) = (3/11)·(8/11) = (24/121)

e. P(R on 2nd draw GIVEN B on 1st draw) = P(R|B) = (24/88) = (3/11)
This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of those outcomes are BR.

f. P(BB) = (64/121)

If you do the same experiment but this time **without replacement**, then the tree diagram is



There are a total of 11 balls in the urn. Draw two balls, one at a time, without replacement. There are 11 × 10 = 110 outcomes; again this is the size of the **sample space**.

a. Using the tree diagram, calculate P(RR)
b. Using the tree diagram, calculate *P*(RB OR BR)..
c. Using the tree diagram, calculate P(R on 1st draw and B on 2nd draw).
d. Using the tree diagram, calculate P(R on 2nd draw GIVEN B on 1st draw)
e. Using the tree diagram calculate P(BB)

Solution for without replacement

a. P(RR) = P(1st Draw Red and 2nd Draw Red) = (3/11)·(2/10) = (6/110) = 3/55

b. *P*(RB OR BR) = (3/11)·(8/10) + (8/11)·(3/10) = (48/110) = 24/55

c. P(R on 1st draw and B on 2nd draw) = (3/11)·(8/10) = (24/110) = 12/55

d. P(R on 2nd draw GIVEN B on 1st draw) = P(R|B) = (3/10)

This is a conditional probability. Since a Blue was draw first that leaves only 10 balls left. From the 10 balls left, all 3 red are still in the urn.

e. P(BB) = (8/11)·(7/10) = (56/110) = 28/55

Try It Σ

3.16 In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. Draw a tree diagram is labeled with all possible probabilities.

A **Venn diagram** is another visual aid that can be used to display the relationships between different events for an experiment, and their probabilities. It generally consists of a box that represents the sample space *S* together with circles or ovals representing events.

Example 3.18

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 displaying this information, letting *C* = "student belongs to a club" and *PT* = "student works part time".

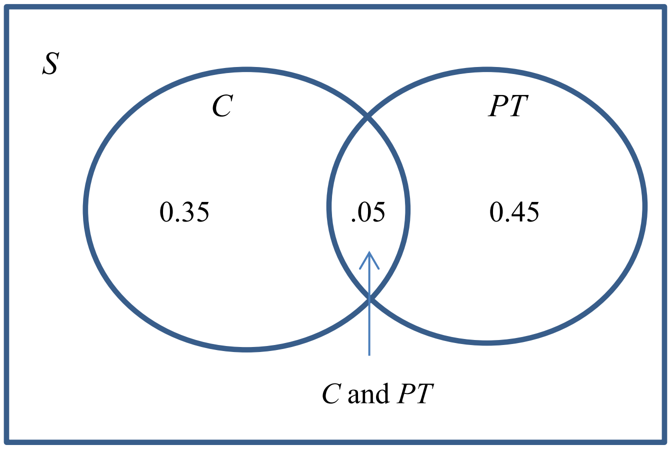
If a student is selected at random, find:

a. The probability that the student belongs to a club **given** that the student works part time.
b. The probability that the student belongs to a club **or** works part time
c. The probability that a student works part time, but does not belong to a club.
d. The probability that a student neither belongs to a club nor works part time.

Solution to 3.18:

We draw a rectangle representing the sample space *S*, and then draw two circles, one for *C* and the other representing *PT*. The overlap between the two circles represents the outcomes that the two events have in common. That is, this region represents the event *C* and *PT*.

Next, we fill in the probabilities, starting with the event *C* and *PT,* whose probability is .05. We are given that *P*(*C*) = .40; so the portion of *C* that does not overlap *PT* must have probability .35. Similarly, the part of *PT* that does not overlap *C* will have probability .45. This gives us the diagram:



a. The probability that the student belongs to a club **given** that the student works part time is:

$$P(C\,|\,PT) = \frac{P(C\text{ and }PT)}{P(PT)} = \frac{.05}{.50} = 0.10$$

b. The probability that the student belongs to a club **or** works part time can be found using the addition rule:

$P(C\text{ or }PT) = P(C) + P(PT) - P(C\text{ and }PT)$ = .40 + .50 – 0.05 = 0.85**.**

Or, we can just add the three probabilities that are inside the two circles:

$P(C\text{ or }PT)$ = .35 + .05 + .45 = **0.85.**

c. The outcomes where a student works part time, but does not belong to a club is represented by the portion of the *PT* circle that does not overlap with *C*. This probability is **0.45**.

d. The event that a student neither belongs to a club nor works part time is represented by the part of the box that is outside both of the circles. That is, this event is the complement of the event *C* or *PT*. By the complement rule,

P(neither in club nor working part time) = 1 – *P*(*C* or *PT*) = 1 – .85 = **0.15**.

Try It Σ

3.17 Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, and 5% work a second job *and* have a spouse who also works. Draw a Venn diagram showing the relationships, letting *W* = works a second job and *S* = spouse also works. Find the probability that a randomly selected worker would neither work a second job nor have a spouse that works.

Example 3.19

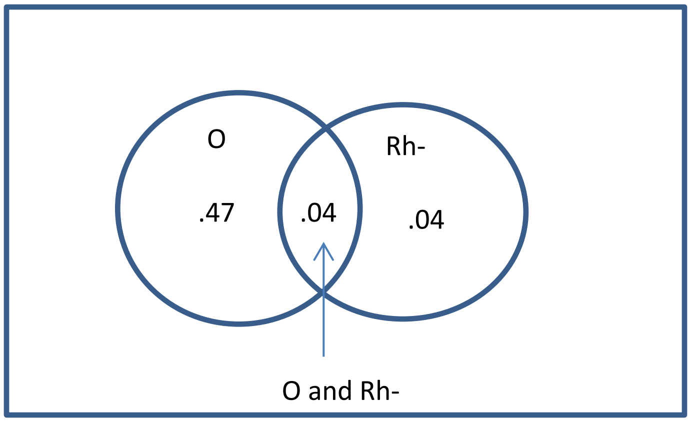
A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. It is known that 4% of African Americans have both type O blood and a negative Rh factor, 8% of African Americans have the Rh- factor, and 51% type O blood. Make a Venn diagram for this situation, using *O* for the set of individuals with type O blood, and Rh- for the individuals with the negative Rh factor.

a. Find the probability that a randomly selected African American has type O blood or negative Rh factor.

b. Find the probability that a randomly selected African American has negative Rh factor, but does not have type O blood.

c. Find the probability that a randomly selected African American has neither type O blood nor a negative Rh factor.

Solution to 3.19:



a. 55% b. 4% c. 45%

Try It Σ

3.18 In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or a non-fiction book.
c. Find the probability that a customer buys a novel but does not buy a non-fiction book.

3.4 | Contingency Tables

A **contingency table** is a commonly used way of displaying data that can facilitate calculating probabilities. The table is particularly helpful for calculating conditional probabilities. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Contingency tables will be used again later in the course as well.

Example 3.20

Suppose a study of speeding violations and drivers who use cell phones produced the data:

	Speeding violation in last year	**No speeding violation in last year**	**Total**
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

As we see in the bottom right corner, 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 we can get the total sample size by adding either the row totals or column totals:

$$305 + 450 = 755 \text{ and } 70 + 685 = 755.$$

Let C = event that person is a cell phone user; so C' = event that person is not a cell phone user. Let V = event that person had a speeding violation in the last year, so V′ = event that person had no speeding violation in the last year. Use the table to calculate the following probabilities:

a. Find P(C).
b. Find P(V').
c. Find P(V' and C).
d. Find P(V' or C).
e. Find P(C given V)
f. Find P(V' given C')

Solution to 3.20

a. $P(C) = \dfrac{\text{number of outcomes in } C}{\text{number of outcomes in } S} = \dfrac{305}{755} \approx 0.404$

b. $P(V') = \dfrac{\text{number of outcomes in } V'}{\text{number of outcomes in } S} = \dfrac{685}{755} \approx 0.907$

c. $P(V' \text{ and } C) = \dfrac{\text{number of outcomes in } (V' \text{ and } C)}{\text{number of outcomes in } S} = \dfrac{280}{755} \approx 0.371$

d. For part d, we can use the addition rule:

$$P(V' \text{ or } C) = \frac{\text{number of outcomes in } (V' \text{ or } C)}{\text{number of outcomes in } S}$$

$$= P(V') + P(C) - P(V' \text{ and } C) = \frac{685}{755} + \frac{305}{755} - \frac{280}{755} \approx 0.940$$

Or, we can find the number of outcomes in simply add the cells that are in either V' or C, being careful not to count any outcome twice:

$$P(V' \text{ or } C) = \frac{\text{number of outcomes in } (V' \text{ or } C)}{\text{number of outcomes in } S} = \frac{280 + 405 + 25}{755} \approx 0.940$$

e. To find $P(C$ given $V)$, we use the conditional probability formula:

$$P(C \mid V) = \frac{\text{number of outcomes in } (C \text{ and } V)}{\text{number of outcomes in } V} = \frac{25}{70} \approx .357$$

Note that the sample space is limited to the number of individuals who had a violation.

f. To find $P(V'$ given $C')$, we again use the conditional probability formula:

$$P(V' \mid C') = \frac{\text{number of outcomesin } (C' \text{ and } V')}{\text{number of outcomesin } C'} = \frac{405}{450} = .90$$

This time the sample space is limited to the number of individuals who were not cell phone users.

Try It Σ

3.19 The following table shows the number of athletes who stretch before exercising and how many injuries within the past year had.

	Injury in last year	**No injury in last year**	**Total**
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. Find the probability that an athlete stretches before exercising.

b. Find the probability that an athlete stretches before exercising given that there was no injury in the last year.

Example 3.21

The table below shows the hiking preferences for a random sample of 100 hikers.

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

a. Complete the table

b. Let F = event "being female" and C = event "Prefers the coastline". Are F and C independent?

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams.

d. Find the probability that a person is female or prefers hiking on mountain peaks.

Solution to 3.21.

a. We start by filling in the grand total; this is the total sample size, 100. This is also the sum of the two row totals: 45 + 55 = 100. Next, we can fill in the entry in the row for "males" and the column "Near lakes and streams". The total for this column is 41, and so the missing value must be 41 – 16 = 25. Next, we fill in the entry in the row for males and the column "The Coastline". The row total is 55, and so this last entry must be 55 – 25 – 14 = 16:

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

Continuing in this way, we get the completed table:

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

When finished, make sure to check that both the row totals and the column totals add to 100.

b. To show that F and C are independent, we must show any one of the following:

$$P(C \mid F) = P(C) \quad \text{or} \quad P(F \mid C) = P(F) \quad \text{or} \quad P(C \text{ and } F) = P(C)P(F)$$

We'll use the last one. From the table we quickly calculate:

$$P(C) = \frac{34}{100} = .34, \;\; P(F) = \frac{45}{100} = .45 \text{ and } P(C \text{ and } F) = \frac{18}{100} = .18$$

Since $P(C \text{ and } F) = .18 \neq (.34)(.45) = P(C)P(F)$, the events are *not* independent.

c. Let L be the event “prefers lakes and streams” and M be the event “being male”. Then we want to calculate $P(M \mid L)$:

$$P(M \mid L) = \frac{\text{number of outcomes in } (M \text{ and } L)}{\text{number of outcomes in } L} = \frac{25}{41} \approx .610$$

d. If we let F = “female” and P = “prefers mountain peaks”, then we want to calculate

$$P(F \text{ or } P) = P(F) + P(P) - P(F \text{ and } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{200} = \frac{59}{100} = .59.$$

Try It Σ

3.20 The following table shows a random sample of 200 cyclists and the routes they prefer.

Gender	**Lake Path**	**Hilly Path**	**Wooded Path**	**Total**
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the male respondents, what is the probability that the cyclist prefers a hilly path?

b) Let M = males and H = hilly path. Are the events M and H *independent*?

Example 3.22

The table below contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Year	**Robbery**	**Burglary**	**Rape**	**Vehicle**	**Total**
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total:					

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

a. Find $P(2009 \text{ and Robbery})$.

b. Find $P(2010 \text{ and Burglary})$.

c. Find $P(2010 \text{ or Burglary})$.

d. Find $P(2011 \mid \text{Rape})$.

e. Find $P(\text{Vehicle} \mid 2008)$.

Solution to 3.22:

a. 0.0294; b. 0.1551; c. 0.7165; d. 0.2365; e. 0.2575

Try It Σ

3.21 The following table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Total				

a. Find the total for each row and column

b. Find the probability that a randomly chosen individual from this group is Tall.

c. Find the probability that a randomly chosen individual from this group is Obese and Tall.

d. Find the probability that a randomly chosen individual from this group is Tall given that the individual is Obese.

e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.

f. Find the probability a randomly chosen individual from this group is Tall and Underweight.

g. Are the events Obese and Tall independent?

3.5 | Counting and Probability

We now present some tools for counting to better utilize the basic definition of theoretical probability. The underlying theme is to represent outcomes in a sample space as "lists" formed from some collection of symbols. The fundamental questions we will need to answer are:

1) Are the lists ordered or unordered?
2) Are the lists drawn with or without replacement?

In most cases, if we can explicitly describe the sample space, and can answer these basic questions, we will be able to count outcomes in our event and in the sample space.

The first counting principle is sometimes called the "Fundamental Principle of Counting"—as its name suggests, it forms the foundation of all our subsequent counting methods.

Fundamental Principle of Counting

Suppose we have a task that can be broken down into a sequence of *k* steps. Further suppose that there are

N_1 ways to do step 1, N_2 ways to do step 2, . . . , and N_k ways to do step *k*. Then there are $N_1 N_2 \cdots N_k$ ways to complete the entire task.

Notice that the Basic Counting Principle can be used only when counting an ordered arrangement—the statement explicitly states that the task is broken down into a *sequence* of steps**.**

Example 3.23

A license plate consists of three letters (from the English alphabet) followed by three digits. How many license plates are possible?

Solution to 3.23:

To form a valid license, we must fill in six slots: ___ ___ ___ ___ ___ ___

For each of the first three, there are 26 choices. For the last three, there are 10 choices each. Thus there are a total of $\underline{26} \text{ x } \underline{26} \text{ x } \underline{26} \text{ x } \underline{10} \text{ x } \underline{10} \text{ x } \underline{10} = 26^3 10^3$ total possibilities.

Example 3.24

a. How many social security numbers are possible?
b. What is the probability that a social security number has no repeated digits?
c. What is the probability that a social security number has at least one repeated digit?

Solution to 3.24:

a. To form a valid social security number, we need to fill in nine slots:

___ ___ ___ - ___ ___ - ___ ___ ___ ___

There are 10 choices for each slot, hence there are 10^9 possible social security numbers.

b. Note that the calculation in part a counted the number of outcomes in the sample space for this problem. Again, we draw nine slots:

___ ___ ___ - ___ ___ - ___ ___ ___ ___

There are 10 choices for the first slot. Since we cannot have repeated digits, there are only 9 choices for the second, only 8 choices for the third, etc. Thus there is a total of

$\underline{10}$ x $\underline{9}$ x $\underline{8}$ x $\underline{7}$ x $\underline{6}$ x $\underline{5}$ x $\underline{4}$ x $\underline{3}$ x $\underline{2}$ = 3,628,800

Possible social security numbers without repeated digits. Thus, the probability that a social security number has no repeated digits is:

$$P(\text{no repeats}) = \frac{\#(\text{no repeats})}{\text{totalnumbef of SSN' s}} = \frac{3{,}628{,}800}{10^9} \approx 0.0036.$$

c. This problem is easy once the previous problem is completed. Noting that "at least one repeat" is the complement of "no repeats", we use the complement rule to get:

Pr (at least one repeat) = 1 – Pr (no repeats)

= 1 – 0.0036 = **0.9964**.

Try It Σ

3.23 A company uses five digit identification numbers for their employees. What is the probability that a randomly selected employee has no repeated digits in his/her ID number?

Factorial Notation

The product of the first n positive integers is often abbreviated by $n!$; this is read as "n factorial". That is,

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

Note that $n! = n\cdot(n-1)!$.

Also, it is convenient to define 0! = 1! = 1.

To understand why $n! = n\cdot(n-1)!$, we consider a simple example and look at 7!.

By definition, $7! = 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1$. But we can also put parentheses around the last six factors to get $7! = 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1 = 7\cdot(6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1) = 7\cdot 6!$

As our examples above suggest, factorial notation is very convenient for counting problems using the basic counting principle; that is, counting ordered arrangements.

Example 3.25:

How many ways can eight people line up for a photograph?

Solution to 3.25:

We have eight slots to fill in: ___ ___ ___ ___ ___ ___ ___ ___ .

There are 8 choices for the first, 7 choices for the second, 6 choices for the third, etc. Thus there is a total of 8 x 7 x 6 x 5 x 4 x 3 x 2 x 1 = **8!** ways to line them up.

The basic counting rules for ordered arrangements can be summarized as follows:

Counting Ordered Arrangements

1) There are $k!$ ways to order k distinct objects.

2) The number of ordered arrangements of k objects, chosen without replacement from a collection of n objects, is

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} .$$

3) The number of ordered arrangements of k objects, chosen with replacement from a collection of n objects, is n^k .

Justification: In each of the three cases, we have k slots to fill:

___ ___ ___ . . . ___ ___ ___

In case (1), there are k choices for the first, $k-1$ choices for the second,...,and only one choice for the last slot. Hence there are $k(k-1)\cdots 2\cdot 1 = k!$ possible orderings.

In case (2), there are n choices for the first, $n-1$ choices for the second, . . ., and $n-(k-1) = n-k+1$ choices for the last slot. Hence there is a total of

$$n(n-1)(n-2)\cdots(n-k+1)$$

possible arrangements. Multiplying and dividing by $(n-k)!$ yields the alternative way of writing the expression.

In case (3), we are choosing *with* replacement, so there are n choices for each slot, hence a total of $n\cdot n\cdots n = n^k$ possible arrangements.

NOTES These formulas are recorded mostly for future reference. When counting ordered arrangements, it is best to simply use the basic counting principle directly—write out the correct number of slots, fill in the number of choices for each slot, and multiply.

Example 3.26

A club at a local community college has 14 men and 18 women. Suppose that the club will elect a president, vice president and treasurer.

a. What is the probability that the president will be female, and the other two officers male?
b. What is the probability that all three officers are female?

Solution to 3.26:

First we note that there are a total of 32 members in the club; and no person can fill two different posts, so there will be no repetitions. By the Fundamental Counting Principle, there is a total of

32 x 31 x 30 = 29,760 ways to choose the officers.

a. There are 18 women to choose from to serve as president. Then there are 14 men, so there are 14 choices for vice president and 13 choices for treasurer. Thus there is a total of 18 x 14 x 13 = 3276 ways to get a female president, and male vice president and treasurer. So the probability is 3276/29,760 = **0.110**.

b. There are 18 women to choose from. So there are 18 choices for president, 17 choices for vice president, and 16 choices for treasurer. Thus there is a total of 18 x 17 x 16 = 4896 ways to get all female officers. So the probability is 4896/29,760 = **0.165**.

Try It Σ

3.25 Suppose that a club has 20 mean and 22 women. If they are to elect a president, vice president and secretary, what is the probability that the officers will all be men? What is the probability that at least one woman is elected as an officer?

Permutations and Combinations:

An ordered arrangement of *k* objects, chosen without replacement from a collection of *n* objects, is called a *permutation*. The total number of such arrangements is denoted by ${}_{n}P_{k}$,. And as we saw above, this number can be calculated using the following formula:

$${}_{n}P_{k} = \frac{n!}{(n-k)!}$$

This formula is not really necessary in practice, because any time we want to count ordered arrangements, we can just use the fundamental counting principle. However, the formula is necessary to develop a formula for counting unordered arrangements.

An unordered arrangement of k objects, chosen without replacement from a collection of n objects, is called a *combination*. The total number of such arrangements is denoted by ${}_nC_{k,}$, and is given by the formula

$${}_nC_k = \frac{n!}{k!(n-k)!} \quad \text{(read as "}n\text{ choose }k\text{")}$$

To understand where this formula comes from, we think about the the number of ordered arrangements of k objects chosen without replacement from a set of n objects. We can such an arrangement in two steps: First, we select a subset of k objects from the set; then we put them in order. By the Fundamental Principle of Counting, we have:

Total # of ordered arrangements = <u>Total # of subsets of size *k*</u> × <u>Number of ways to order *k* objects</u>

We know that the number on the left is ${}_nP_k = \dfrac{n!}{(n-k)!}$. And we know that there are k! ways to order k distinct objects. Finally, the number of subsets of size k is ${}_nC_{k,}$, which is what we are trying to find. So the equation above becomes: $\dfrac{n!}{(n-k)!} = {}_nC_k \cdot k!$ Dividing both sides by k!, we get ${}_nC_k = \dfrac{n!}{k!(n-k)!}$.

NOTE: The numbers ${}_nC_k$ are often called binomial coefficients, because they appear in the Binomial Theorem, which you may remember from your high school algebra class. In the next chapter they will also be used for an important probability distribution called the Binomial distribution.

Note also that the formulas for ${}_nP_{k,}$ and ${}_nC_{k,}$ are built into the TI-83 and TI-84 calculators:

<u>Using the TI-83, 83+, 84, 84+ Calculator</u>

To calculate ${}_nP_k$ and ${}_nC_k$ we go press the MATH button, and scroll to the PRB menu.

To find ${}_nP_{k,}$, enter n, then press MATH >> PRB; press 2 (nPr) and then enter k and press the ENTER key.

To find ${}_nC_{k,}$, enter n, then press MATH >> PRB; press 3 (nPr) and then enter k and press the ENTER key.

Try It Σ

Use the calculator to find each of the following:

${}_{10}P_3$, ${}_{10}C_3$, ${}_8P_4$, and ${}_8C_4$.

Example 3.27

A shipment of 40 DVD players contains eight defective units. Suppose that a sample of four players is selected, and that the entire shipment will be rejected if at least one DVD player in the sample is defective.

a. What is the probability that the sample contains no defective players?
b. What is the probability that the shipment will be rejected?

Solution to 3.27:

There are ${}_{40}C_4$ = 91,390 ways to choose a sample of four DVD players from the shipment.

a. There are 36 non-defective players, so there are ${}_{36}C_4$ = 58,905 ways to choose four non-defective players. Thus, the probability is P(no defective players) = 58,905/91,390 = **0.6445**, or about 64.5%.

b. The probability calculated in part a is the probability that the shipment is *not* rejected. So we can use the complement rule to get:

P(rejected) = 1 – P(no defective players) = 1 – 0.6445 = **0.3555.**

Example 3.28

A committee of four is to be selected from a class of 7 men and 9 women.

a. In how many ways can the committee be selected?
b. How many ways can the committee be chosen to have 3 women?
c. What is the probability that the committee has 3 women?
d. What is the probability that the committee has at least three women?
e. What is the probability that the committee has at least one man?

Solution to 3.28:

a. A committee is an unordered arrangement without replacement, so the formula above applies: We are choosing 4 objects from a set of 16 objects, so the total number of committees is ${}_{16}C_4$. Using the calculator, we press 16, then go to MATH >> PRB and select item 3 (nCr); then press 4 and ENTER to get ${}_{16}C_4$ = **1820**.

b. We will choose the committee in two stages: First select three women, then select one man to complete the committee. There are ${}_{9}C_3$ = 84 ways to choose three women from among the seven. There are ${}_{7}C1$ = 7ways to choose the man and complete the committee. Thus there are $({}_{9}C_3)({}_{7}C_1)$ = 84 x 7 = **588** committees with three women.

c. The probability that a committee has three women is

$$\text{P (three women)} = \frac{\#(\text{committeeswith threewomen})}{\text{total\# committees}} = \frac{588}{1820} = 0.323.$$

d. The probability that a committee has *at least* three women there are two possibilities: either there are exactly three women, or there are exactly four women. So the probability we want is P (three women) + P (four women).

We have already calculated the probability of getting three women on the committee, so we must now calculate the probability of getting four women on the committee. There are ${}_7C_4 = 35$ ways to choose the committee consisting of all women (verify this), and from part (a), there are 1820 committees altogether.

Thus the probability that the committee consists entirely of women is 35/1820 = .01923. Combining this with the result from part c we have:

P (three women) + P (four women) = 0.323 + 0.01923 = **0.34223**.

e. We will do this one using the complement rule. If there is not one or more men on the committee, then the committee must consist entirely of women. I.e. the complement of the event "at least one man" is the event "all women". Thus the probability that the committee has at least one man is

P(at least one man) = 1 – P(all women).

Using the result from part (d), the probability that the committee has at least one man is

P (at least one man) = 1 – Pr (all women) = 1 – 0.01923 = **0.98077**.

Some important things to notice about this example:

- Often we are using more than one counting principle at a time
- The overall strategy for computing probabilities is more important than the formulas.

KEY TERMS

Complement The complement of event *A* (denoted as *A′*) the set of outcomes that are *not* in *A*.

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

Contingency table A 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.

Dependent Events If two events are NOT independent, then we say that they are dependent.

Equally Likely Each outcome of an experiment has the same probability.

Event A subset of the set of all outcomes of an experiment; that is, an event is a subset of the sample space *S*. Standard notations for events are capital letters such as *A*, *B*, *C*, and so on.

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 another 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 Two events are mutually exclusive if they have no outcomes in common. Thus, if events *A* and *B* are mutually exclusive, then $P(A \text{ and } B) = 0$.

Outcome a particular result of an experiment

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

Sample Space the set of all possible outcomes of an experiment

Sampling with Replacement means that each member of a population is replaced after it is picked, so that there is a possibility that the same item is chosen more than once.

Sampling without Replacement means that items are not put back after being chosen, so each member of a population may be chosen only once.

The AND Event The event *A* and *B* consists of all outcome that are common to both events. So an outcome is in the event *A* and *B* if the outcome is in both *A* and in event *B*.

Tree Diagram a visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram A visual representation of a sample space and events in the form of circles or ovals showing their intersections.

FORMULA REVIEW

Theoretical Probability: For a sample space with equally likely outcomes,

$$\mathrm{P}(A) = \frac{\text{number of outcomesin } A}{\text{number of outcomesin } S}$$

Empirical Probability: $\mathrm{P}(A) = \dfrac{\text{number of times} A \text{ occurs}}{\text{totalnumber of trials}}$

Conditional Probability: $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(B)}$

Complement Rule: $P(A') = 1 - P(A)$

Addition Rules:

- For *any* events *A, B*: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- If A, B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$

Multiplication Rules:

- For *any* events *A, B*: $P(A \text{ and } B) = P(A)P(B \mid A)$
- If A, B are independent, then $P(A \text{ and } B) = P(A)P(B)$

Fundamental Principle of Counting.
Suppose we have a task that can be broken down into a sequence of *k* steps. Further suppose that there are N_1 ways to do step 1, N_2 ways to do step 2, . . . , N_k ways to do step *k*.
Then there are $N_1 N_2 \cdots N_k$ ways to do the entire task.

A *permutation* is an ordered arrangement of *k* objects, chosen without replacement from a collection of n objects. The total number of such arrangements is: ${}_nP_k = \dfrac{n!}{(n-k)!}$.

A *combination* is an ordered arrangement of *k* objects, chosen without replacement from a collection of n objects. The total number of such arrangements is: ${}_nC_k = \dfrac{n!}{k!(n-k)!}$

Exercises for Chapter 3

1. If the experiment is drawing two colored marbles from a box of 8 blue marbles, 10 green marbles, and 5 purple marbles, then state the sample space.

2. If the experiment is testing 3 items from a shipment box that contain defective (D) and non-defective (N) items. State the sample space.

3. A bag with 7 letters (A, B, C, D, E, F, G) is shaken and two letters are drawn without replacement. State the sample space.

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

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

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

5. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let H = the event of getting a hat.

Let N = the event of getting a noisemaker. Let F = the event of getting a finger trap.

Let C = the event of getting a bag of confetti.

a. Find P(H).
b. Find P(N).
c. Find P(F).
d. Find P(C).

6. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
   a. Let B = the event of getting a blue jelly bean. Find P(B)
   b. Let G = the event of getting a green jelly bean. Find P(G)
   c. Let O = the event of getting an orange jelly bean. Find P(O)
   d. Let P = the event of getting a purple jelly bean. Find P(P)
   e. Let R = the event of getting a red jelly bean. Find P(R)
   f. Let Y = the event of getting a yellow jelly bean. Find P(Y)

7. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

   a. Let A = the event that a country is in Asia. Find P(A).
   b. Let E = the event that a country is in Europe. Find P(E).
   c. Let F = the event that a country is in Africa. Find P(F).
   d. Let N = the event that a country is in North America. Find P(N).
   e. Let O = the event that a country is in Oceania. Find P(O).
   f. Let S = the event that a country is in South America. Find P(S).

8. A card is drawn from a standard deck. Find the following probabilities.
   a. What is the probability of a red card is drawn?
   b. What is the probability of a club is drawn?
   c. What is the probability of a 9 or diamond is drawn?
   d. What is the probability of a 4 or king is drawn?
   e. What is the probability of a 6 and heart is drawn?

9. The experiment is rolling a fair, six-sided die numbered one through six.
   a. What is the probability of rolling an even number of dots?
   b. What is the probability of rolling a prime number of dots?

10. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.
    a. Let B = the event of landing on blue. Find P(B)
    b. Let R = the event of landing on red. Find P(R′)
    c. Let G = the event of landing on green. Find P(G′)
    d. Let Y = the event of landing on yellow. Find P(Y).

11. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters. Let I = the event that a player in an infielder. Let O = the event that a player is an outfielder. Let H = the event that a player is a great hitter. Let N = the event that a player is not a great hitter.
    a. Write the symbols for the probability that a player is not an outfielder.
    b. Write the symbols for the probability that a player is an outfielder or is a great hitter.
    c. Write the symbols for the probability that a player is an infielder and is not a great hitter.
    d. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

e. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.
f. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.
g. Write the symbols for the probability that of all the great hitters, a player is an outfielder.
h. Write the symbols for the probability that a player is an infielder or is not a great hitter.
i. Write the symbols for the probability that a player is an outfielder and is a great hitter.
j. Write the symbols for the probability that a player is an infielder.

12. What is the word for the set of all possible outcomes?

13. What is conditional probability?

14. A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book. Let F = event that book is fiction. Let N = event that book is nonfiction. What is the sample space?

15. What is the sum of the probabilities of an event and its complement?

16. Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.
    a. What does P(E|M) mean in words?
    b. What does P(E OR M) mean in words?

17. Let E and F be mutually exclusive events. P(E) = 0.4; P(F) = 0.5. Find P(E|F).

18. Let J and K be independent events. P(J|K) = 0.3. Find P(J).

19. Let U and V be mutually exclusive events. P(U) = 0.26; P(V) = 0.37. Find:
    a. P(U AND V) =
    b. P(U|V) =
    c. P(U OR V) =

20. Let Q and R be independent events. P(Q) = 0.4 and P(Q AND R) = 0.1. Find P(R).

21. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino. Let C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder. Let L = Latino Californians. Suppose that one Californian is randomly selected.
    a. Find P(C).
    b. Find P(L).
    c. Find P(C|L).
    d. In words, what is C|L?

e. Find P(L AND C).
f. In words, what is L AND C?
g. Are L and C independent events? Show why or why not.
h. Find P(L OR C).
i. In words, what is L OR C?
j. Are L and C mutually exclusive events? Show why or why not.

22. The following table shows a random sample of musicians and how they learned to play their instruments.

	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

a. Find P(musician is a female).
b. Find P(musician is a male AND had private instruction).
c. Find P(musician is a female OR is self-taught).
d. Are the events “being a female musician” and “learning music in school” mutually exclusive events?

23. The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: C = a man develops cancer in his lifetime; P = man has at least one false positive. Construct a tree diagram of the situation.

24. An article in the New England Journal of Medicine, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

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

a. Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

b. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

c. Find the probability that the person was Latino.

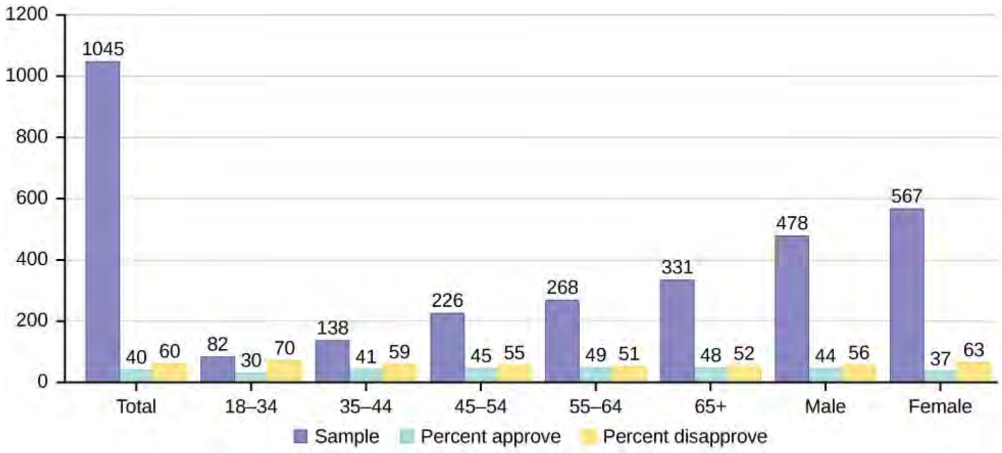
d. In words, explain what it means to pick one person from the study who is “Japanese American AND smokes 21 to 30 cigarettes per day.” Also, find the probability.

e. In words, explain what it means to pick one person from the study who is “Japanese American OR smokes 21 to 30 cigarettes per day.” Also, find the probability.

f. In words, explain what it means to pick one person from the study that is “Japanese American GIVEN that person smokes 21 to 30 cigarettes per day.” Also, find the probability.

g. Prove that smoking level/day and ethnicity are dependent events.

25. The graph in Figure 3.11 displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford’s actions in office. The total number in the sample of all the age groups is 1,045.



a. Define three events in the graph.

b. Describe in words what the entry 40 means.

c. Describe in words the complement of the entry in question 2.

d. Describe in words what the entry 30 means.

e. Out of the males and females, what percent are males?

f. Out of the females, what percent disapprove of Mayor Ford?

g. Out of all the age groups, what percent approve of Mayor Ford?

h. Find P(Approve|Male).

i. Out of the age groups, what percent are more than 44 years old?

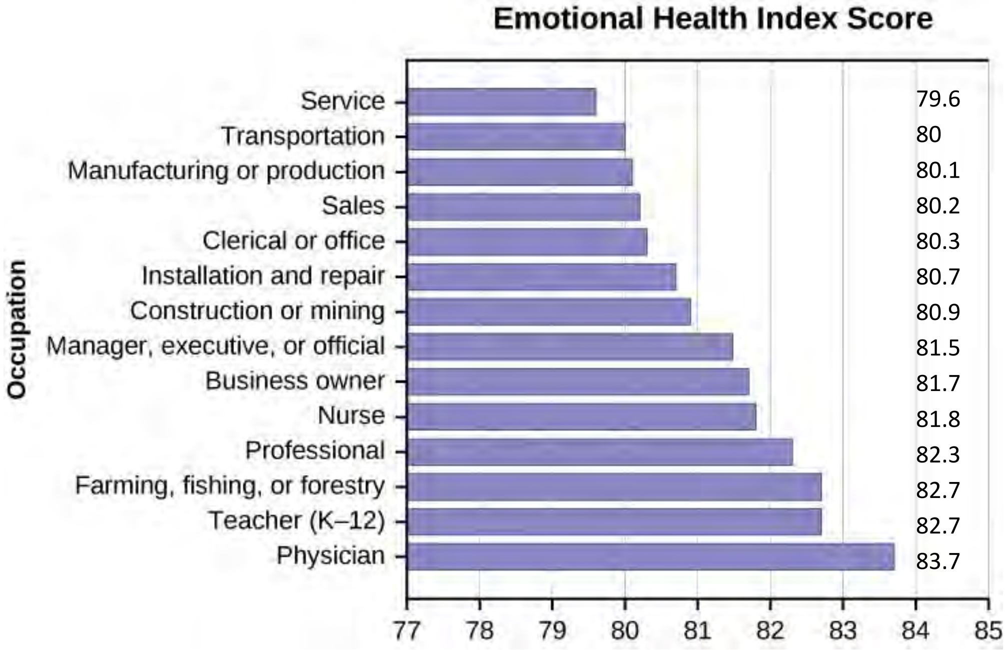
j. Find P(Approve|Age < 35).

26. Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.

b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

27. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.



a. Find the probability that an Emotional Health Index Score is 82.7.
b. Find the probability that an Emotional Health Index Score is 81.0.
c. Find the probability that an Emotional Health Index Score is more than 81?
d. Find the probability that an Emotional Health Index Score is between 80.5 and 82?
e. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?
f. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?
g. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81?
h. What occupation has the highest emotional index score?
i. What occupation has the lowest emotional index score?
j. What is the range of the data?
k. Compute the average EHIS.
l. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

28. On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

- C = California registered voters who support same-sex marriage.

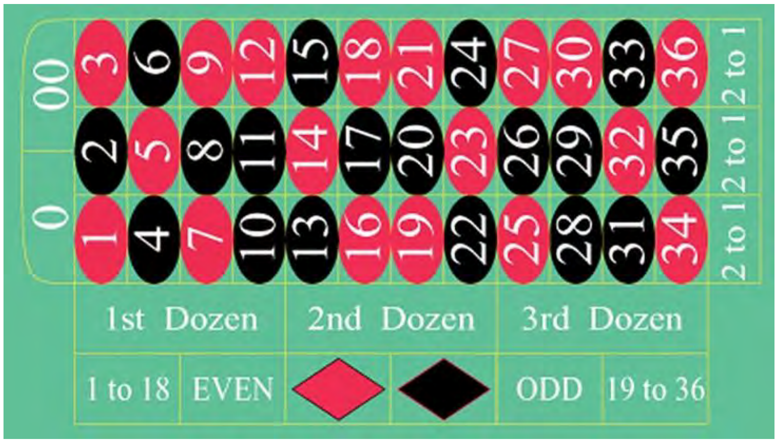
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

a. Find P(C).
b. Find P(B).
c. Find P(C|A).
d. Find P(B|C).
e. In words, what is C|A?
f. In words, what is B|C?
g. Find P(C AND B).
h. In words, what is C AND B?
i. Find P(C OR B)
j. Are C and B mutually exclusive events? Show why or why not.

29. After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:
    - In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
    - In mid-2011, 57 percent of the population approved of his actions.
    - In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

30. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.



(Photo credit: film8ker/wikibooks)

(Photo credit: gary128025/flicker)

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find P(red).

c. You bet on -1st 12- (1st Dozen). Find P(-1st 12-).
d. You bet on an even number. Find P(even number).
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

31. Using the information of the roulette wheel, compute the probability of winning the following types of bets:
a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

32. U sing the information of the roulette wheel, compute the probability of winning the following types of bets:
a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

33. Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.
a. List the sample space.
b. P(G) =
c. P(G|E) =
d. P(G AND E) =
e. P(G OR E) =
f. Are G and E mutually exclusive? Justify your answer numerically.

34. Roll two fair dice. Each die has six faces.
a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find P(A).
c. Let B be the event that the sum of the two rolls is at most seven. Find P(B).
d. In words, explain what "P(A|B)" represents. Find P(A|B).
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

35. A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.
a. List the sample space.

b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find P(A).
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

36. An experiment consists of first rolling a die and then tossing a coin.
a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find P(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

37. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.
a. List the sample space.
b. Let A be the event that there are at least two tails. Find P(A).
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

38. Consider the following scenario: Let P(C) = 0.4. Let P(D) = 0.5. Let P(C|D) = 0.6.
a. Find P(C AND D).
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find P(C OR D).
e. Find P(D|C).

39. Let Y and Z are independent events.
a. Rewrite the basic Addition Rule P(Y OR Z) = P(Y) + P(Z) - P(Y AND Z) using the information that Y and Z are independent events.
b. Use the rewritten rule to find P(Z) if P(Y OR Z) = 0.71 and P(Y) = 0.42.

40. Let G and H are mutually exclusive events. P(G) = 0.5 P(H) = 0.3
a. Explain why the following statement MUST be false: P(H|G) = 0.4.
b. Find P(H OR G).
c. Are G and H independent or dependent events? Explain in a complete sentence.

41. Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish. Let: E = speaks English at home; E′ = speaks another language at home; S = speaks Spanish; Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. P(E′)	i. 0.8043
b. P(E)	ii. 0.623
c. P(S and E′)	iii. 0.1957
d. P(S | E′)	iv. 0.1219

42. In 1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.
   a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
   b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
   c. Are G and F independent or dependent events? Justify your answer numerically
   d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

43. Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned. Let: R = money returned; E = economics classes; O = other classes
   a. Write a probability statement for the overall percent of money returned.
   b. Write a probability statement for the percent of money returned out of the economics classes.
   c. Write a probability statement for the percent of money returned out of the other classes.
   d. Is money being returned independent of the class? Justify your answer numerically
   e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

44. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	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	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

45. United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.
    a. Find the probability that a person has both type O blood and the Rh- factor.
    b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

46. At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.
    a. Find the probability that a course has a final exam or a research project.
    b. Find the probability that a course has NEITHER of these two requirements.

47. In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.
    a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
    b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

48. A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student.
    a. Find P(D AND E).
    b. Find P(E|D).
    c. Find P(D OR E).
    d. Using an appropriate test, show whether D and E are independent.
    e. Using an appropriate test, show whether D and E are mutually exclusive.

49. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Independent	Total
2016	10	24	0	
2018	23	8	2	
Total				

    a. What is the probability that a randomly selected senator has an "Independent" affiliation?
    b. What is the probability that a randomly selected senator is up for reelection in November 2016?
    c. What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?
    d. What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2018?
    e. Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?
    f. Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2018, knowing that this senator is a Republican?

g. The events “Republican” and “Up for reelection in 2016” are

i. mutually exclusive.
ii. independent.
iii. both mutually exclusive and independent.
iv. neither mutually exclusive nor independent.

h. The events “Independent” and “Up for reelection in November 2016” are

i. mutually exclusive.
ii. independent.
iii. both mutually exclusive and independent.
iv. neither mutually exclusive nor independent.

50. The table below gives the number of suicides estimated in the U.S. for a recent year by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Race and Sex	1 – 14	15 – 24	25 – 64	Over 64	Totals
white, male	210	3360	13,610		22,050
white, female	80	580	3380		4930
black, male	10	460	1060		1670
black, female	0	40	270		330
all others					
Totals	310	4650	18,780		29,760

Do not include "all others" for parts f and g.

a. Fill in the column for the suicides for individuals over age 64.

b. Fill in the row for all other races.

c. Find the probability that a randomly selected individual was a white male.

d. Find the probability that a randomly selected individual was a black female.

e. Find the probability that a randomly selected individual was black

f. Find the probability that a randomly selected individual was male.

g. Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.

51. The table of data obtained from www.baseball-almanac.com shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

a. Find P(hit was made by Babe Ruth).

i. $\frac{1518}{2873}$

ii. $\frac{2873}{12351}$

iii. $\frac{583}{12351}$

iv. $\frac{4189}{12351}$

Name	Single	Double	Triple	HR	Total
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

b. Find P(hit was made by Ty Cobb|The hit was a Home Run).

i. $\frac{4189}{12351}$

ii. $\frac{114}{1720}$

iii. $\frac{1720}{4189}$

iv. $\frac{114}{12351}$

52. The following table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	**Brown**	**Blond**	**Black**	**Red**	**Totals**
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

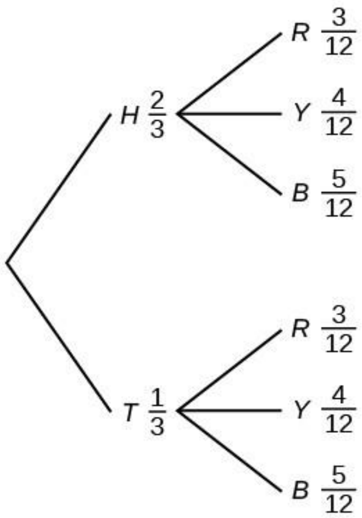
a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

53. In a previous year, the weights of the members of the San Francisco 49ers and the Dallas Cowboys were published in the San Jose Mercury News. The factual data were compiled into the following table. For the following, suppose that you randomly select one player from the 49ers or Cowboys.

Shirt #	**≤ 210**	**211 - 250**	**251 - 290**	**> 290**
1 - 33	21	5	0	0
34 - 66	6	18	7	4
67 - 99	6	12	22	5

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

54. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (R), four yellow (Y) and five blue (B) beads. For the coin, P(H) = $\frac{2}{3}$ and P(T) = $\frac{1}{3}$ where H is heads and T is tails.



a. Find P(tossing a Head on the coin AND a Red bead)
b. Find P(Blue bead).

55. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find P(S).
e. Let T be the event that the cookies selected were different flavors. Find P(T) by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find P(U).

56. A previous year, the weights of the members of the San Francisco 49ers and the Dallas Cowboys were published in the San Jose Mercury News. The factual data are compiled into the following table:

Shirt #	**≤ 210**	**211-250**	**251 – 290**	**290 ≤**
1 – 33	21	5	0	0
34 – 66	6	18	7	4
67 – 99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys. If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about P(Shirt# 1–33|≤ 210 pounds)?

57. The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.
    a. P(C) =
    b. P(P|C) =
    c. P(P|C') =
    d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

58. Given events G and H: P(G) = 0.43; P(H) = 0.26; P(H AND G) = 0.14
    a. Find P(H OR G).
    b. Find the probability of the complement of event (H AND G).
    c. Find the probability of the complement of event (H OR G).

59. Given events J and K: P(J) = 0.18; P(K) = 0.37; P(J OR K) = 0.45
    a. Find P(J AND K).
    b. Find the probability of the complement of event (J AND K).
    c. Find the probability of the complement of event (J or K).

60. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled. Suppose that you randomly draw two cards, one at a time, with replacement. Let G1 = first card is green. Let G2 = second card is green
    a. Draw a tree diagram of the situation.
    b. Find P(G1 AND G2).
    c. Find P(at least one green).
    d. Find P(G2|G1).
    e. Are G2 and G1 independent events? Explain why or why not.

61. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled. Suppose that you randomly draw two cards, one at a time, without replacement. G1 = first card is green G2 = second card is green
    a. Draw a tree diagram of the situation.
    b. Find P(G1 AND G2).
    c. Find P(at least one green).
    d. Find P(G2|G1).
    e. Are G2 and G1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

62. Complete the following.
    a. Construct a table or a tree diagram of the situation.
    b. Find P(driver is female).
    c. Find P(driver is age 65 or over|driver is female).
    d. Find P(driver is age 65 or over AND female).
    e. In words, explain the difference between the probabilities in part c and part d.
    f. Find P(driver is age 65 or over).
    g. Are being age 65 or over and being female mutually exclusive events? How do you know?

63. Suppose that 10,000 U.S. licensed drivers are randomly selected.
    a. How many would you expect to be male?
    b. Using the table or tree diagram, construct a contingency table of gender versus age group.
    c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

64. Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.
    a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
    b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
    c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
    d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

65. When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.
    a. Based on the given data, find P(H) and P(T).
    b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
    c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
    d. Use the tree to find the probability of obtaining at least one head.

66. Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	
Male	2146	463	60	135	
Totals					

a. Find P(Person is female).
b. Find P(Person has a risk factor heterosexual contact).
c. Find P(Person is female OR has a risk factor of IV drug user).
d. Find P(Person is female AND has a risk factor of homosexual/bisexual).
e. Find P(Person is male AND has a risk factor of IV drug user).
f. Find P(Person is female GIVEN person got the disease from heterosexual contact).
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

67. Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.
a. Find P(Person is female).
b. Find P(Person obtained the disease through heterosexual contact).
c. Find P(Person is female GIVEN person got the disease from heterosexual contact)
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

68. Suppose 22% of the population are 65 or older, 28% of those 65 or older have loans, and 55% of those younger than 65 have loans. Find the probabilities that a person fits into the following categories.
a. 65 or older and has a loan.
b. Has a loan
c. Younger than 65 and doesn't have a loan.
d. Are the events that a person is 65 or older and that the person has a loan independent.

69. In an election with 4 candidates for one office and 8 candidates for another office, how many different ballots may be printed?

70. How many different 4-letter call letters for a student radio station can be made if

a. The first letter must be a M or a K and no letter may be repeated?
b. Repeats are allowed but the first letter is a M or K?

71. How many different 6 space license plates can be made if
    a. The first three positions must be letters and the last three positions digits.
    b. The first three positions must be letters that do not repeat and the last three positions are digits.

72. A financial advisor gives her client 9 potential investments and asked her to select and rank her top 5. In how many different ways can she do this?

73. An internet installer is given 12 customers that need installation to be completed by the end of the day. How many different possible ways can the list of customers be arranged?

74. There are 5 defective parts in a crate of 28 parts.

    a. How many samples of 3 parts can be drawn from the crate?
    b. How many sample of 3 parts can be drawn in which all three are not defective?
    c. How many sample of 3 parts can be drawn in which 2 are not defective and 1 is defective?

75. In a club with 12 male and 10 female members, a 5 member committee will be randomly chosen. Find the probability that the committee contains the following
    a. All women
    b. 3 men and 2 women
    c. At least 4 women

REFERENCES

3.1 Terminology

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

3.2 Independent and Mutually Exclusive Events

Lopez, Shane, Preety Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

3.3 Two Basic Rules of Probability

DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at http://www.field.com/

fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

"Mayor's Approval Down." News Release by Forum Research Inc. Available online at http://www.forumresearch.com/
forms/News Archives/News Releases/74209_TO_Issues_-

_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013). "Roulette." Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013). Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2, 2013).

3.4 Contingency Tables

"Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood- types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

"Human Blood Types." Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at http://www.ehow.com/ facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

"United States: Uniform Crime Report – State Statistics from 1960–2011." The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Tree and Venn Diagrams

Data from Clara County Public H.D. Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013). Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce. Data from USA Today.

"Environment." The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

"Search for Datasets." Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

4 | DISCRETE RANDOM VARIABLES

Figure 4.1 You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Introduction

Chapter Objectives

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

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.
- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** is a numerical variable that describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment. Thus, a random variable is a variable whose values are determined by chance.

Random Variable Notation

We will use upper case letters such as *X* or *Y* to 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 described in words, and the value *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*}. So the values of *X* are *x* = 0, 1, 2, 3. 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.

4.1 | Probability Distribution Function (PDF) for a Discrete Random Variable

If we have a random variable that has only finitely many outcomes, then we can make a table that shows the values *x* in one column and the corresponding probabilities in another column. Such a table is called a *probability distribution function* (PDF), and is very similar to a relatively frequency distribution, like those we saw in Chapters 2 and 3. In particular, a discrete probability distribution function has two key characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

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 per week a newborn baby's crying wakes its mother after midnight.
For this example, *x* = 0, 1, 2, 3, 4, 5. Let *P*(*x*) = probability that *X* takes on a value *x*. Then the following is the probability distribution for *X* :

x	P(*x*)
0	2/50
1	11/50
2	23/50
3	9/50
4	4/50
5	1/50

This is a discrete distribution function because:

a. Each $P(x)$ is between zero and one, inclusive.

b. The sum of the probabilities is one, that is, $\frac{2}{50}+\frac{11}{50}+\frac{23}{50}+\frac{9}{50}+\frac{4}{50}+\frac{1}{50}=1$

Try It Σ

4.1 A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, the possible values are x = 0, 1, 2, 3, 4, 5. $P(x)$ = the probability that X takes on value x. For a random sample of 50 patients, the following information was obtained:

x	P(x)
0	4/50
1	8/50
2	16/50
3	14/50
4	6/50
5	2/50

Why is this a discrete probability distribution function?

These two examples show that we can think of a relative frequency distribution as an approximation to a probability distribution. But using ideas from Chapter 3, we can calculate the actual PDF for many probability experiments.

Example 4.2

Suppose that we toss a coin 4 times in succession. Then we can write the sample space as follows:

TTTT, HTTT, THTT, TTHT, TTTH, HHTT, THHT, TTHH,
THTH, HTHT, HTTH, THHH, HTHH, HHTH, HHHT, HHHH**.**

Let X = the number of heads in four tosses; so the values of X are x = 0, 1, 2, 3, 4.
Find the PDF for the random variable X.

Solution to 4.2:

X	**P(x)**
0	$\frac{1}{16}$
1	$\frac{4}{16}$
2	$\frac{6}{16}$
3	$\frac{4}{16}$
4	$\frac{1}{16}$

Try It Σ

4.2 Suppose that a fair coin is tossed three times in succession. List the outcomes in the sample space as sequences of H's and T's. Let X = the number of heads in three tosses. Construct the PDF for X.

Example 4.3

Suppose that we toss a pair of fair dice. Then we can represent the sample space using the table:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1,6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 5)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Let X = number of dots facing up in a single toss. Find the PDF for the random variable X.

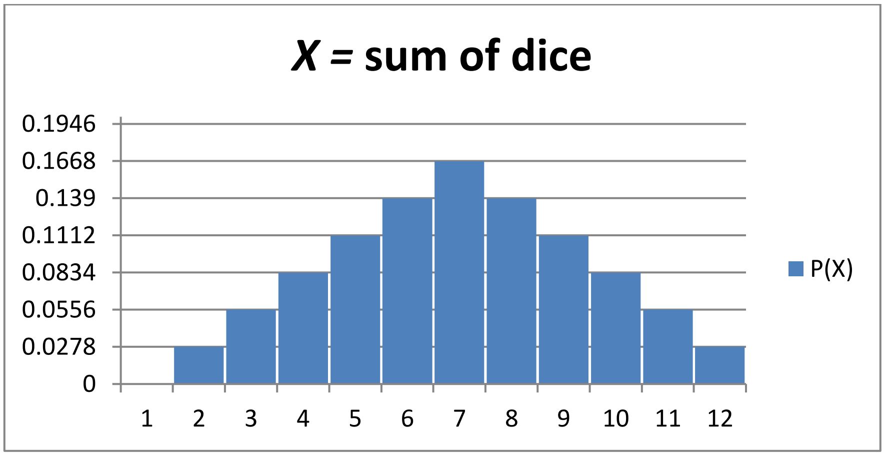
Solution to 4.3:

The possible values of X are $x = 2, 3, \ldots, 12$.

Using the sample space above, we see that $P(x=2)=\frac{1}{36}, P(x=3)=\frac{2}{36}$, etc. and we get the table:

x	2	3	4	5	6	7	8	9	10	11	12
P(*x*)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Finally, we can represent a PDF graphically using a histogram. For example, the histogram for the PDF obtained by rolling two fair dice would be:



4.2 | Mean, Expected Value and Standard Deviation

The **expected value** or **mean** of a random variable can be thought of as the "long-term" average of the variable. That is, if we performed the probability experiment over and over, we would expect the average of the numerical values to be this amount.

For example, suppose that we toss a fair coin and record the result. Although the probability of getting heads is 1/2, this does not mean that in multiple trials that exactly half the tosses would land as heads. I.e. if you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As we learned in Chapter 3, probability does not describe the short-term results of an experiment. Instead, it gives information about what can be expected in the long term. To demonstrate this principle, the British statistician Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together). When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This "long-term average" is known as the **mean** or **expected value** of the experiment and is denoted by the Greek letter μ. In other words, after conducting many trials of an experiment, you would expect this average value.

As we showed in Chapter 2, we can find the mean of a relative frequency distribution by multiplying each data value by its relative frequency, and adding all of these products up. Given the close relationship of relative frequency distributions and probability distributions, it is not surprising that the same basic procedure:

Mean/Expected Value of a Discrete Random Variable

Let *X* be random variable, and $x_1, x_2, \ldots x_n$, the list of possible outcomes for *X*.
Then the *mean* of the distribution and *expected value* of *X* are the same quantity, given by:

$$\mu = \mathbf{E}(X) = \Sigma\, x_i \, \mathbf{P}(x_i)$$

That is, to find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

Example 4.4

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.

Solution to 4.4:

To do the problem, first let the random variable *X* = the number of days the men's soccer team plays soccer per week. Then X takes on the values 0, 1, 2. Construct a PDF table, and add an extra column labeled x*P(x). In this column, you will multiply each x value by its probability:

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

This table is called an expected value table, and helps you calculate the expected value or long-term average. In this case, we have

$$E(X) = \mu = 0*(0.2) + 1*(0.5) + 2*(0.3) = 0 + 0.5 + 0.6 = \mathbf{1.1}.$$

This means that the men's soccer team would, on average, expect to play 1.1 days per week.

We can also calculate the variance and standard deviation of a random variable; again this will be similar to the way we found these statistics for a relative frequency distribution in Chapter 2. That is, we will take a weighted average of the squared deviations from the mean.

Variance and Standard Deviation of a Discrete Random Variable

Let X be random variable, and $x_1, x_2, \ldots x_n$, the list of possible outcomes for X.

- The *variance* of the random variable is: $\sigma^2 = \Sigma(x_i - \mu)^2\, P(x_i)$
- The *standard deviation* is the square root of the variance: $\sigma = \sqrt{\sum (x_i - \mu)^2 P(x_i)}$

Example 4.5

Refer to Example 4.1 Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

Solution to 4.5:

We start with the PDF from Example 4.1; then add columns column showing xP(x), and $(x - \mu)^2$P(x):

x	P(x)	xP(x)	$(x - \mu)^2$P(x)
0	0.04	0	$(0 - 2.1)^2(0.04) = 0.1764$
1	0.22	0.22	$(1 - 2.1)^2(0.22) = 0.2662$
2	0.46	0.92	$(2 - 2.1)^2(0.46) = 0.0046$
3	0.18	0.54	$(3 - 2.1)^2(0.18) = 0.1458$
4	0.08	0.32	$(4 - 2.1)^2(0.08) = 0.2888$
5	0.02	0.10	$(5 - 2.1)^2(0.02) = 0.1682$

Summing the entries in the third column, we get $\mu = E(X) = 2.1$. So on average, we expect a newborn to wake its mother after midnight 2.1 times per week. We then use this value to complete the fourth column. Add the values in the fourth column of the table to get the variance:

$$\sigma^2 = 0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$$

The standard deviation of X is the square root of this sum: $\sigma \approx 1.0247$.

While doing an example using these tables is instructive, we can do these calculations quickly and efficiently using the TI-84. The procedure is exactly as we did for calculating the mean and standard deviation of a frequency distribution in Chapter 2:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the mean and standard deviation of a discrete probability distribution

Go to STAT >> EDIT and clear lists L1 and L2.
Enter the x values into L1, and enter the probabilities P(x) into L2.
Then go to STAT >> CALC. Select 1-VarStats, type L1, L2 and hit ENTER.

The mean will appear at the top of the screen as $\bar{x}$.
The standard deviation will appear as σx.

Try It Σ

4.5 A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained.

x	P(x)
0	0.08
1	0.16
2	0.32
3	0.28
4	0.12
5	0.04

Find the expected value and the standard deviation of this random variable.

Example 4.6

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay $2 to play and could profit $100,000 if you match all five numbers in order (you get your $2 back plus $100,000). Over the long term, what is your **expected** profit of playing the game?

Solution 4.6 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 possible values of x are 100,000 dollars and −2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. We are sampling without replacement so the selections are independent of one another; thus the probability of choosing all five correctly is $\left(\frac{1}{10}\right)^{10}$ =0.00001. Therefore, the probability of winning is 0.00001 and the probability of losing is

$1 - 0.00001 = 0.99999$. Thus we have the PDF:

	x	P(x)
Win	100,000	0.00001
Loss	-2	0.99999

From the calculator, we have μ = Expected Value = **-0.99998**.

Since –0.99998 is about –1, on average you should expect to lose approximately $1 for each game you play. Note that we will never actually lose $1; the only potential outcomes are winning $100,000 or losing $2. But if we play the game many, many times, the average loss will be about $1 per game.

Try It Σ

4.5 You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

Example 4.7

Toss a pair of fair, six-sided dice. Recall from Example 4.3 that the sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1,6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 5)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Let X = the number of faces that show an even number.
Calculate the mean μ and standard deviation σ of X.

Solution to 4.7:

Using the sample space, we have the following PDF:

X	**P(*x*)**
0	$\frac{9}{36}$
1	$\frac{18}{36}$
2	$\frac{9}{36}$

From the calculator, we get $\mu = 1$, and $\sigma = 0.7071$.

Example 4.8

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20. Let X = the amount of profit from a bet.

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X?

Solution to 4.8:

We are given that P(win) = P(one moderate earthquake will occur) = .2142.
Thus, we also have P(loss) = P(one moderate earthquake will *not* occur) = 1 – .2142 = .7858.

And if we win, the we win \$50, so x = 50. If we lose, we pay \$20, so x = -20.
Thus we have the PDF:

	x	P(x)
Win	50	0.2142
Loss	-20	0.7859

From the calculator, we have:

μ = Expected Value = -5.006, and the standard deviation is $\sigma = 28.72$.

If you make this bet many times under the same conditions, the long term outcome will be an average *loss* of $5.01 per bet.

The standard deviation is $\sigma \approx 28.7186$.

Try It Σ

4.8 On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. As in **Example 4.8**, you bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win $100. If you lose the bet, you pay $10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X.

So far we have presented discrete probability distribution functions in table form. But for some distributions, we are actually able to come up with a formula that describes the probabilities in the distribution. In the next few sections, we will investigate some well-known discrete probability functions: the binomial, geometric, and Poisson distributions.

A probability distribution function is a pattern. Each distribution has its own special characteristics. Learning the characteristics each of these enables us to distinguish among the different distributions, and to match a given probability problem into the correct pattern.

4.3 | Binomial Distribution

Binomial distribution is a special discrete probability distribution. There are **four conditions** that the experiment has to meet to be considered a **binomial experiment:**

Conditions for a Binomial Experiment

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.

2. There are only two possible outcomes, called "success" and "failure," for each trial.

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.

4. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial, so $p + q = 1$. Since the trials are independent, p remains the same for each trial.

Any experiment that has characteristics two, three, four, 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 4.9

Randomly guessing at a multiple choice question with 4 possible answers has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose there are 6 multiple choice questions. Joe guesses on each question with no pattern. Determine if this is a binomial experiment.

Solution to 4.9

Condition 1 is met since there are a fixed number of trials: $n = 6$ questions

Condition 2 is met since there are two outcomes: each question is either correct or incorrect

Condition 3 is met since each trial is independent: Joe guessing with no pattern on each question doesn't help in predicting the outcome of another question.

Condition 4 is met since the probability of success (correct) is constant for each trial: $p = .25$ is the probability of guessing a question correct and remains constant for each question.

Example 4.10

Randomly guessing at a multiple choice question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose there are 6 multiple choice questions. The first 3 questions have 4 choices and the last 3 questions have 5 choices each. Joe guesses on each question with no pattern. Determine if this is a binomial experiment.

Solution to 4.10

Condition 1 is met since there are a fixed number of trials: n = 6 questions.

Condition 2 is met since there are two outcomes: each question is either correct or incorrect

Condition 3 is met since each trial is independent: Joe guessing with no pattern on each question doesn't help in predicting the outcome of another question.

Condition 4 is not met since the probability of success (correct) is not constant for each trial: p = .25 (1 out of the 4 choices is correct) is the probability of guessing a question correct for the first 3 question but changes to p = .20 (1 out of 5 choices is correct) for the last 3 questions. The probability of success (correct) does not remain constant for each trial.

Try It Σ

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

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. To find the probability of a specific value of the random variable, the **binomial formula** is used.

Notation and Properties of the Binomial Distribution

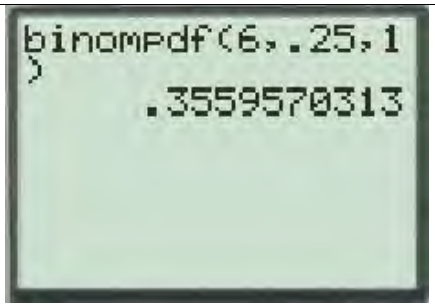
- $X \sim B(n, p)$ This notation states the random variable X is a binomial distribution with n trials and the probability of success , p.

- $P(x = k) = {}_nC_k \cdot p^k \cdot q^{n-k}$ Binomial formula
- The **mean** of a binomial probability distribution, $\mu = n \cdot p$, and
- **Variance** of a binomial probability distribution, $\sigma^2 = n \cdot p \cdot q$.
- The **standard deviation**, $\sigma = \sqrt{npq}$

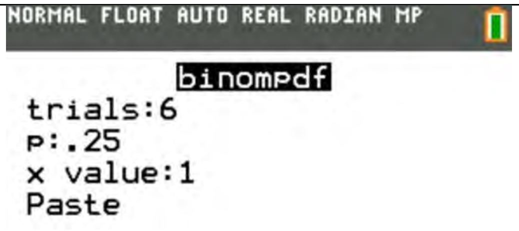
Using the TI-83, 83+, 84, 84+ Calculator

To calculate binomial formula

Go into 2nd DISTR: $P(x = k)$ = binompdf(n, p, k)



```
binompdf(6,.25,1
)
         .3559570313
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
        binompdf
 trials:6
 p:.25
 x value:1
 Paste
```

The above screenshots are for the following situation:

X~B(6, .25) which means that the random variable is Binomial with $n = 6$, $p = .25$, and the formula finds $P(x = 1) = {}_6C_1 \cdot .25^1 \cdot .75^5 = .356$

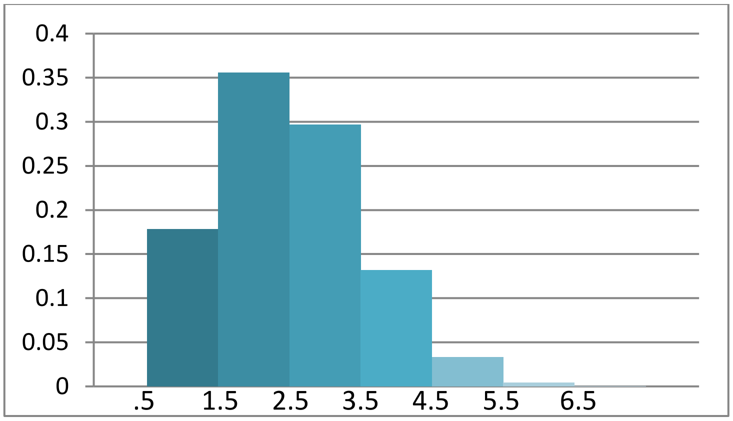
Example 4.11

Randomly guessing at a multiple choice question with 4 possible answers has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose there are 6 multiple choice questions. Joe guesses on each question with no pattern. Create a binomial probability distribution. Draw a histogram.

Solution to 4.11

$n = 6$, $p = .25$ which means $q = .75$

X	P(X)
0	Binompdf(6, .25, **0**) = .178
1	Binompdf(6, .25, **1**) = .356
2	Binompdf(6, .25, **2**) = .297
3	Binompdf(6, .25, **3**) = .132
4	Binompdf(6, .25, **4**) = .033
5	Binompdf(6, .25, **5**) = .00439
6	Binompdf(6, .25, **6**) = .00024



Example 4.12

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.

a.) If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.
b.) Find the mean number of wins.
c.) Find the standard deviation of wins.

Solution to 4.12

a. Let *X* be the number of wins (0, 1, 2, 3, ..., 20). The probability of a success is $p = 0.55$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15) = \text{Binompdf}(20, .55, 15) = .0365$

b. $\mu = n{\cdot}p = 20{\cdot}.55 = 11$ wins is the mean number of wins for 20 trials.

c. $\sigma = \sqrt{npq} = 2.22$

Example 4.13

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 10 attempts, you want to find the probability that the dolphin succeeds at most 5 times. State the probability question mathematically.

Solution to 4.13

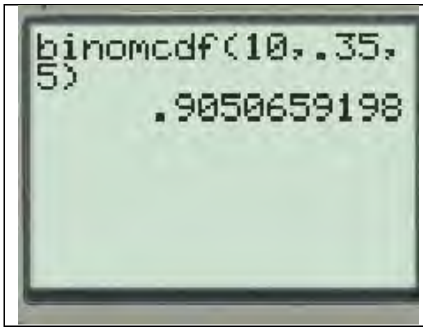
Here, if you define X as the number of successful performances, then X takes on the values 0, 1, 2, 3, ..., 10. The probability of success is p = .35. The probability question can be stated mathematically as $P(x \leq 5)$. Here we want to add all the probabilities from 0 to 5 as seen in the table below:

X	P(X)
0	Binompdf(10, .35, **0**) = .0135
1	Binompdf(10, .35, **1**) = .0725
2	Binompdf(10, .35, **2**) = .176
3	Binompdf(10, .35, **3**) = .252
4	Binompdf(10, .35, **4**) = .238
5	Binompdf(10, .35, **5**) = .154
6	Binompdf(10, .35, **6**) = .0689
7	Binompdf(10, .35, **7**) = .0212
8	Binompdf(10, .35, **8**) = .00428
9	Binompdf(10, .35, **9**) = .00051
10	Binompdf(10, .35, **10**) = .000028

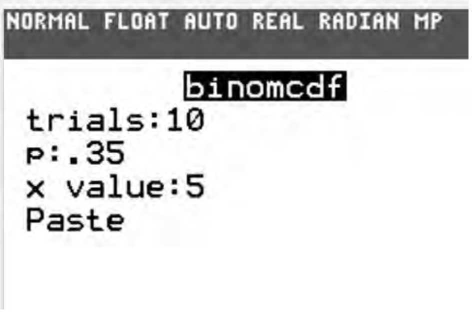
$P(x \leq 5) = .0135 + .0725 + .176 + .252 + .238 = .905$

Using the TI-83, 83+, 84, 84+ Calculator

A simpler way of finding this same probability is to use the cumulative binomial function on the calculator. The function **binomcdf (*n*, *p*, *x*)** calculates the cumulative probability. In this case the calculator automatically sums up the probabilities from $x = 0$ to $x = 5$.



```
binomcdf(10,.35,
5)
      .9050659198
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
          binomcdf
trials:10
p:.35
x value:5
Paste
```

Example 4.14

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically. Find the probability.

Solution to 4.14

$\text{P}(x > 10) = \text{P}(x = 11) + \text{P}(x = 12) + \text{P}(x = 13) + \text{P}(x = 14) + \text{P}(x = 15)$

$= 1 – \text{P(complement)}$
$= 1 – \text{P}(x \leq 10)$
$= 1 – \text{binomcdf}(15, .5, 10)$
$= 1 - .941 = .059$

Try It Σ

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically. Find the probability.

Example 4.15

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? Create a histogram.

Solution to 4.15

Let X = the number of workers who have a high school diploma but do not pursue any further education. X takes on the values 0, 1, 2, ..., 20 where $n = 20$, $p = 0.41$, and $q = 1 – 0.41 = 0.59$.

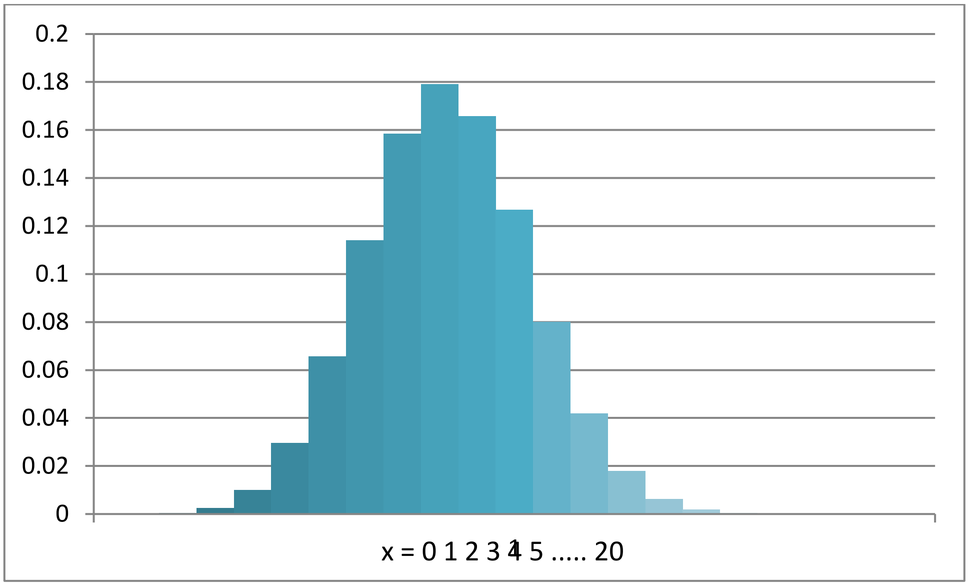
$X \sim \text{B}(20, 0.41)$
$\text{P}(x \leq 12) = \text{binomcdf}(20, .41, 12) = 0.9738.$

Recap

If you want to find $\text{P}(x = \text{k})$, which is the probability of an exact value, use binompdf(n, p, k).
If you want to find $\text{P}(x \leq \text{k})$, use binomcdf($n$, p, k)
If you want to find $\text{P}(x < \text{k})$, use binomcdf(n, p, k - 1)
If you want to find $\text{P}(x > \text{k})$,use 1 – P(Complement) = 1 –$\text{P}(x \leq \text{k})$ = 1 - binomcdf(n, p, k).
If you want to find $\text{P}(x \geq \text{k})$, use 1 – P(Complement) = 1 –$\text{P}(x \leq \text{k} - 1)$ = 1 - binomcdf(n, p, k - 1).

To answer the question how many adult workers do you expect to have a high school diploma but do not pursue any further education:

$\mu = n{\cdot}p = 20(.41) = 8.2$ adult workers.



Try It Σ

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Example 4.16

In the 2013 Jerry's Artarama art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution?
c. Find the following probabilities:

i. the probability that two pages feature signature artists
ii. the probability that at most six pages feature signature artists
iii. the probability that more than three pages feature signature artists.

d. Using the formulas, calculate the mean and standard deviation.

Solution to 4.16

a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
b. $X \sim B(100, 8/560)$
c. probabilities:

i. P($x = 2$) = binompdf(100, 8/560, 2) = .2466
ii. P($x \leq 6$) = binomcdf(100, 8/560, 6) = .9994
iii. P($x > 3$) = 1 – P(x ≤ 3) = 1 – binomcdf(100, 8/560, 3) = .0557

d. $\mu = n \cdot p = 100(8/560) \approx 1.43$ pages

$$\sigma = \sqrt{npq} = \sqrt{100\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.19$$

Try It Σ

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?

b. Use your calculator to find the following probabilities:

   i. the probability that 25 adults in the sample prefer saving over spending

   ii the probability that at most 20 adults prefer saving

   iii. the probability that more than 30 adults prefer saving

c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

Example 4.17

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

a. What is the probability distribution for X?

b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

c. Use your calculator to find the probability that at most eight people develop pancreatic cancer

d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Solution to 4.17

a. $X \sim B(200, 0.0128)$

b. Mean and standard deviation

   i. $\mu = np = 200(0.0128) = 2.56$

   ii. $\sigma = \sqrt{npq} = (200)(0.0128)(0.9853) \approx 1.5897$

c. $P(x \leq 8) = \text{binomcdf}(200, 0.0128, 8) = 0.9988$

d. $P(x = 5) = \text{binompdf}(200, 0.0128, 5) = 0.0707$

e. $P(x = 6) = \text{binompdf}(200, 0.0128, 6) = 0.0298$ So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Try It Σ

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

Example 4.18

The following example illustrates a problem that is not binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn without replacement. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is 6/16. The probability of a student on the second draw is 5/15, when the first draw selects a student. The probability is 6/15, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Try It Σ

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

4.4 | The Geometric Distribution

There are four main characteristics of a geometric experiment:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, we keep repeating the experiment until the first success. Then we stop.

2. The trials are identical and independent.

3. The probability of a success, p, is the same for each trial. Thus, the probability of a failure, q, the same for each trial. Of course, $q = 1 - p$.

4. The random variable is X = the number of independent trials until the first success.

For example, you throw a dart at a bullseye until you hit the bullseye. The first time you hit the bullseye is a "success" so you stop throwing the dart. It might take six tries until you hit the bullseye. You can think of the trials as failure, failure, failure, failure, failure, success, STOP. There must be at least one trial; in theory, the number of trials could go on forever.

As another example, suppose we roll one fair die; then the probability of rolling a three is $\left(\frac{1}{6}\right)$. This is true no matter how many times you roll the die. Suppose you want to know the probability of getting the first three on the fifth roll. On rolls one through four, you do not get a three. The probability for each of the rolls is $q = \left(\frac{5}{6}\right)$. And since the trials are independent, the probability of getting a three on the fifth roll is $\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = 0.0804$.

Example 4.19

You play a game of chance that you can either win or lose until you lose. Your probability of losing is $p = 0.57$. What is the probability that it takes five games until you lose? Let X = the number of games you play until you lose (includes the losing game). Then X takes on the values 1, 2, 3, ... (could go on indefinitely). The probability question is $P(x = 5)$.

Solution to 4.19

The probability of losing is $p = 0.57$, so the probability of winning is $q = 1 – 0.57 = 0.43$.
If the first loss occurs on the fifth play, then the first four plays must be wins. Thus,

$$P(x = 5) = (0.43)(0.43)(0.43)(0.43)(0.57) = \mathbf{0.0195}.$$

Try It Σ

4.19 You throw darts at a board until you hit the center area. Your probability of hitting the center area is $p = 0.17$. Find the probability that it takes eight throws until you hit the center.

Example 4.20

A safety engineer feels that 25% of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) until she finds one that shows an accident caused by failure of employees to follow instructions. On average, how many reports would the safety engineer expect to look at until she finds a report showing an accident caused by employee failure to follow instructions? What is the probability that the safety engineer will have to examine at least three reports until she finds a report showing an accident caused by employee failure to follow instructions?

Solution to 4.20

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 us to find the **expected value** or the mean. If the probability of success is 0.25, then on average we would expect to check about 4 reports before we had a success. I.e. the mean is $\mu = 4 = \dfrac{1}{0.25}$.

The second question asks us to find $P(x \geq 3)$. (At least three means *three or more.*) We can calculate this using the complement rule:

$$P(x \geq 3) = 1 - P(x \leq 2) = 1 - P(x = 2) - P(x = 1)$$

$$= 1 - (0.75)(0.25) - 0.25 = \mathbf{0.5625}.$$

The calculations for the geometric distribution can be done very easily using the calculator.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate probabilities involving Geometric Distribution

Go into 2nd DISTR. The syntax is as follows:

To calculate $P(X = x)$: geometpdf(p, x)

To calculate $P(X \leq x)$: geometcdf(p, x)

Try It Σ

4.20 An instructor feels that 15% of students get below a C on their final exam. She decides to look at final exams (selected randomly and replaced in the pile after reading) until she finds one that shows a grade below a C. We want to know the probability that the instructor will have to examine at most ten exams until she finds one with a grade below a C. Use the calculator to find P($x \leq 10$).

Notation and Properties of the Geometric Distribution

- We write $X \sim G(p)$ to indicate that the random variable X has a geometric distribution.
- The parameter is p = probability of success for a single trial.
- The distribution function is $\mathrm{P}(X = x) = (1-p)^{x-1}\, p$
- The mean of the distribution is $\mu = \dfrac{1}{p}$
- The variance is $\sigma^2 = \left(\dfrac{1}{p}\right)\left(\dfrac{1}{p} - 1\right)$
- The standard deviation is $\sigma = \sqrt{\left(\dfrac{1}{p}\right)\left(\dfrac{1}{p} - 1\right)}$

Example 4.20

Assume that the probability of a defective computer component is 0.02. Components are randomly selected and tested until one component fails.

a. Find the probability that the first defect is caused by the seventh component tested.
b. How many components do you expect to test until one is found to be defective?

Solution to 4.20

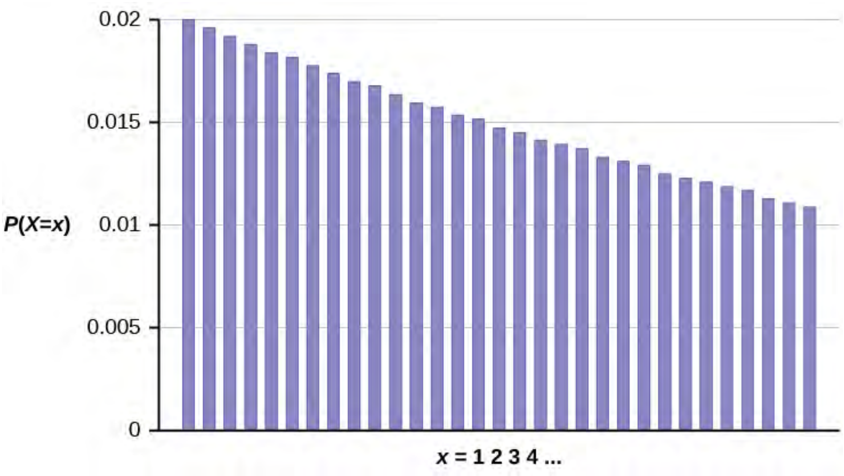
Let X = the number of computer components tested until the first defect is found. Then X takes on the values 1, 2, 3, ... And the probability of a defective component is $p = 0.02$, so $X \sim \mathrm{G}(0.02)$.

a. Here we want the mean: $\mu = \dfrac{1}{p} = \dfrac{1}{.02} = 50$. I.e. we expect that we would have to inspect 50 components before we found a defective one.

b. Here we want to find $P(x = 7)$. Go to 2nd DISTR, and select the geometpdf :

$P(x = 7)$ = geometpdf(.02, 7) = **0.0177**.

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



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

Try It Σ

4.20 The probability of a defective steel rod is 0.01. Steel rods are selected at random. Find the probability that the first defect occurs on the ninth steel rod. Use the TI-83+ or TI-84 calculator to find the answer.

Example 4.21

The lifetime risk of developing pancreatic cancer is about one in78 (1.28%). Let X = the number of people you ask until one says he or she has pancreatic cancer. Then X is a discrete random variable with a geometric distribution: $X \sim G(0.0128)$.

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

Solution to 4.21

a. P(x = 10) = geometpdf(0.0128, 10) = **0.0114**

b. $P(x = 20)$ = geometpdf(0.0128, 20) = **0.01**

c. The mean is $\mu = \frac{1}{p} = \frac{1}{.0128} = \mathbf{78}$

The standard Deviation is $\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{78(78-1)} = \mathbf{77.6234}$.

Try It Σ

4.21 The literacy rate for a nation measures the proportion of people age 15 and over who can read and write. The literacy rate for women in Afghanistan is 12%. Let *X* = the number of Afghani women you ask until one says that she is literate.

a. What is the probability distribution of *X*?
b. What is the probability that you ask five women before one says she is literate?
c. What is the probability that you must ask ten women?
d. Find the mean and standard deviation of *X*.

4.5 | Poisson Distribution

Both Binomial and Poisson are discrete probability distributions. In Binomial the goal was to look for the probability of a specific value of success in *n* trials. Now we want to look for the specific number of occurrences in a specific amount of time or space.

Conditions of a Poisson Experiment.

1. The experiment consists of counting the 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.

2. The probability of the event remains constant for each interval of equal length.

3. The number of occurrences in one fixed interval is independent of the number of occurrences in other fixed intervals.

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

For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.

Notation and Properties for the Poisson Distribution

- $X \sim \text{P}(\mu)$ where μ is the mean number of occurrences per fixed interval.
- $\mu = np$
- $\sigma = \sqrt{\mu}$
- The probability of exactly x occurrences in an interval is

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Where ***e*** is an irrational number approximately 2.718

Example 4.22

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the average number of loaves of bread put on a shelf in 5 minutes?

Solution to 4.22

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

Try It Σ

4.22 The average number of fish caught in an hour is eight. Of interest is the number of fish caught in 15 minutes. The time interval of interest is 15 minutes. What is the average number of fish caught in 15 minutes?

Using the TI-83, 83+, 84, 84+ Calculator

To calculate poisson distribution

Press 2nd VARS, Arrow down to poissonpdf(. Press ENTER

P(x) = poissonpdf(λ, x); P($X < x$) = poisson**c**df(λ, x)

Example 4.23

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in 5 minutes is three?

Solution to 4.23

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

From the previous example, we found μ = 2 for the average number of loaves to put on the shelf in 5 minutes.

P(x = 3) = poissonpdf(2, 3) = .180

Example 4.24

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

Solution to 4.24

X ~ P(6). We want to find P(x < 5) = P($x \leq 4$) = poisssoncdf(6, 4) = **0.2851**.

Try It Σ

4.24 An electronics store expects to have ten returns per day on average. The manager wants to know the probability of the store getting fewer than eight returns on any given day. State the probability question mathematically.

Example 4.25

You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

a. What is the interval of interest?

b. What is the average number of times the news reporter says "uh" during one broadcast?

c. What does X represent? Write the correct notation for Poisson distribution

d. Write a mathematical statement for the probability question.

e. Find the probability that the news reporter says "uh" more than two times per broadcast.

Solution to 4.25

a. one broadcast is the fixed interval

b. $\mu = 2$

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

$x = 0, 1, 2, 3, \ldots$

$X \sim P(2)$

d. $P(x > 2)$

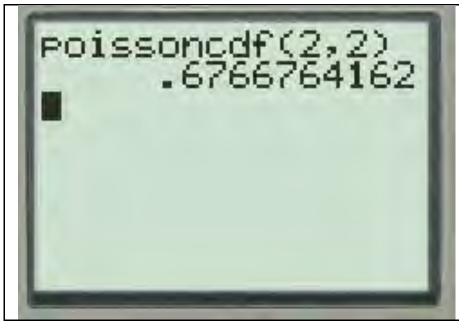
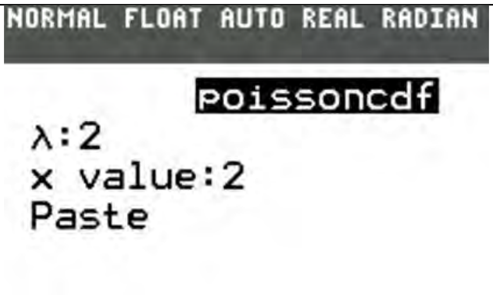
Using the TI-83, 83+, 84, 84+ Calculator

2nd VARS, scroll up to poissoncdf(, Press enter

NOTE: similar to binomcdf, poissoncdf(λ, x) gives the cumulative probability from 0 to x.

e. $P(x > 2) = P(x \geq 3) = 1 - P(\text{Complement}) = 1 - P(x \leq 2)$

$= 1 - \text{poissoncdf}(2, 2) = .677$

 poissoncdf(2,2) .6766764162	 NORMAL FLOAT AUTO REAL RADIAN poissoncdf λ:2 x value:2 Paste

Try It Σ

4.25 An emergency room at a particular hospital gets an average of five patients per hour. A doctor wants to know the probability that the ER gets more than five patients per hour. Give the reason why this would be a Poisson distribution.

Example 4.26

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

Let X = the number of calls Leah receives in 15 minutes. (The interval of interest is 15 minutes or ¼ hour)

Solution to 4.26

x = 0, 1, 2, 3, ...

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

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

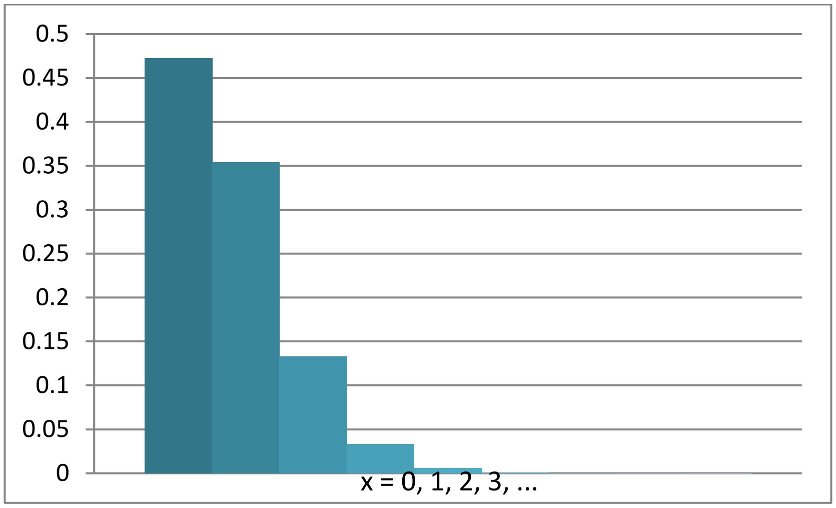
$X \sim P(0.75)$

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

Find $P(x > 1)$.

$P(x > 1) = 1 - P(\text{complement}) = 1 - P(x \leq 1) = 1 - \text{poissoncdf}(.75, 1) = 0.1734$

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



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

Try It Σ

4.26 A customer service center receives about ten emails every half-hour. What is the probability that the customer service center receives more than four emails in the next six minutes? Use the TI-83+ or TI-84 calculator to find the answer.

Example 4.27

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

a. What is the probability that an email user receives exactly 160 emails per day?

b. What is the probability that an email user receives at most 160 emails per day?

c. What is the standard deviation?

Solution to 4.27

a. P(x = 160) = poissonpdf(147, 160) ≈ 0.0180

b. P(x ≤ 160) = poissoncdf(147, 160) ≈ 0.8666

c. Standard Deviation = $\sigma = \mu = 147 \approx 12.1244$

Try It Σ

4.27 According to a recent poll by the Pew Internet Project, girls between the ages of 14 and 17 send an average of 187 text messages each day. Let X = the number of texts that a girl aged 14 to 17 sends per day. The discrete random variable X takes on the values x = 0, 1, 2 …. The random variable X has a Poisson distribution: X ~ P(187). The mean is 187 text messages.

a. What is the probability that a teen girl sends exactly 175 texts per day?
b. What is the probability that a teen girl sends at most 150 texts per day?
c. What is the standard deviation?

Example 4.28

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

a. How many text messages does a text message user receive or send per hour?

b. What is the probability that a text message user receives or sends two messages per hour?

c. What is the probability that a text message user receives or sends more than two messages per hour?

Solution to 4.28:

a. Let X = the number of texts that a user sends or receives in one hour. The average number of texts received per hour is $\frac{41.5}{24} \approx 1.7292$.

b. $X \sim P(1.7292)$, so $P(x = 2) = \text{poissonpdf}(1.7292, 2) \approx 0.2653$

c. $P(x > 2) = 1 - P(x \leq 2) = 1 - \text{poissoncdf}(1.7292, 2) \approx 1 - 0.7495 = 0.2505$

Try It Σ

4.28 Atlanta’s Hartsfield-Jackson International Airport is the busiest airport in the world. On average there are 2,500 arrivals and departures each day.

a. How many airplanes arrive and depart the airport per hour?
b. What is the probability that there are exactly 100 arrivals and departures in one hour?
c. What is the probability that there are at most 100 arrivals and departures in one hour?

Example 4.29

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 24 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution to 4.29

Let X = the number of days with low seismic activity. Using the binomial distribution:

- $P(x = 10)$ = binompdf(200, .0102, 10) ≈ 0.000039

Using the Poisson distribution:

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

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

Try It Σ

On May 13, 2013, starting at 4:30 PM, the probability of moderate seismic activity for the next 48 hours in the Kuril Islands off the coast of Japan was reported at about 1.43%. Use this information for the next 100 days to find the probability that there will be low seismic activity in five of the next 100 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

KEY TERMS

Bernoulli Trials an experiment with the following characteristics:

1. There are only two possible outcomes called "success" and "failure" for each trial.
2. The probability p of a success is the same for any trial.
   So the probability $q = 1 - p$ of a failure is also the same for any trial.

Binomial Experiment a statistical experiment that satisfies the following four conditions:

1. There are a fixed number of identical trials, n.
2. There are only two possible outcomes, called "success" and, "failure," for each trial.
3. The probability p of a success on one trial remains the same for all trials. Similarly, the probability q of a failure on one trial remains the same for all trials.

4 The n trials are independent of one another.

Binomial Distribution a discrete random variable that arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial random variable X is defined as the number of successes in n trials.

The notation is: $X \sim B(n, p)$.

The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$.

The probability of exactly x successes in n trials is $P(X = x) = {}_nC_k\, p^x q^{n-x}$

Expected Value: The expected arithmetic average when an experiment is repeated many times; also called the *mean* of the distribution.

The expected value is denoted by either E(X) or μ. For a discrete random variable with probability distribution function $P(x)$, we calculate the expected value by E(X) = μ = $\sum xP(x)$.

Geometric Experiment a statistical experiment with the following properties:

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

Geometric Distribution a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success.

Notation: $X \sim G(p)$.

The mean is $\mu = \dfrac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\dfrac{1}{p} \cdot \left(\dfrac{1}{p} - 1\right)}$.

The probability of exactly x failures before the first success is given by the formula:

$$P(X = x) = p(1-p)^{x-1}.$$

Mean of a Probability Distribution: The long-term average of the random variable over many trials of a statistical experiment. The mean is also called the *expected value*.

Poisson Probability Distribution: A discrete random variable that counts the number of times a certain event will occur in a specific interval. Characteristics of the variable:

- The probability *p* that the event occurs in a given interval is the same for all intervals of equal length.
- The events occur with a known mean and independently of the time since the last event.
- The distribution is defined by the mean λ of the event in the interval.

  Notation: $X \sim P(\mu)$.

  The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{\mu}$.

  The probability of having exactly *x* successes a fixed time interval is $P(X = x) = e^{-\mu} \dfrac{\mu^x}{x!}$.

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

Probability Distribution Function (PDF): A mathematical description of a discrete random variable, 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.

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,* and *z*. For example, if *X* is the number of children in a family, then *x* represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in algebra in the two following ways.

- The domain of the random variable is not necessarily a numerical set; the domain may be categorical in nature and expressed in words; for example, if *X* = hair color then the domain is {black, blond, gray, green, orange}.

- The values of the random variable are determined by chance. So we can tell what specific value *x* the random variable *X* takes only after performing the experiment.

Standard Deviation of a Probability Distribution: A numerical measure of how dispersed the outcomes of a statistical experiment are from the mean of the distribution.

The Law of Large Numbers: As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

FORMULA REVIEW

Let X be a discrete random variable, and $x_1, x_2, \ldots x_n$, the list of possible outcomes for X.

- The *mean* of the random variable is: $\mu = \Sigma x_i \mathrm{P}(x_i)$
- The *variance* of the random variable is: $\sigma^2 = \Sigma(x_i - \mu)^2 \mathrm{P}(x_i)$
- The *standard deviation* is the square root of the variance: $\sigma = \sqrt{\sum (x_i - \mu)^2 \mathrm{P}(x_i)}$

Binomial Distribution:

Notation: $X \sim B(n, p)$.

Mean: is $\mu = np$

PDF: $P(X = x) = {}_nC_k p^x q^{n-x}$

Standard deviation: $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$

Geometric Distribution:

Notation: $X \sim G(p)$.

Mean: $\mu = \dfrac{1}{p}$

PDF: $P(X = x) = p(1-p)^{x-1}$

Standard deviation: $\sigma = \sqrt{\dfrac{1}{p} \cdot \left(\dfrac{1}{p} - 1\right)}$.

Poisson Distribution:

Notation: $X \sim P(\mu)$.

Mean: $\mu = np$

PDF: $P(X = x) = e^{-\mu} \dfrac{\mu^x}{x!}$

Standard deviation: $\sigma = \sqrt{\mu}$

Exercises for Chapter 4

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let X = the number of years a new hire will stay with the company.
Let $P(x)$ = the probability that a new hire will stay with the company x years.

1. Complete the table using the data provided:

x	P(x)
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

2. $P(x = 4)$ = _________

3. $P(x \geq 5)$ = _________

4. On average, how long would you expect a new hire to stay with the company?

5. What does the column "$P(x)$" sum to?

Use the following information to answer the next four exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

X	P(x)
1	0.15
2	0.35
3	0.40
4	0.10

6. Define the random variable X.

7. What is the probability the baker will sell more than one batch? $P(x > 1)$ =

8. What is the probability the baker will sell exactly one batch? $P(x = 1)$ =

9. On average, how many batches should the baker make?

Use the following information to answer the next two exercises: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

10. Define the random variable *X*.

11. Construct a probability distribution table for the data.

12. We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

13. Define the random variable *X*.

14. What values does *x* take on?

15. Construct a PDF table.

16. Find the probability that Javier volunteers for less than three events each month:
That is, calculate P($x < 3$).

17. Find the probability that Javier volunteers for at least one event each month.
That is, calculate P($x \geq 1$).

18. Complete the expected value table:

x	P(*x*)	*x** P(*x*)
0	0.2	
1	0.2	
2	0.4	
3	0.2	

19. Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in

x	P(*x*)
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. In words, define the random variable *X*.
b. What does it mean that the values zero, one, and two are not included for *x* in the PDF?
c. On average, how many years do you expect it to take for an individual to earn a B.S.?

20. Find the mean and standard deviation of the probability distributions:

a.

x	**P(*x*)**
2	0.1
4	0.3
6	0.3
8	0.2
10	0.1

b.

x	**P(*x*)**
1	0.15
2	0.25
3	0.30
4	0.20
5	0.15

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution

x	P(x)
1	0.35
2	0.20
3	0.15
4	
5	0.10
6	0.05

21. Define the random variable *X*.

22. Define *P*(*x*), or the probability of *x*.

23. Find the probability that a physics major will do post-graduate research for four years, $P(x = 4)$.

24. Find the probability that a physics major will do post-graduate research for at most three years. I.e. find $P(x \leq 3)$.

25. On average, how many years would you expect a physics major to spend doing post-graduate research?

Use the following information to answer the next five exercises: A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Let *X* = the number of years a student will study ballet with the teacher.
Over the years, she has established the following probability distribution:

x	P(x)
1	0.10
2	0.05
3	0.10
4	
5	0.30
6	0.20
7	0.10

26. Complete the table using the data provided.

27. Calculate P($x \geq 5$).

28. Calculate P($x = 4$) =

29. Calculate P($x < 4$) =

30. On average, how many years would you expect a child to study ballet with this teacher

31. What does the column "P(x)" sum to and why?

35. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

36. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

Use the following information to answer the next eight exercises:
The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. Of those surveyed, 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

37. In words, define the random variable X.

38. Identify the distribution of X: $X \sim$ ____ (___,___)

39. What values does the random variable X take on?

40. Construct the probability distribution function (PDF).

41. On average, how many would you expect to answer yes?

42. What is the standard deviation σ?

43. What is the probability that at most five of the freshmen reply "yes"?

44. What is the probability that at least two of the freshmen reply "yes"?

45. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win $6
- If the card is a face card, and the coin lands on Tails, you win $2
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

46. You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

47. A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

48. Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. Let X = the number of children married people have. The results of the survey are compiled and are used as theoretical probabilities:

x	**P(*x*)**
0	0.10
1	0.20
2	0.30
3	
4	0.10
5	0.05
6 or more	0.05

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

49. A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money. b. No, I expect to come out behind in money.
c. It doesn’t matter. I expect to break even.

50. Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable *X* equal the size of the student’s class. Define the PDF for *X*.
c. Find the mean of *X*.
d. Find the standard deviation of *X*.

51. In a lottery, there are 250 prizes of $5, 50 prizes of $25, and ten prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

52. According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed.
Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

53. Define the random variable and list its possible values.

54. State the distribution of *X*.

55. Find the probability that at least four of the 25 patients actually have the flu.

56. On average, for every 25 patients calling in, how many do you expect to have the flu?

57. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the table below. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	0	1	2	3	4	5
P(x)	0.03	0.50	0.24		0.07	0.04

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

58. A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ ____(__,__)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.
f. Find the probability that more than two students will attend.

Use the following information to answer the next two exercises: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.
Let X = the number of games won in that upcoming month.

59. Find the expected number of wins for that upcoming month.

60. What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476 b. 0.2336 c. 0.7664 d. 0.8903

61. What is the probability that the San Jose Sharks win at least five games in the upcoming month?

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

62. A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

63. A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

64. Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(__,__)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one?
Justify your answer numerically.

65. More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(___,___)
d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most ten offer such courses.
f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

66. Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(___,___)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

67. At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(___,___)
d. How many are expected to **not** to use the foil as their main weapon?

e. Find the probability that six do **not** use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

68. Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ ____ (____,____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

69. The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ ____ (____,____)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

70. It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ ____ (____,____)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

71. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ ____ (____,____)
d. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player. g. Determine who has the advantage, the player or the house.

72. According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

a. What is the probability distribution for X?
b. Using the formulas, calculate the mean and standard deviation of X.
c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most ten people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

73. The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.

a. Sketch a graph of the probability distribution of X.
b. Using the formulas, calculate the mean and standard deviation of X.
c. Find the probability that more than five people in the sample are literate. Is it is more likely that three people or four people are literate.

4.4 Geometric Distribution

Use the following information to answer the next three exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly select freshman from the study until you find one who replies “yes.” You are interested in the number of freshmen you must ask.

74. In words, define the random variable X and identify the distribution of X.
75. On average, how many freshmen would you expect to have to ask until you found one who replies "yes?
76. What is the probability that you will need to ask fewer than three freshmen?

77. Suppose that the probability that an adult in America will watch the Super Bowl is 40%. Each person is considered independent. We are interested in the number of adults in America we must survey until we find one who will watch the Super Bowl.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Identify the distribution of *X*. *X* ~ ___(___,___)
d. How many adults in America do you expect to survey until you find one who will watch the Super Bowl?
e. Find the probability that you must ask seven people.
f. Find the probability that you must ask three or four people.

78. It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose we are interested in the number of California residents we must survey until we find a resident who does **not** have adequate earthquake supplies.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(___,___)
d. What is the probability that we must survey just one or two residents until we find a California resident who does not have adequate earthquake supplies?
e. What is the probability that we must survey at least three California residents until we find a California resident who does not have adequate earthquake supplies?
f. How many California residents do you expect to need to survey until you find a California resident who **does not** have adequate earthquake supplies?
g. How many California residents do you expect to need to survey until you find a California resident who **does** have adequate earthquake supplies?

79. In one of its Spring catalogs, L.L. Bean® advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked more than once.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(______,______)
d. How many pages do you expect to advertise footwear on them?
e. Is it probable that all twenty will advertise footwear on them? Why or why not?
f. What is the probability that fewer than ten will advertise footwear on them?
g. Reminder: A page may be picked more than once. We are interested in the number of pages that we must randomly survey until we find one that has footwear advertised on it. Define the random variable *X* and give its distribution.
h. What is the probability that you only need to survey at most three pages in order to find one that advertises footwear on it?
i. How many pages do you expect to need to survey in order to find one that advertises footwear?

80. The World Bank records the prevalence of HIV in countries around the world. According to their data, “Prevalence of HIV refers to the percentage of people ages 15 to 49 who are infected with HIV.”[1] In South Africa, the prevalence of HIV is 17.3%. Let *X* = the number of people you test until you find a person infected with HIV.

a. Sketch a graph of the distribution of the discrete random variable *X*.
b. What is the probability that you must test 30 people to find one with HIV?
c. What is the probability that you must ask ten people?
d. Find the mean and standard deviation of the distribution of *X*.

81. According to a recent Pew Research poll, 75% of millennials (people born between 1981 and 1995) have a profile on a social networking site. Let *X* = the number of millennials you ask until you find a person without a profile on a social networking site.

a. Describe the distribution of *X*.
b. Find the mean and standard deviation of *X*.
c. What is the probability that you must ask ten people to find one person without a social networking site?
d. What is the probability that you must ask 20 people to find one person without a social networking site?
e. What is the probability that you must ask *at most* five people?

82. A consumer looking to buy a used red Miata car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28%. We are interested in the number of dealerships she must call.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(______,______)
d. On average, how many dealerships would we expect her to have to call until she finds one that has the car?
e. Find the probability that she must call at most four dealerships.
f. Find the probability that she must call three or four dealerships.

[1] ”Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value 2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May15, 2013).

4.5 Poisson Distribution

Use the following information to answer the next six exercises: 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.

83. Assume the event occurs independently in any given day. Define the random variable *X*.

84. *X* ~ ____ (___,___)

85. What values does *X* take on?

86. For the given values of the random variable *X*, fill in the corresponding probabilities.

87. Is it likely that there will be no teens killed from motor vehicle injuries on any given day in the U.S? Justify your answer numerically.

88. Is it likely that there will be more than 20 teens killed from motor vehicle injuries on any given day in the U.S.? Justify your answer numerically.

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

89. Assume the event occurs independently in any given day. Define the random variable *X* and identify the distribution.

90. What is the probability of getting 150 customers in one day?

91. What is the probability of getting 35 customers in the first four hours? Assume the store is open 12 hours each day.

92. What is the probability that the store will have more than 12 customers in the first hour?

93. What is the probability that the store will have fewer than 12 customers in the first two hours?

94. The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let *X* = the number of births in an hour.

a. Find the mean and standard deviation of *X*.
b. Sketch a graph of the probability distribution of *X*.
c. What is the probability that the maternity ward will deliver three babies in one hour?
d. What is the probability that the maternity ward will deliver at most three babies in one hour?
e. What is the probability that the maternity ward will deliver more than five babies in one hour?

95. The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let *X* = the number of calls received at noon.

a. Find the mean and standard deviation of *X*.
b. What is the probability that the office receives at most six calls at noon on Monday?
c. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
d. What is the probability that the office receives more than eight calls at noon?

96. The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(______,______)
d. Find the probability that she has no children.
f. Find the probability that she has more children than the Japanese average.

97. The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.

a. In words, define the Random Variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(______,______)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Spanish average.
f. Find the probability that she has more children than the Spanish average.

98. Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____
d. Find the probability that she has no litters in one year.
e. Find the probability that she has at least two litters in one year.
f. Find the probability that she has exactly three litters in one year.

99. The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to answer the following questions.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Identify the distribution of *X*. *X* ~ _(___,___)
d. How many are expected to be audited?
e. Find the probability that no one was audited.
f. Find the probability that at least three were audited.

100. An article about directory assistance stated that "internal surveys paid for by directory assistance providers show that even the most accurate companies give out wrong numbers 15% of the time." Assume that you are testing such a provider by making 20 directory assistance requests and also assume that the provider gives the wrong phone number 15% of the time.

a. Give the distribution of X. $X \sim$ ______(______,______)

b. Find the probability of getting 3 wrong numbers
c. Find the probability of getting at most 7 wrong numbers
d. Find the probability of getting at least 5 wrong numbers

101. A company prices its natural disaster insurance using the following assumptions. In one calendar year, there is at most 1 natural disaster. Let's say the probability of a natural disaster is .15. Lastly, the number of natural disasters in any calendar year is independent of the number of natural disasters in any other calendar year.

a. Give the distribution of X. $X \sim$ ______(______,______)

b. Find the probability that there are fewer than 4 natural disasters in a 10 year period.

c. Find the probability that there are at least 6 natural disasters in a 10 year period.

REFERENCES

4.2 Mean, Expected Value and Standard Deviation

Class Catalogue at the Florida State University. Available online at https://apps.oti.fsu.edu/RegistrarCourseLookup/ SearchFormLegacy (accessed May 15, 2013).

"World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.world- earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

4.3 Binomial Distribution

"Access to electricity (% of population)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/ EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

"Distance Education." Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

"NBA Statistics – 2013," ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. "Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income," GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/americans- enjoy-saving-rather-spending.aspx (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/ TheAmericanFreshman2011.pdf (accessed May 15, 2013).

"The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the- world-factbook/geos/af.html (accessed May 15, 2013).

"What are the key statistics about pancreatic cancer?" American Cancer Society, 2013. Available online at http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).

4.4 Geometric Distribution

"Millennials: A Portrait of Generation Next," PewResearchCenter. Available online at http://www.pewsocialtrends.org/ files/2010/10/millennials-confident-connected-open-to-change.pdf (accessed May 15, 2013).

"Millennials: Confident. Connected. Open to Change." Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-to- change/ (accessed May 15, 2013).

"Prevalence of HIV, total (% of populations ages 15-49)," The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_ value-last&sort =desc (accessed May15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011.* Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/ TheAmericanFreshman2011.pdf (accessed May 15, 2013).

"Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan," The European Union and ICON-Institute. Available online at http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf (accessed May 15, 2013).

"The World FactBook," Central Intelligence Agency. Available online at https://www.cia.gov/library/ publications/the- world-factbook/geos/af.html (accessed May 15, 2013).

"UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills," UNICEF Television. Video available online at http://www.unicefusa.org/assets/video/afghan-female- literacy-centers.html (accessed May 15, 2013).

4.5 Poisson Distribution

"ATL Fact Sheet," Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at http://www.atlanta-airport.com/Airport/ATL/ATL_FactSheet.aspx (accessed May 15, 2013).

Center for Disease Control and Prevention. "Teen Drivers: Fact Sheet," Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html (accessed May 15, 2013).

"Children and Childrearing," Ministry of Health, Labour, and Welfare. Available online at http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html (accessed May 15, 2013).

"Eating Disorder Statistics," South Carolina Department of Mental Health, 2006. Available online at http://www.state.sc.us/ dmh/anorexia/statistics.htm (accessed May 15, 2013).

"Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day," the guardian, 2013. Available online at http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2 (accessed May 15, 2013).

"How Americans Use Text Messaging," Pew Internet, 2013. Available online at http://pewinternet.org/Reports/2011/Cell- Phone-Texting-2011/Main-Report.aspx (accessed May 15, 2013).

Lenhart, Amanda. "Teens, Smartphones & Testing: Texting volume is up while the frequency of voice calling is down. About one in four teens say they own smartphones," Pew Internet, 2012. Available online at http://www.pewinternet.org/~/media/ Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf (accessed May 15, 2013).

"One born every minute: the maternity unit where mothers are THREE to a bed," MailOnline. Available online at http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers- bed.html (accessed May 15, 2013).

Vanderkam, Laura. "Stop Checking Your Email, Now." CNNMoney, 2013. Available online at http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/ (accessed May 15, 2013).

"World Earthquakes: Live Earthquake News and Highlights," World Earthquakes, 2012. http://www.world- earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

5|Continuous Random Variable

Figure 5.1 The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Introduction

Chapter Objectives

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

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

5.1|Continuous Probability Distributions

NOTE

The values of discrete and continuous random variables can be unclear. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, and then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Recall Discrete Probability Distributions have two rules

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

Example 5.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$. Let $P(x)$ = probability that X takes on a value x. Then the following is the probability distribution for X :

x	P(x)
0	2/50
1	11/50
2	23/50
3	9/50
4	4/50
5	1/50

Recall P(x < 2) = $\frac{2}{50} + \frac{11}{50}$

For continuous random variables, the values of X can't be placed in a table. The outcomes are measured, not counted.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated as pdf). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as cdf). The cumulative distribution function is used to evaluate probability as area.

Area under the curve

- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of *x* values rather than for individual *x* values.
- P(c < *x* < d) is the probability that the random variable *X* is in the interval between the values c and d. P(c < *x* < d) is the area under the curve, above the x-axis, to the right of c and the left of d. P(c < *x* < d) is the same as P(c ≤ *x* ≤ d) because probability is equal to area.
- P(*x* = c) = 0. The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between *x* = c and *x* = c has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.

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, the formulas were found by using the techniques of integral calculus. However, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way. In this chapter and the next, we will study the continuous **uniform distribution**, the **exponential distribution**, and the **normal distribution**. The following graphs illustrate these distributions respectively.

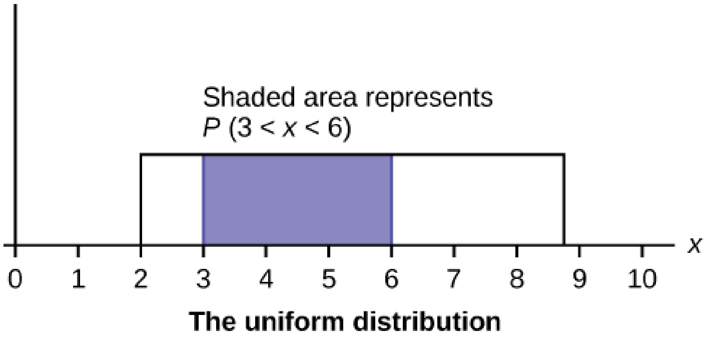


Figure 5.2 Graph shows a Uniform Distribution with the shaded area between *x* = 3 and *x* = 6 to represent the probability that the value of the random variable *X* is in the interval between 3 and 6.

Remember that the x-axis has the values of the continuous random variable.

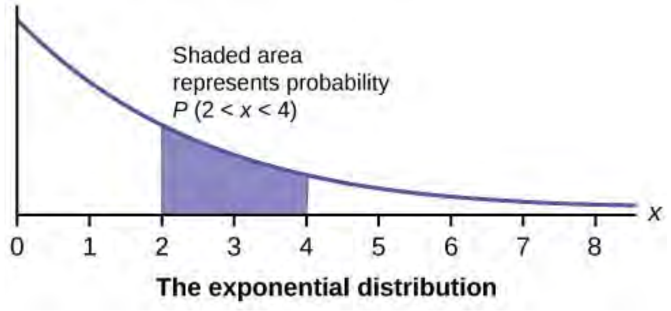


Figure 5.3 Graph above shows an Exponential Distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between 2 and 4.

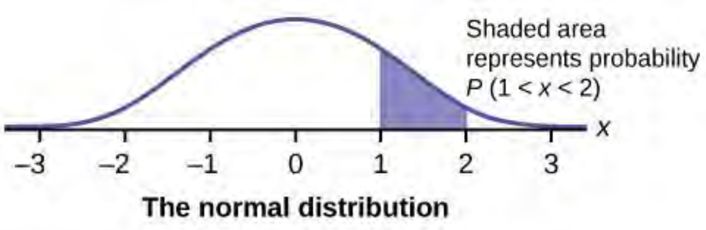


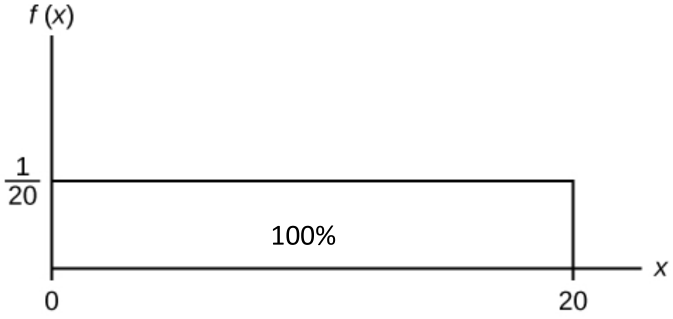
Figure 5.4 Graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between 1 and 2.

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 5.2

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and

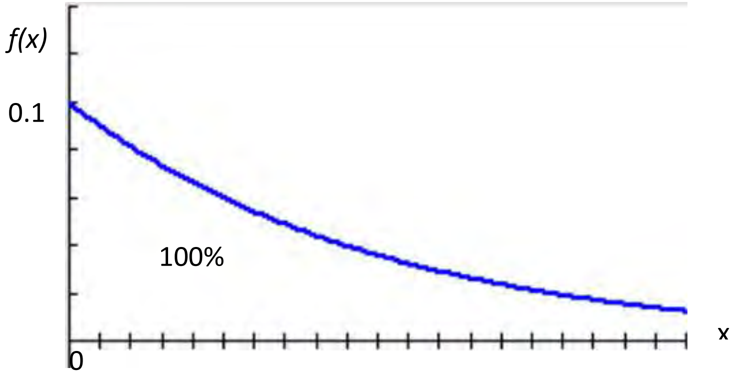
$x = 20$, inclusive.



Example 5.3

Consider the function *f*(*x*) = $.1e^{-.1x}$ for $x \geq 0$. *x* = a real number. The graph of *f*(*x*) = $.1e^{-.1x}$ is an exponential function that decreases starting at *x* = 0.

NOTE: The x-axis continuous to positive infinity on the right hand side. The exponential curve never touches the x-axis.



5.2 | Uniform Distribution

The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive.

Notation and Properties of the Uniform Distribution

- $X \sim \text{U}(a, b)$ where a = lowest value of *x* and b = highest value of *x*.
- The probability density function is $f(x) = \dfrac{1}{b-a}$
- The **mean** of a uniform probability distribution, $\mu = \dfrac{a+b}{2}$, and
- **Variance** of a uniform probability distribution, $\sigma^2 = \dfrac{(b-a)^2}{12}$
- The **standard deviation**, $\sigma = \dfrac{(b-a)}{\sqrt{12}}$
- Probability = Area = Base·height

Example 5.4

Consider the function $f(x) = \frac{1}{20}$ for $0 \le x \le 20$. x = a real number.

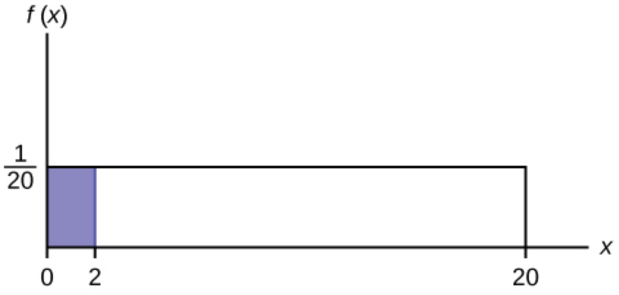
a. State the probability distribution in correct notation.
b. Find the mean and standard deviation.
c. Find $P(x < 2)$.

Solution 5.4

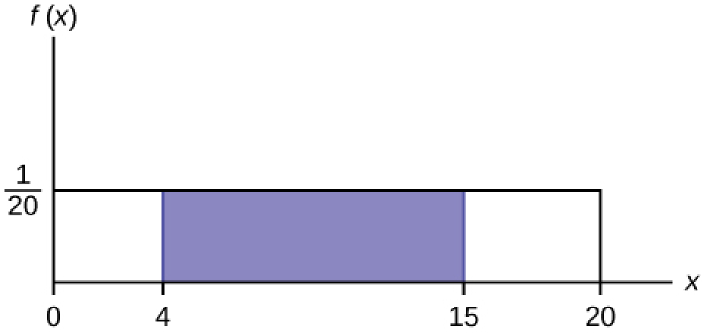
a. Here we have a uniform distribution: U(0, 20) where a = 0 and b = 20

b. $\mu = \frac{a+b}{2} = \frac{0+20}{2} = 10\,;\ \sigma = \frac{(b-a)}{\sqrt{12}} = \frac{20-0}{\sqrt{12}} = 5.77$

c. $P(x < 2)$ = area of rectangle = base·height = $2 \cdot \frac{1}{20} = .1 = 10\%$



Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $4 < x < 15$.

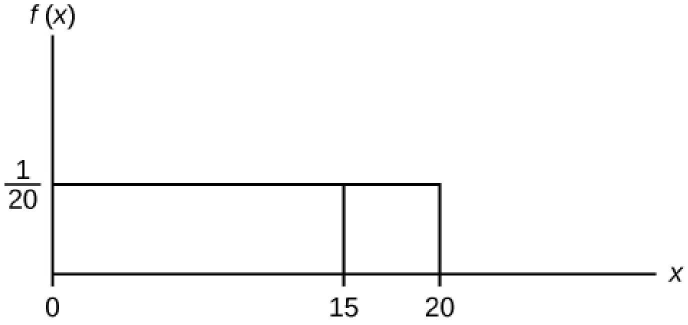


Area = (Base)(Height)

Area = $(15 - 4)\left(\frac{1}{20}\right) = 0.55$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

NOTE: Suppose we want to find P(x = 15). On an x-y graph, x = 15 is a vertical line. A vertical line has no width (or zero width). Therefore, P(x = 15) = (base)(height) = (0) $\left(\frac{1}{20}\right)$ = 0



P($x \leq$ k) (can be written as P($x <$ k) 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($x >$ k). The CDF gives "area to the left" and P($x >$ k) gives "area to the right." We calculate P($x >$ k) for continuous distributions as follows: P($x >$ k) = 1 – P ($x <$ k).

Try It Σ

5.1 The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are equally likely. State the values of a and b. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

1	12	4	10	4	14	11	7	11	4	13	2	4	6	3	10	0	12
6	9	10	5	13	4	10	14	12	11	6	10	11	0	11	13	2	

Example 5.5

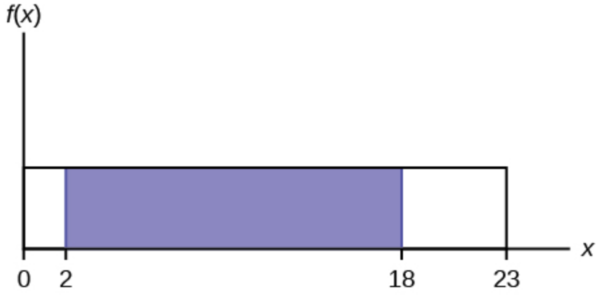
Let the smiling times of an 8-week old baby, in seconds; follow a uniform distribution between 0 and 23 seconds inclusive.

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

b. Find the 90th percentile for an 8-week old baby's smiling time.

c. Find the probability that a random eight-week-old baby smiles more than 12 seconds knowing that the baby smiles more than eight seconds.

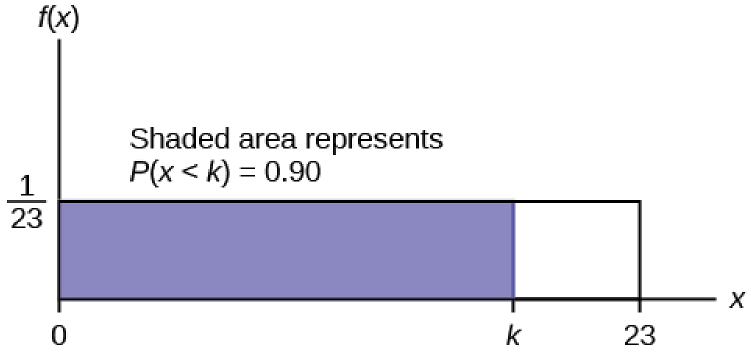
Solution 5.5



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

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

b. 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

$$(k - 0)\left(\frac{1}{23}\right) = 0.90$$

k = 20.7 minutes

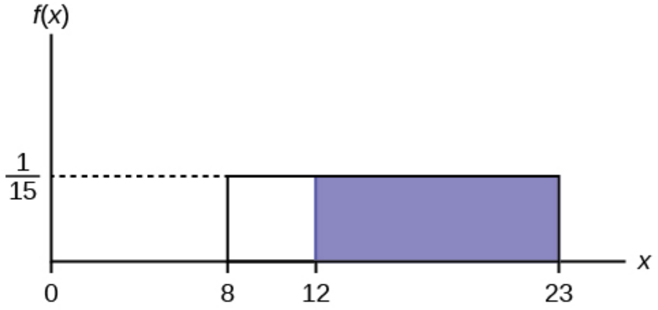
c. This probability question is a **conditional**. You are asked to find the probability that an eight-week-old baby smiles more than 12 seconds when you already know the baby has smiled for more than eight seconds.

Find P(x > 12|x > 8) There are two ways to do the problem:

For the first way, use the fact that this is a conditional and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than eight seconds.

Write a new f(x):

$$f(x) = \frac{1}{23-8} = \frac{1}{15} \text{ for } 8 < x < 23$$



P(x > 12|x > 8) = base·height

$$= (23 - 12)\left(\frac{1}{15}\right) = \frac{11}{15}$$

For the second way, use the conditional formula from Probability Topics with the original distribution $X \sim$ U(0, 23):

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

For this problem, A is (x > 12) and B is (x > 8).

So, $\text{P}(x > 12|x > 8) = \dfrac{P\left(x > 12 \text{ AND } x > 8\right)}{P(x > 8)}$

$$= \frac{P\left(x > 12\right)}{P(x > 8)} = \frac{\left(23-12\right)\left(\frac{1}{15}\right)}{\left(23-8\right)\left(\frac{1}{15}\right)} = \frac{11}{15}$$

Try It Σ

A distribution is given as X ~ U (0, 20). What is P(2 < x < 18)? Find the 90th percentile.

Example 5.6

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

a. What is the probability that a person waits fewer than 12.5 minutes?

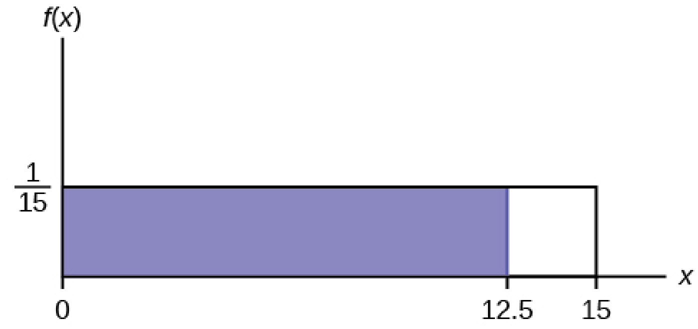
b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

c. Ninety percent of the time, the time a person must wait falls below what value?

Solution 5.6

a. Let X = the number of minutes a person must wait for a bus. a = 0 and b = 15. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \le x \le 15$.

Find P ($x < 12.5$). Draw a graph.



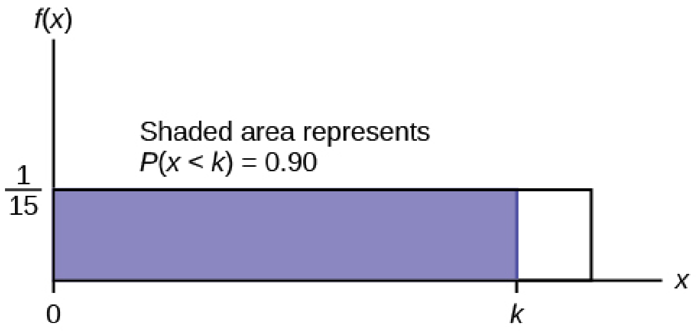
$$P(x < k) = (\text{base})(\text{height}) = (12.5 - 0)\left(\frac{1}{15}\right) = 0.8333$$

The probability a person waits less than 12.5 minutes is 0.8333.

b. $\mu = (a + b)/2 = (15 + 0)/2 = 7.5$ minutes. On the average, a person must wait 7.5 minutes.

$\sigma = \frac{(b-a)}{\sqrt{12}} = \frac{15-0}{\sqrt{12}} = 4.3$. The Standard deviation is 4.3 minutes.

c. This asks for the 90th percentile. Draw a graph. Let k = the 90th percentile.



P(x < k) = (base)(height)

$$0.9 = (\mathrm{k} - 0)\left(\frac{1}{15}\right)$$

$$0.9 = \mathrm{k}\left(\frac{1}{15}\right)$$

k = (0.90)(15) = 13.5 k is sometimes called a **critical value**.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.

Try It Σ

5.4 The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

a. Find *a* and *b* and describe what they represent.
b. Write the distribution.
c. Find the mean and the standard deviation.
d. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
e. What is the 65th percentile for the duration of games for a team for the 2011 season?

Example 5.7

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).

a. What is the probability that a randomly selected nine-year old child eats a donut in at least two minutes?
b. Find the probability that a different nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes.

Solution 5.7

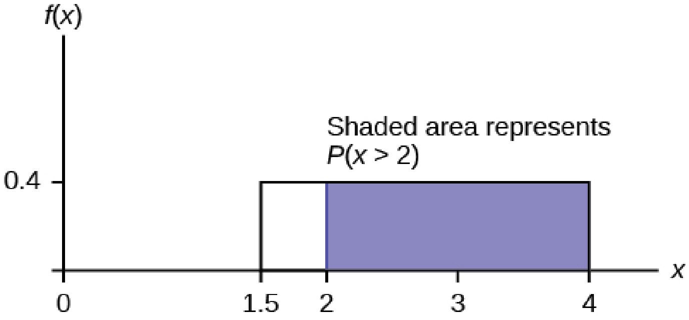
a. 0.5714

b. The second question has a conditional probability. You are asked to find the probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes. First way: Since you know the child has already been eating the donut for more than 1.5 minutes, you are no longer starting at *a* = 0.5 minutes. Your starting point is 1.5 minutes.

Write a new *f*(*x*):

$$f(x) = \frac{1}{4-1.5} = \frac{1}{2.5} = .4 \text{ for } 1.5 \le x \le 4.$$

Find P(*x* > 2|*x* > 1.5). Draw a graph.



$$P(x > 2|x > 1.5) = \text{(base)(new height)} = (4 - 2)(.4) = .8$$

The probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes is .8 = 80%.

Second way: Draw the original graph for *X* ~ U (0.5, 4). Use the conditional formula

$$P(x > 2|x > 1.5) = \frac{P(x>2 \ \ and \ \ x>1.5)}{P(x>1.5)} = \frac{2\left(\frac{1}{3.5}\right)}{2.5\left(\frac{1}{3.5}\right)} = .8$$

Try It Σ

Suppose the time it takes a student to finish a quiz is uniformly distributed between six and 15 minutes, inclusive. Let *X* = the time, in minutes, it takes a student to finish a quiz. *X* ~ U (6, 15). Find the probability that a randomly selected student needs at least eight minutes to complete the quiz. Then find the probability that a different student needs at least eight minutes to finish the quiz given that she has already taken more than seven minutes.

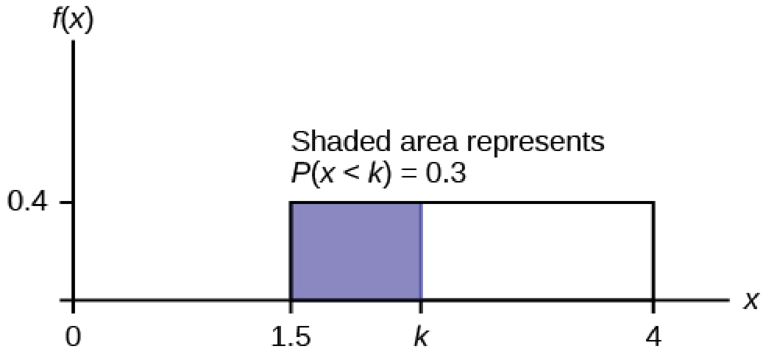
Example 5.8

Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let X = the time needed to fix a furnace. Then $X \sim U(1.5, 4)$.

a. Find the 30th percentile of furnace repair times.
b. The longest 25% of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25% of repair times.) What percentile does this represent?
c. Find the mean and standard deviation

Solution 5.8

a. 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) = (base)(height) = (k – 1.5)(0.4)

0.3 = (k – 1.5) (0.4); Solve to find k:

0.75 = k – 1.5, obtained by dividing both sides by 0.4

k = 2.25, obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.5 hours or less.

b. 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) = (base)(height)

0.25 = (4 – k)(0.4); Solve for k:

0.625 = 4 − k, obtained by dividing both sides by 0.4

−3.375 = −k, obtained by subtracting four from both sides: k = 3.375

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the 75th percentile of furnace repair times.

c. $\mu = (a + b)/2 = (1.5 + 4)/2 = 2.75$ hours

$$\sigma = \frac{(b-a)}{\sqrt{12}} = \frac{4-1.5}{\sqrt{12}} = 0.7217$$

Try It Σ

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let X = the time needed to change the oil on a car.

a. Write the random variable X in words. .

b. Write the distribution.

c. Graph the distribution.

d. Find P ($x > 19$).

e. Find the 50th percentile.

5.3 | 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 who spend small amounts of money and fewer people who spend large amounts of money.

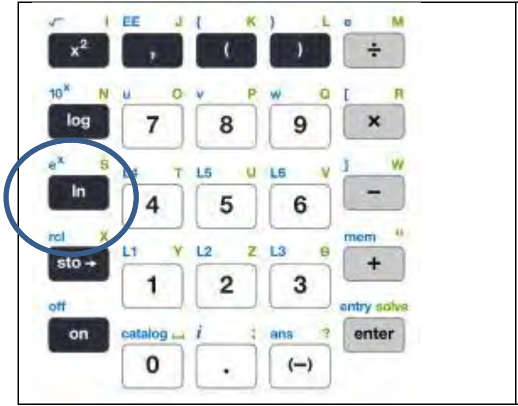
The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

Notation and Properties of the Exponential Distribution

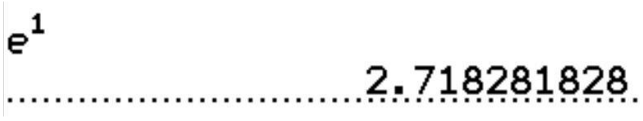
- $X \sim Exp(\text{m})$ where m = is the rate of decay (m > 0)
- The probability density function is $f(x) = me^{-mx}$
- The **mean** of an exponential probability distribution, $\mu = \dfrac{1}{m}$
- The **standard deviation,** $\sigma = \mu$
- $\text{P}(x > \text{k}) = e^{-mx}$
- $\text{P}(x < \text{k}) = 1 - e^{-mx}$
- kth percentile $= \dfrac{\ln(\textit{area to right})}{-m}$

Graphing an exponential probability density function, $f(x) = me^{-mx}$, you will have a y-intercept of (0, m). Since m is positive and there is a negative sign in front of it, as x increases the function decreases. 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 one for x, the calculator will display the value e.

Using the TI-83, 83+, 84, 84+ Calculator



2$^{\text{nd}}$ ln



```
e^1
.....................2.718281828
```

Example 5.9

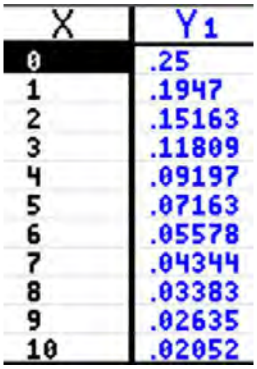
Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes.

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.

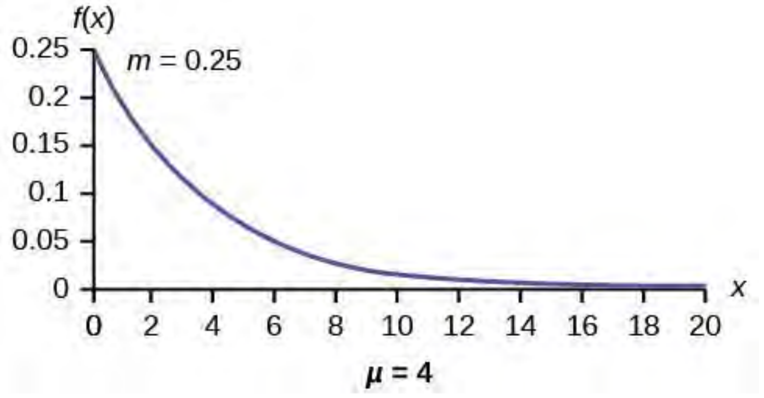
$\mu = \frac{1}{m} \rightarrow m = \frac{1}{\mu}$ Therefore m = .25

The distribution notation is $X \sim Exp(\text{m})$. Therefore, $X \sim Exp(0.25)$.

$f(x) = .25e^{-.25x}$; where x is at least zero. For example, $f(5) = 0.25e^{-(0.25)(5)} = 0.072$.



X	Y1
0	.25
1	.1947
2	.15163
3	.11809
4	.09197
5	.07163
6	.05578
7	.04344
8	.03383
9	.02635
10	.02052



The standard deviation, σ, is the same as the mean. μ = σ.

Try It Σ

The amount of time spouses shop for anniversary cards can be modeled by an exponential distribution with the average amount of time equal to eight minutes. Write the distribution, state the probability density function, and graph the distribution.

Example 5.10

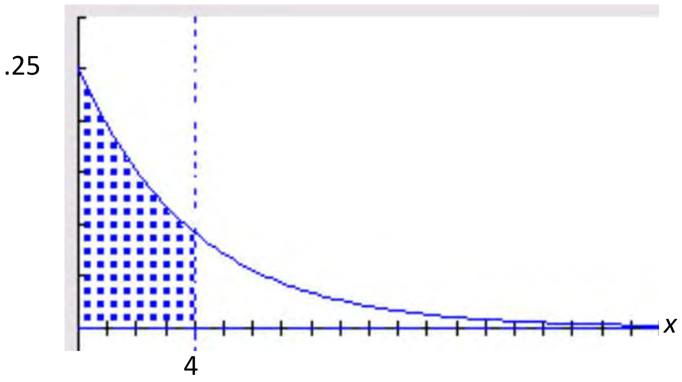
a. Using the information in Example 5.9, find the probability that a clerk spends less than four minutes with a randomly selected customer.

b. Half of all customers are finished within how long? (Find the 50th percentile).

Solution 5.10

a. Find $P(x < 4)$.

The cumulative distribution function (CDF) gives the area to the left.



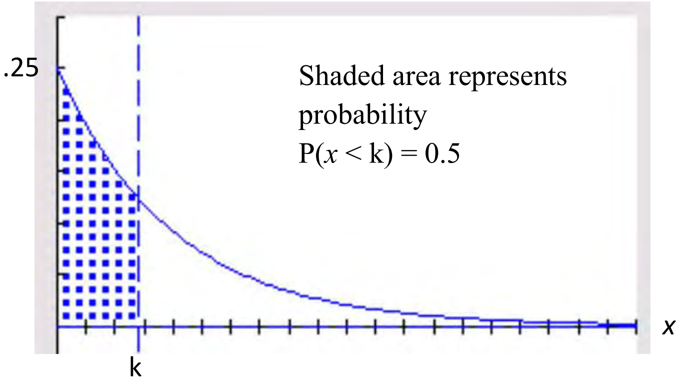
$P(x < k) = 1 - e^{-mk}$

$P(x < 4) = 1 - e^{(-0.25)(4)} = 0.6321$

The probability that a postal clerk spends less than four minutes with a randomly selected customer is 63.21%.

NOTE: $\mu = 4$ is the mean and 63.21% is below it. It is unlike uniform distribution where μ is the median too.

b. Here it is asking us to find the 50th percentile (the median).



$P(x < k) = 0.50$

kth percentile = $\dfrac{\ln(\mathit{area\ to\ right})}{-m}$ where ln is the natural logarithm.

50th percentile = $\dfrac{\ln(.5)}{-.25}$

k = 2.8 minutes Half of all customers are finished within 2.8 minutes.

NOTE: The mean is larger. The mean is 4 minutes and the median is 2.8 minutes.

Try It Σ

The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than ten days in advance. How many days do half of all travelers wait?

Example 5.11

On the average, a certain computer part lasts ten years. The length of time the computer part lasts is exponentially distributed.

a. What is the probability that a computer part lasts more than 7 years?

b. On the average, how long would five computer parts last if they are used one after another?

c. Eighty percent of computer parts last at most how long?

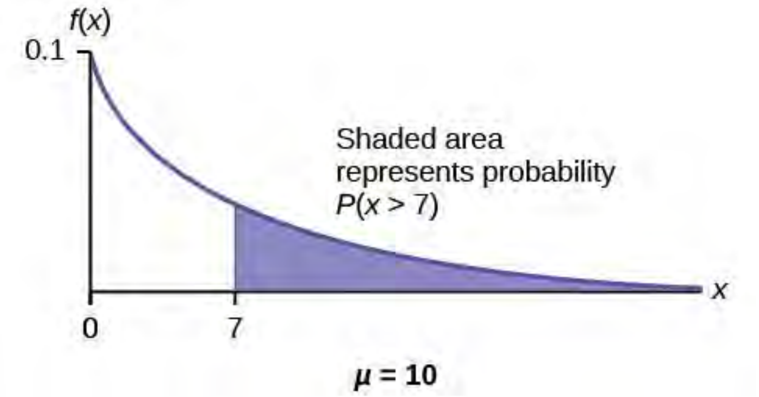
d. What is the probability that a computer part lasts between nine and 11 years?

Solution 5.11

a. Let X = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = 0.1$

Find $P(x > 7)$. Draw the graph.

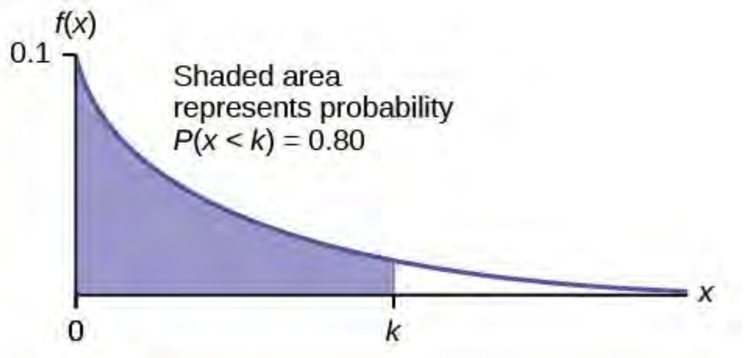


$P(x > k) = e^{-mk}$

$P(x > 7) = e^{(-0.1)(7)} = 0.4966.$

b. On the average, one computer part lasts ten years. Therefore, five computer parts, if they are used one right after the other would last, on the average, (5)(10) = 50 years.

c. Find the 80th percentile. Draw the graph. Let k = the 80th percentile.



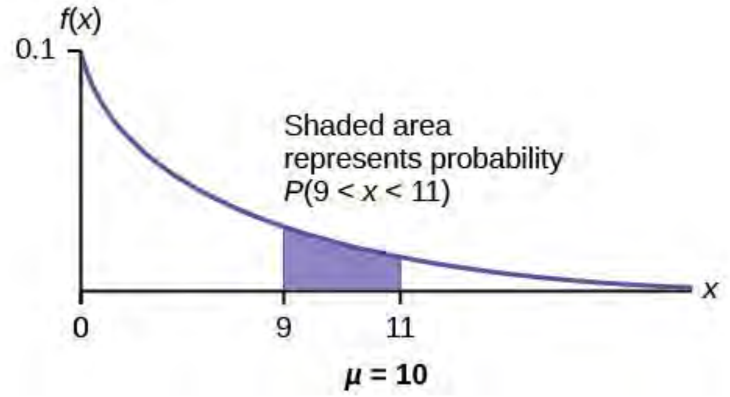
Solve for k: 80% is below *k* and 20% is above *k*.

kth percentile = $\frac{\ln(\mathit{area\ to\ right})}{-m}$

k = $\frac{\ln(.2)}{-.1}$ = 16.1 years

Eighty percent of the computer parts last at most 16.1 years.

d. Find P(9 < *x* < 11). Draw the graph.



P(9 < *x* < 11) = P(*x* < 11) – P(*x* < 9)

P(*x* < 11) = 1 – $e^{(-0.1)(11)}$ = 0.6671

P(*x* < 9) = 1 – $e^{(-0.1)(9)}$ = 0.5934

P(9 < *x* < 11) =0.6671 – 0.5934 = 0.0737.

The probability that a computer part lasts between nine and 11 years is 7.37%

Try It Σ

On average, a pair of running shoes can last 18 months if used every day. The length of time running shoes last is exponentially distributed. What is the probability that a pair of running shoes last more than 15 months? On average, how long would six pairs of running shoes last if they are used one after the other? Eighty percent of running shoes last at most how long if used every day?

Example 5.12

The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that an average of 30 customers per hour arrive at a store and the time between arrivals is exponentially distributed.

a. On average, how many minutes elapse between two successive arrivals.
b. When the store first opens, how long on average does it take for three customers to arrive?
c. After a customer arrives, find the probability that it takes less than one minute for the next customer to arrive.
d. After a customer arrives, find the probability that it takes more than five minutes for the next customer to arrive.
e. Seventy percent of the customers arrive within how many minutes of the previous customer?
f. Is an exponential distribution reasonable for this situation?

Solution 5.12

a. Since we expect 30 customers to arrive per hour (60 minutes), we expect on average one customer to arrive every two minutes on average.
b. Since one customer arrives every two minutes on average, it will take six minutes on average for three customers to arrive.
c. Let X = the time between arrivals, in minutes. By part a, μ = 2, so m = ½ = .5

   $X \sim Exp(0.5)$. Exponential distribution property states $P(x < k) = 1 - e^{(-0.5k)}$

   Therefore $P(x < 1) = 1 - e^{(-0.5)(1)} \approx 0.3935$.

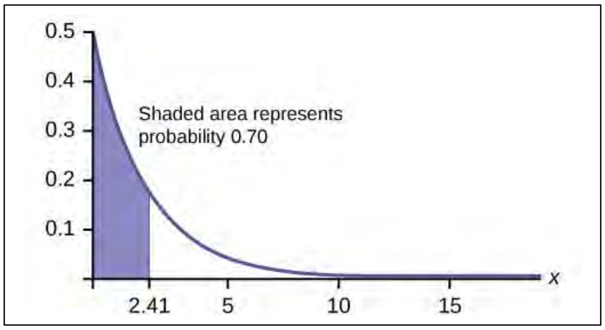
d. Exponential distribution property states and $P(x > k) = e^{(-0.5k)}$

   Therefore $P(x > 5) = e^{(-5)(0.5)} \approx 0.0821$.

e. We want to solve for the 70th percentile

   $$\text{kth percentile} = \frac{\ln(\mathit{area\ to\ right})}{-m}$$

   $$k = \frac{\ln(.3)}{-.5} = 2.41$$



Thus, seventy percent of customers arrive within 2.41 minutes of the previous customer.

f. This model assumes that a single customer arrives at a time, which may not be reasonable since people might shop in groups, leading to several customers arriving at the same time. It also assumes that the flow of customers does not change throughout the day, which is not valid if some times of the day are busier than others.

Try It Σ

Suppose that on a certain stretch of highway, cars pass at an average rate of five cars per minute. Assume that the duration of time between successive cars follows the exponential distribution.

a. On average, how many seconds elapse between two successive cars?
b. After a car passes by, how long on average will it take for another seven cars to pass by?
c. Find the probability that after a car passes by, the next car will pass within the next 20 seconds.
d. Find the probability that after a car passes by, the next car will not pass for at least another 15 seconds.

Memorylessness of the Exponential Distribution

In **Example 5.9**, recall that the amount of time between customers is exponentially distributed with a mean of two minutes ($X \sim Exp\ (0.5)$). Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case–the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**.

Memoryless property says that

$$\mathrm{P}\ (x > \mathrm{r} + \mathrm{t} \mid x > \mathrm{r}) = \mathrm{P}\ (x > \mathrm{t}) \text{ for all } \mathrm{r} \geq 0 \text{ and } \mathrm{t} \geq 0$$

For example, if five minutes has elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using r = 5 and t = 1 in the foregoing equation.

$$\mathrm{P}(x > 5 + 1 \mid x > 5) = \mathrm{P}(x > 1) = \mathrm{e}^{(-0.5)(1)} \approx 0.6065.$$

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

The exponential distribution is often used to model the longevity of an electrical or mechanical device. In **Example 5.11,** the lifetime of a certain computer part has the exponential distribution with a mean of ten years ($X \sim Exp(0.1)$). The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is P(x > 17|x > 10) = P(x > 7) = 0.4966.

Example 5.13

Refer to Example 5.9 where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

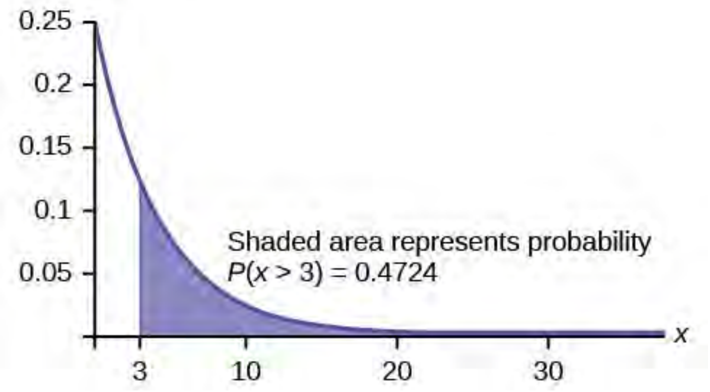
Solution 5.13

The decay parameter of X is m = .25 so $X \sim Exp(0.25)$.

Exponential distribution property states P(x > k) = $e^{-0.25k}$.

We want to find P(x > 7|x > 4). The **memoryless property** says that P(x > 7|x > 4) = P (x > 3), so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is P(x > 3) = $e^{-0.25.3} \approx 0.4724$.



Try It Σ

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years. If a bulb has already lasted 12 years, find the probability that it will last a total of over 19 years.

Relationship between the Poisson and the Exponential Distribution

There is an interesting relationship between the exponential distribution and the Poisson distribution. Suppose that the time that elapses between two successive events follows the exponential distribution with a mean of μ units of time. Also assume that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a Poisson distribution with mean λ =1/μ. Recall from the chapter on Discrete Random Variables that if X has the Poisson distribution with mean λ, then P(x = k) = $\frac{\lambda^k e^{-\lambda}}{k!}$

Conversely, if the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution.

Example 5.14

At a police station in a large city, calls come in at an average rate of four calls per minute. Assume that the time that elapses from one call to the next has the exponential distribution. Take note that we are concerned only with the rate at which calls come in, and we are ignoring the time spent on the phone. We must also assume that the times spent between calls are independent. This means that a particularly long delay between two calls does not mean that there will be a shorter waiting period for the next call. We may then deduce that the total number of calls received during a time period has the Poisson distribution.

a. Find the average time between two successive calls.
b. Find the probability that after a call is received, the next call occurs in less than ten seconds.
c. Find the probability that exactly five calls occur within a minute.
d. Find the probability that less than five calls occur within a minute.
e. Find the probability that more than 40 calls occur in an eight-minute period.

Solution 5.14

a. On average there are four calls occur per minute, so 15 seconds, or 15/60 = 0.25 minutes occur between successive calls on average.

b. Let T = time elapsed between calls. From part a, μ = 0.25, so m = $\frac{1}{\mu}$ = 4. Thus, T ~*Exp*(4). The cumulative distribution function is P(t < k) = 1 – e^{-4t}.

The probability that the next call occurs in less than ten seconds (ten seconds = 1/6 minute) is $P\left(t<\frac{1}{6}\right)=1-e^{-4\cdot\left(\frac{1}{6}\right)}\approx 0.4866$

c. Let X = the number of calls per minute. As previously stated, the number of calls per minute has a Poisson distribution, with a mean of four calls per minute. Therefore, X ~ Poisson(4), and so
P(x = 5) = $\frac{4^5e^{-4}}{5!}$ = poissonpdf(4, 5) ≈ 0.1563 (NOTE: 5! = 5·4·3·2·1 = 120)

d. Keep in mind that X must be a whole number, so P(x < 5) = P(x ≤ 4).To compute this, we could take P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4). Using technology, we see that P(x ≤ 4) = poisssoncdf(4, 4) = 0.6288

e. Let Y = the number of calls that occur during an eight minute period.

Since there is an average of four calls per minute, there is an average of (8)(4) = 32 calls during each eight minute period. Hence, Y ~ Poisson(32). Therefore, P(y > 40) = 1 – P (y ≤ 40) = 1 – poissoncdf(32, 40). = 0.0707

KEY TERMS

Decay parameter The decay parameter describes the rate at which probabilities decay to zero for increasing values of *x*. It is the value m in the probability density function f(x) = me$^{(-mx)}$ of an exponential random variable. It is also equal to m = $\frac{1}{\mu}$, where μ is the mean of the random variable.

Exponential Distribution a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is *X* ~ *Exp*(m). The mean is μ = $\frac{1}{m}$ and the standard deviation is σ = $\frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is P($x \leq$ k) = 1 − e^{-mx}.

Memoryless property for an exponential random variable *X*, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds *x* + k, given that it has exceeded *x*, is the same as the probability that *X* would exceed k if we had no knowledge about it. In symbols we say that P(x > r + k|x > r) = P(x > k).

Poisson distribution If there is a known average of λ events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of k events occurring in one unit time is equal to P(x = k) = $\frac{\lambda^k e^{-\lambda}}{k!}$.

Uniform Distribution a continuous random variable (RV) that has equally likely outcomes over the domain, a < *x* < b; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. Notation: *X* ~ U(a,b). The mean is μ = $\frac{a+b}{2}$ and the standard deviation is σ = $\frac{b-a}{\sqrt{12}}$. The probability density function is *f*(*x*) = $\frac{1}{b-a}$ for a < *x* < b or a ≤ *x* ≤ b. The cumulative distribution is P($x \leq$ k) = $\frac{k-a}{b-a}$.

Formula Review

5.2 The Uniform Distribution

- $X \sim$ U(a, b) where a = lowest value of x and b = highest value of x.
- The **probability density function** is $f(x) = \frac{1}{b-a}$
- The **mean** of a uniform probability distribution, $\mu = \frac{a+b}{2}$, and
- **Variance** of a uniform probability distribution, $\sigma^2 = \frac{(b-a)^2}{12}$
- The **standard deviation**, $\sigma = \frac{(b-a)}{\sqrt{12}}$
- P(x < k) = $(k-a)\frac{1}{b-a}$
- P(c < x < d) = $(d-c)\frac{1}{b-a}$

Probability = Area

Area = Base·Height

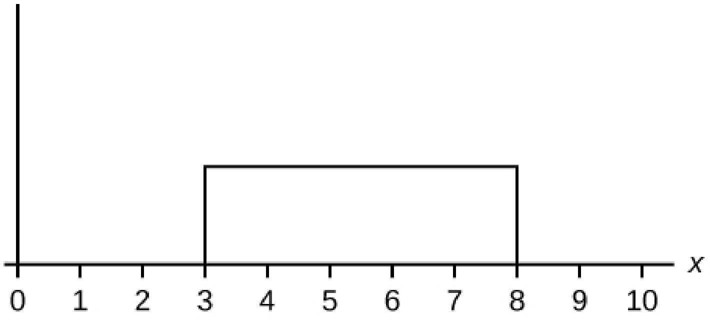
5.3 The Exponential Distribution

- $X \sim Exp$(m) where m = is the rate of decay
- The probability density function is $f(x) = me^{-mx}$ where x ≥ 0 and m > 0
- The **mean** of an exponential probability distribution, $\mu = \frac{1}{m}$
- The **standard deviation,** $\sigma = \mu$
- P(x > k) = e^{-mx}
- P(x < k) = 1 - e^{-mx}
- kth percentile = $\frac{\ln(area\ to\ right)}{-m}$
- Memoryless Property: P(x > r + k|x > r) = P(x > k)
- Poisson Probability: P(x = k) = $\frac{\lambda^k e^{-\lambda}}{k!}$.
- k! = k·(k-1)·(k-2)·…·3·2·1

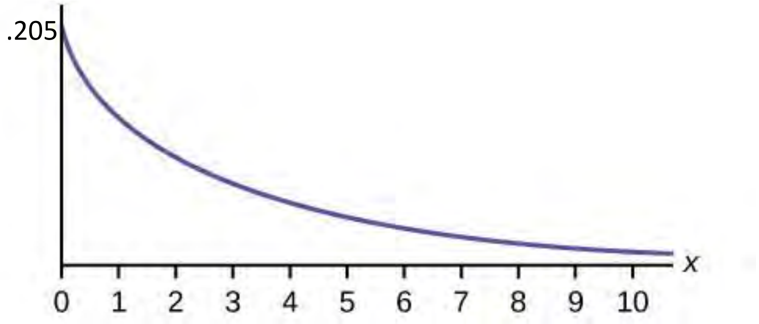
EXERCISES FOR CHAPTER 5

For each probability and percentile problem, draw the picture.

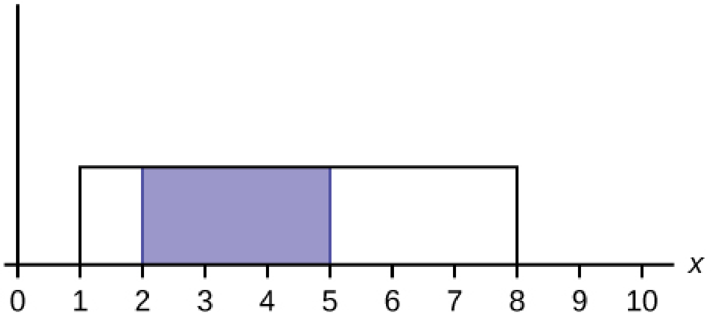
1. Write the notation for the following distribution.



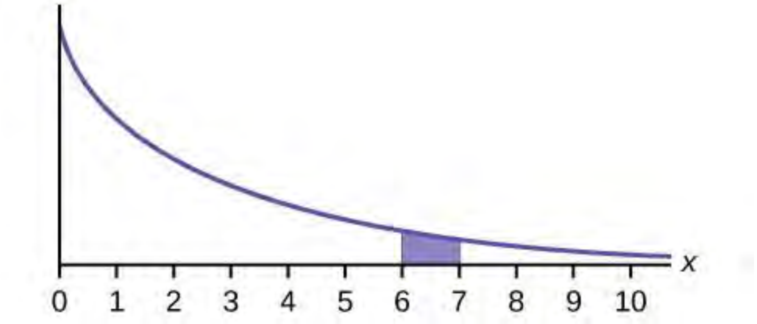
2. Write the notation for the following distribution.



3. What does the shaded area represent? P(< x <)



4. What does the shaded area represent? P(< x <)



5. For a continuous probability distribution, $0 \le x \le 15$. What is $P(x > 15)$?

6. What is the area under $f(x)$ if the function is a continuous probability density function?

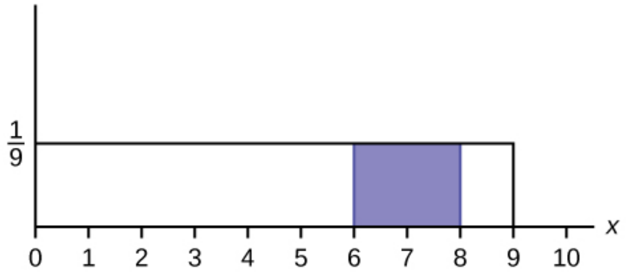
7. For a continuous probability distribution, $0 \le x \le 10$. What is $P(x = 7)$?

8. A continuous probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

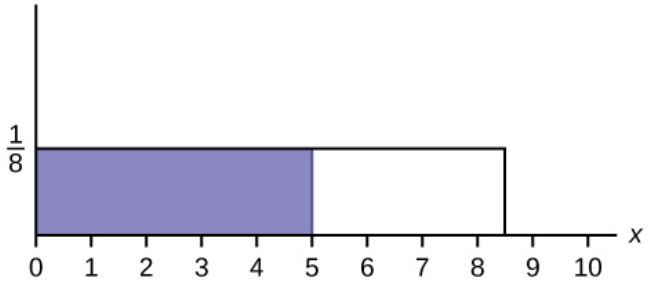
9. $f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \le x \le 5$. What is $P(x < 0)$?

10. $f(x)$, a continuous probability function, is equal to $\frac{1}{12}$ and the function is restricted to $0 \le x \le 12$. What is $P(0 < x < 12)$?

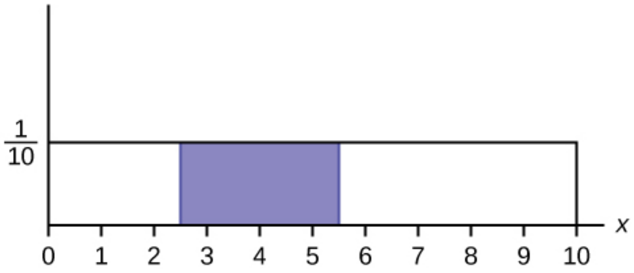
11. Find the probability that x falls in the shaded area:



12. Find the probability that x falls in the shaded area:



13. Find the probability that x falls in the shaded area:



14. $f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \leq x \leq 4$. Describe P($x > \frac{3}{2}$)

15. The uniform distribution $X \sim$ U(1.5, 4.5) describes the square footage (in 1000 feet squared)
   a. What is the height of $f(x)$ for the continuous probability distribution?
   b. What are the constraints for the values of x?
   c. Graph P($2 < x < 3$).
   d. What is P($2 < x < 3$)?
   e. What is P($x < 3.5 | x < 4$)?
   f. What is P($x = 1.5$)?
   g. What is the 90th percentile of square footage for homes?
   h. Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.

16. A distribution is given as $X \sim$ U(0, 12)
   a. What is a? What does it represent?
   b. What is b? What does it represent?
   c. What is the probability density function?
   d. What is the theoretical mean?
   e. What is the theoretical standard deviation?
   f. Draw the graph of the distribution for P(x > 9).
   g. Find P(x > 9).
   h. Find the 40th percentile.

17. 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.
   a. What is being measured here?
   b. In words, define the random variable X.
   c. Are the data discrete or continuous?
   d. The interval of values for x is
   e. The distribution for X is
   f. Write the probability density function.
   g. Graph the probability distribution.
   h. Find the probability that a randomly chosen car in the lot was less than four years old.
   i. Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.
   j. What changed in the previous two problems that made the solutions different?
   k. Find the third quartile of ages of cars in the lot.

18. A customer service representative must spend different amounts of time with each customer to resolve various concerns. The amount of time spent with each customer can be modeled by the following distribution: $X \sim Exp(0.2)$
   a. What type of distribution is this?
   b. Are outcomes equally likely in this distribution? Why or why not?

c. What is m? What does it represent?
d. What is the mean?
e. What is the standard deviation?
f. State the probability density function.
g. Graph the distribution.
h. Find P(2 < x < 10).
i. Find P(x > 6).
j. Find the 70th percentile.

19. A distribution is given as X ~ Exp(0.75).
   a. What is m?.
   b. What is the probability density function?
   c. What is the cumulative distribution function?
   d. Draw the distribution.
   e. Find P(x < 4).
   f. Find the 30th percentile.
   g. Find the median.
   h. Which is larger, the mean or the median?

20. Carbon-14 is a radioactive element with a half-life of about 5,730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121. We start with one gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

   a. What is being measured here?
   b. Are the data discrete or continuous?
   c. In words, define the random variable X.
   d. What is the decay rate (m)?
   e. The distribution for X is
   f. Find the amount (percent of one gram) of carbon-14 lasting less than 5,730 years. This means, find P(x < 5,730).
   g. Find the percentage of carbon-14 lasting longer than 10,000 years.
   h. Thirty percent (30%) of carbon-14 will decay within how many years?

21. Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer "yes" or "no." You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.
   a. What part of the experiment will yield discrete data?
   b. What part of the experiment will yield continuous data?

22. When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

23. Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).
    a. *X* ~
    b. Graph the probability distribution.
    c. *f*(*x*) =
    d. μ =
    e. σ =
    f. Find the probability that a person is born at the exact moment week 19 starts. That is, find P(x = 19) =
    g. P(2 < x < 31) =
    h. Find the probability that a person is born after week 40.
    i. P(12 < x|x < 28) =
    j. Find the 70th percentile.
    k. Find the minimum for the upper quarter.

24. A random number generator picks a number from one to nine in a uniform manner.
    a. X ~
    b. Graph the probability distribution.
    c. *f*(*x*) =
    d. μ =
    e. σ =
    f. P(3.5 < x < 7.25) =
    g. P(x > 5.67)
    h. P(x > 5|x > 3) =
    i. Find the 90th percentile.

25. According to a study by Dr. John McDougall of his live-in weight loss program at St. Helena Hospital, the people who follow his program lose between six and 15 pounds a month until they approach trim body weight. Let's suppose that the weight loss is uniformly distributed. We are interested in the weight loss of a randomly selected individual following the program for one month.
    a. Define the random variable. *X* =
    b. *X* ~
    c. Graph the probability distribution.
    d. *f*(*x*) =
    e. μ =
    f. σ =
    g. Find the probability that the individual lost more than ten pounds in a month.
    h. Suppose it is known that the individual lost more than ten pounds in a month. Find the probability that he lost less than 12 pounds in the month.
    i. P(7 < x < 13|x > 9) = . State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

26. A subway train on the Red Line arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution.
    a. Define the random variable. $X =$
    b. $X \sim$
    c. Graph the probability distribution.
    d. $f(x) =$
    e. $\mu =$
    f. $\sigma =$
    g. Find the probability that the commuter waits less than one minute.
    h. Find the probability that the commuter waits between three and four minutes.
    i. Sixty percent of commuters wait more than how long for the train? State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

27. The age of a first grader on September 1 at Garden Elementary School is uniformly distributed from 5.8 to 6.8 years. We randomly select one first grader from the class.
    a. Define the random variable. $X =$
    b. $X \sim$
    c. Graph the probability distribution.
    d. $f(x) =$
    e. $\mu =$
    f. $\sigma =$
    g. Find the probability that she is over 6.5 years old.
    h. Find the probability that she is between four and six years old.
    i. Find the 70th percentile for the age of first graders on September 1 at Garden Elementary School.

28. The Sky Train from the terminal to the rental–car and long–term parking center is supposed to arrive every eight minutes. The waiting times for the train are known to follow a uniform distribution.
    a. What is the average waiting time (in minutes)?
    b. Find the 30th percentile for the waiting times (in minutes).
    c. The probability of waiting more than seven minutes given a person has waited more than four minutes is?

29. The time (in minutes) until the next bus departs a major bus depot follows a distribution with $f(x) = \frac{1}{20}$ where x goes from 25 to 45 minutes.
    a. Define the random variable. $X =$
    b. $X \sim$
    c. Graph the probability distribution.
    d. The distribution is (name of distribution). Determine if it is discrete or continuous.
    e. $\mu =$
    f. $\sigma =$

g. Find the probability that the time is at most 30 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.

h. Find the probability that the time is between 30 and 40 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.

i. P(25 < x < 55) = . State this in a probability statement, similarly to parts g and h, draw the picture, and find the probability.

j. Find the 90th percentile. This means that 90% of the time, the time is less than minutes.

k. Find the 75th percentile. In a complete sentence, state what this means. (See part j.)

l. Find the probability that the time is more than 40 minutes given (or knowing that) it is at least 30 minutes.

30. Suppose that the value of a stock varies each day from $16 to $25 with a uniform distribution.

a. Find the probability that the value of the stock is more than $19.

b. Find the probability that the value of the stock is between $19 and $22.

c. Find the upper quartile - 25% of all days the stock is above what value? Draw the graph.

d. Given that the stock is greater than $18, find the probability that the stock is more than $21.

31. A fireworks show is designed so that the time between fireworks is between one and five seconds, and follows a uniform distribution.

a. Find the average time between fireworks.

b. Find probability that the time between fireworks is greater than four seconds.

32. The number of miles driven by a truck driver falls between 300 and 700, and follows a uniform distribution.

a. Find the probability that the truck driver goes more than 650 miles in a day.

b. Find the probability that the truck drivers goes between 400 and 650 miles in a day.

c. At least how many miles does the truck driver travel on the furthest 10% of days?

33. Suppose that the length of long distance phone calls, measured in minutes, is known to have an exponential distribution with the average length of a call equal to eight minutes.

a. Define the random variable. $X =$

b. $X \sim$

c. Determine if the variable is discrete or continuous.

d. $\mu =$

e. $\sigma =$

f. Draw a graph of the probability distribution. Label the axes.

g. Find the probability that a phone call lasts less than nine minutes.

h. Find the probability that a phone call lasts more than nine minutes.

i. Find the probability that a phone call lasts between seven and nine minutes.

j. If 25 phone calls are made one after another, on average, what would you expect the total to be? Why?

34. Suppose that the useful life of a particular car battery, measured in months, decays with parameter 0.025. We are interested in the life of the battery.

a. Define the random variable. $X =$.

b. Is X continuous or discrete?
c. $X \sim$
d. On average, how long would you expect one car battery to last?
e. On average, how long would you expect nine car batteries to last, if they are used one after another?
f. Find the probability that a car battery lasts more than 36 months.
g. Seventy percent of the batteries last at least how long?

35. The percent of persons (ages five and older) in each state who speak a language at home other than English is approximately exponentially distributed with a mean of 9.848. Suppose we randomly pick a state.
a. Define the random variable. $X =$.
b. Is X continuous or discrete?
c. $X \sim$
d. $\mu =$
e. $\sigma =$
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that the percent is less than 12.
h. Find the probability that the percent is between eight and 14.
i. The percent of all individuals living in the United States who speak a language at home other than English is 13.8.
i. Why is this number different from 9.848%?
ii. What would make this number higher than 9.848%?

36. The time (in years) after reaching age 60 that it takes an individual to retire is approximately exponentially distributed with a mean of about five years. Suppose we randomly pick one retired individual. We are interested in the time after age 60 to retirement.
a. Define the random variable. $X =$.
b. Is X continuous or discrete?
c. $X \sim$
d. $\mu =$
e. $\sigma =$
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that the person retired after age 70.
h. Do more people retire before age 65 or after age 65?
i. In a room of 1,000 people over age 80, how many do you expect will NOT have retired yet?

37. The cost of all maintenance for a car during its first year is approximately exponentially distributed with a mean of $150.
a. Define the random variable. $X =$.
b. $X \sim$
c. $\mu =$
d. $\sigma =$
e. Draw a graph of the probability distribution. Label the axes.
f. Find the probability that a car required over $300 for maintenance during its first year.

38. The average lifetime of a certain new cell phone is three years. The manufacturer will replace any cell phone failing within two years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.
    a. The decay rate is:
    b. What is the probability that a phone will fail within two years of the date of purchase?
    c. What is the median lifetime of these phones (in years)?

39. Let X ~ Exp(0.1).
    a. decay rate =
    b. μ =
    c. Graph the probability distribution function.
    d. On the graph, shade the area corresponding to $P(x < 6)$ and find the probability.
    e. Sketch a new graph, shade the area corresponding to $P(3 < x < 6)$ and find the probability.
    f. Sketch a new graph, shade the area corresponding to $P(x < 7)$ and find the probability.
    g. Sketch a new graph, shade the area corresponding to the 40th percentile and find the value.
    h. Find the average value of x.

40. Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years.
    a. Find the probability that a light bulb lasts less than one year.
    b. Find the probability that a light bulb lasts between six and ten years.
    c. Seventy percent of all light bulbs last at least how long?
    d. A company decides to offer a warranty to give refunds to light bulbs whose lifetime is among the lowest two percent of all bulbs. To the nearest month, what should be the cutoff lifetime for the warranty to take place?
    e. If a light bulb has lasted seven years, what is the probability that it fails within the 8th year.

41. At a 911 call center, calls come in at an average rate of one call every two minutes. Assume that the time that elapses from one call to the next has the exponential distribution.
    a. On average, how much time occurs between five consecutive calls?
    b. Find the probability that after a call is received, it takes more than three minutes for the next call to occur.
    c. Ninety-percent of all calls occur within how many minutes of the previous call?
    d. Suppose that two minutes have elapsed since the last call. Find the probability that the next call will occur within the next minute.
    e. Find the probability that less than 20 calls occur within an hour.

42. In major league baseball, a no-hitter is a game in which a pitcher, or pitchers, doesn't give up any hits throughout the game. No-hitters occur at a rate of about three per season. Assume that the duration of time between no-hitters is exponential.
    a. What is the probability that an entire season elapses with a single no-hitter?
    b. If an entire season elapses without any no-hitters, what is the probability that there are no no-hitters in the following season?
    c. What is the probability that there are more than 3 no-hitters in a single season?

43. During the years 1998–2012, a total of 29 earthquakes of magnitude greater than 6.5 have occurred in Papua New Guinea. Assume that the time spent waiting between earthquakes is exponential.
    a. What is the probability that the next earthquake occurs within the next three months?
    b. Given that six months has passed without an earthquake in Papua New Guinea, what is the probability that the next three months will be free of earthquakes?
    c. What is the probability of zero earthquakes occurring in 2014?
    d. What is the probability that at least two earthquakes will occur in 2014?

44. According to the American Red Cross, about one out of nine people in the U.S. have Type B blood. Suppose the blood types of people arriving at a blood drive are independent. In this case, the number of Type B blood types that arrive roughly follows the Poisson distribution.
    a. If 100 people arrive, how many on average would be expected to have Type B blood?
    b. What is the probability that over 10 people out of these 100 have type B blood?
    c. What is the probability that more than 20 people arrive before a person with type B blood is found?

45. A web site experiences traffic during normal working hours at a rate of 12 visits per hour. Assume that the duration between visits has the exponential distribution.
    a. Find the probability that the duration between two successive visits to the web site is more than ten minutes.
    b. The top 25% of durations between visits are at least how long?
    c. Suppose that 20 minutes have passed since the last visit to the web site. What is the probability that the next visit will occur within the next 5 minutes?
    d. Find the probability that less than 7 visits occur within a one-hour period.

46. At an urgent care facility, patients arrive at an average rate of one patient every seven minutes. Assume that the duration between arrivals is exponentially distributed.
    a. Find the probability that the time between two successive visits to the urgent care facility is less than 2 minutes.
    b. Find the probability that the time between two successive visits to the urgent care facility is more than 15 minutes.
    c. If 10 minutes have passed since the last arrival, what is the probability that the next person will arrive within the next five minutes?
    d. Find the probability that more than eight patients arrive during a half-hour period.

REFERENCES

5.2 The Uniform Distribution

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

5.3 The Exponential Distribution

Data from the United States Census Bureau.

Data from World Earthquakes, 2013. Available online at http://www.world-earthquakes.com/ (accessed June 11, 2013).

"No-hitter." Baseball-Reference.com, 2013. Available online at http://www.baseball-reference.com/bullpen/No-hitter(accessed June 11, 2013).

Zhou, Rick. "Exponential Distribution lecture slides." Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).

6 | THE NORMAL DISTRIBUTION

Figure 6.1 If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Introduction

Chapter Objectives

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

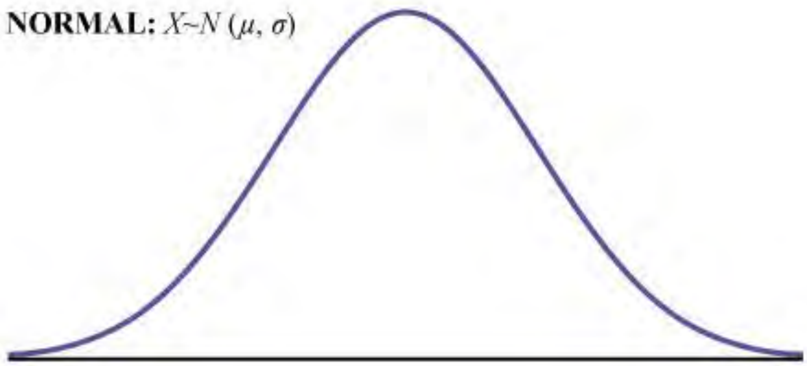
- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.
- Find value of the random variable given area under the curve.

The normal distribution is a continuous distribution, and is the single most important of all the distributions discussed in this text. It is widely used and even more widely abused. Its graph is bell-shaped, and we see the bell curve in many 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, as are many standardized test scores. In this chapter, we will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures of the population), the mean, μ, and the standard deviation, σ. If *X* is a quantity to be measured that has a normal distribution with mean μ and standard deviation σ, we designate this by writing *X* ~ *N*(μ, σ). The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary for calculations and is included only for completeness:

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma}}{\sigma\sqrt{2\pi}}$$

The graph of this function exhibits the bell-shape of the normal distribution:



The cumulative distribution function is P(*X* < *x*). It can be calculated by a calculator, statistical software, or by using a normal distribution table. Modern technology has made the tables virtually obsolete, and we will focus mainly on the use of the TI-84 family of calculators for calculations. But there is still some insight to be gained from working with the tables so we will discuss these briefly.

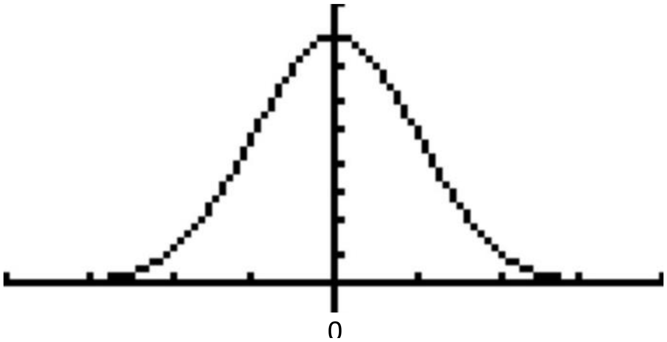
The curve is perfectly symmetrical about a vertical line drawn through the mean, μ. Thus the mean is the same as the median. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. The smaller the standard deviation is, the narrower the normal curve appears. 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. But we can understand all of them by understanding the **standard normal distribution**.

6.1 | The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of standardized values called ***z*-scores**. Recall from Chapter 2 that *z*-score is measured in units of the standard deviation. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean, since

$$x = \mu + z\sigma = 5 + 3{*}2 = 11$$

So the *z*-score corresponding to *x* = 11 is *z* = 3. If the value *x* comes from a normal distribution with mean μ and standard deviation, σ, then the transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. That is, the mean for the standard normal distribution is zero, and the standard deviation is one. Therefore, zero is in the center.



z-Scores

If *X* is a normally distributed random variable and *X* ~ *N*(μ, σ), then the *z*-score is: $z = \frac{x - \mu}{\sigma}$.

The *z*-score tells you how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean μ. Values of *x* that are larger than the mean have positive *z*-scores, and values of *x* that are smaller than the mean have negative *z*-scores. If *x* equals the mean, then *x* has a *z*-score of zero.

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 the corresponding *z*-score is:

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

This means that *x* = 17 is **two standard deviations above** the mean of μ = 5.

Now suppose *x* = 1; then the corresponding *z*-score is:

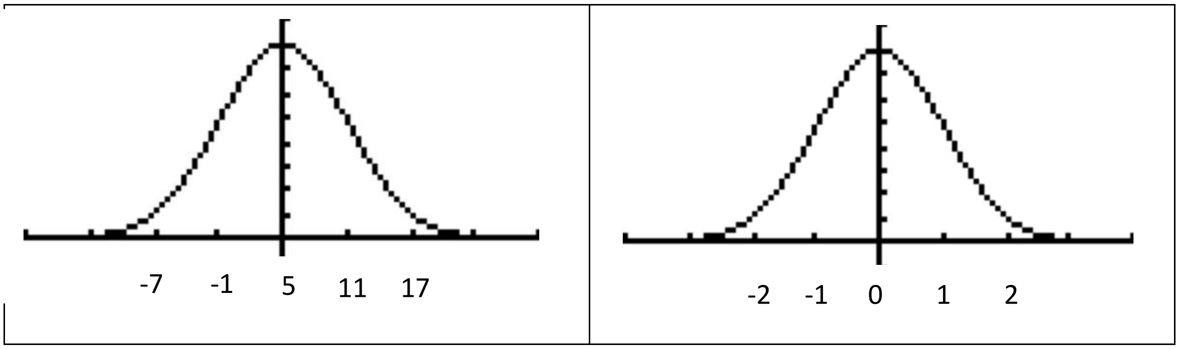
$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67.$$

This means that $x = 1$ is **0.67 standard deviations below** the mean $\mu = 5$**.**

Summarizing,

- When z is positive, x is above, or to the right of μ
- When z is negative, x is to the left of, or below μ.

Graphical interpretation of Example 6.1. The first graph displays $x = 17$ on $X \sim N(5, 6)$. The second graph displays $z = 2$ on $X \sim N(0, 1)$.



Try It Σ

6.1 What is the z-score of x, when $x = 1$ and $X \sim N(12,3)$?

Example 6.2

Some doctors believe that a person can lose five pounds, on average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is _______ standard deviations to the _______ (right or left) of the mean of ___ .

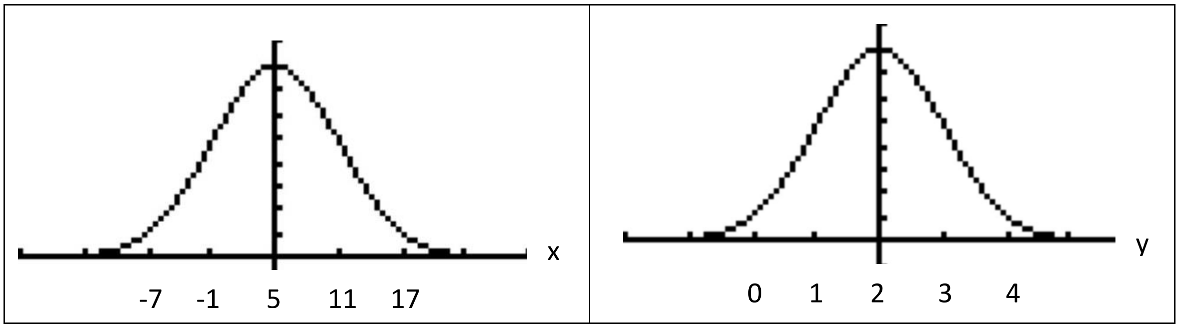
b. Suppose a person **gained** three pounds (a negative weight loss). Then z = _________ . This z-score tells you that $x = –3$ is _______ standard deviations to the ________ (right or left) of the mean.

Solution 6.2

a. This z-score tells you that $x = 10$ is **2.5** standard deviations to the **right** of the mean $\mathbf{\mu = 5}$.

b. $z = \mathbf{–4}$. This z-score tells you that $x = –3$ is **four** standard deviations to the **left** of the mean.

One of the most important aspects of *z*-scores is that they allow us to compare data that are scaled differently**.** To understand the concept, suppose $X \sim N(5,6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.



Try It Σ

6.2 Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The *z*–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is ____standard deviations to the _____(right or left) of the mean______(What is the mean?).

Example 6.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution.
Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010, so $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. Calculate and interpret the *z*-score for this individual.
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = 1.27$. What was the male's height?

Solution 6.3

a. $z = -0.32$. The individual was 0.32 standard deviations to the *left* of the mean $\mu = 170$.

b. $x = \mu + z\sigma = 170 + 1.27(6.28) =$ **177.98 cm**.

Try It Σ

6.3 Use the information in **Example 6.3** to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile in this time period was 176 cm tall. What does that tell us about the individual's height relative to the population?
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = -2$. What was the individual's height?

Example 6.4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.
The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the *z*-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each *z*-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Solution 6.4

The *z*-score for $x = 160.58$ is $z = –1.5$. The *z*-score for $y = 162.85$ is $z = –1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction. So these individuals had the same heights, relative to their respective populations.

Try It Σ

6.4 In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$. Find the *z*-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each *z*-score; what can we say about these two scores?

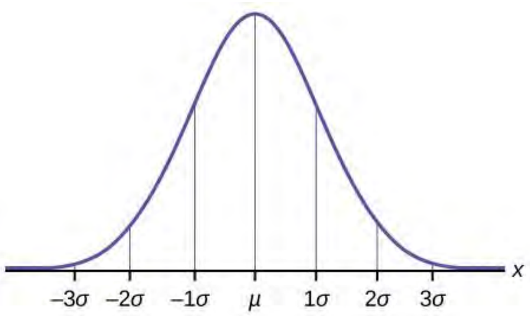
The Empirical Rule

Let X be a random variable that has a normal distribution with mean μ and standard deviation σ. Then the **Empirical Rule** says the following:

- About 68% of the *x* values lie within one standard deviation of the mean. That is, about 68% of all data values lie between $\mu - \sigma$ and $\mu + \sigma$.
- About 95% of the *x* values lie within two standard deviations of the mean. That is, about 95% of all data values lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

- About 99.7% of the *x* values lie within three standard deviations of the mean. That is, about 99.7% of all data values lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Notice that almost all the *x* values lie within three standard deviations of the mean. The empirical rule is also known as the 68-95-99.7 rule.



Example 6.5

Suppose *x* has a normal distribution with mean 50 and standard deviation 6.

- Then $\mu - \sigma = 44$ and $\mu + \sigma = 56$. So about 68% of the *x* values lie between 44 and 56.
- Similarly, $\mu - 2\sigma = 38$ and $\mu + 2\sigma = 62$. So about 95% of the *x* values lie between 38 and 62.
- Finally, $\mu - 3\sigma = 32$ and $\mu + 3\sigma = 68$. So about 99.7% of the *x* values lie between 32 and 68.

Try It Σ

6.5 Suppose *X* has a normal distribution with mean 25 and standard deviation five. Between what values of *x* do 68% of the values lie?

Example 6.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let *Y* = the height of 15 to 18-year-old males in 1984 to 1985, so $Y \sim N(172.36, 6.34)$.

a. About 68% of the *y*-values fall between what two values? What are the respective *z*-scores of these values?
b. About 95% of the *y* values lie between what two values? What are the respective *z*-scores of these values?
c. About 99.7% of the *y* values lie between what two values? What are the respective *z*-scores of these values?

Solution 6.6

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.

b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.

c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Try It Σ

6.6 The scores on a college entrance exam have an approximate normal distribution with mean $\mu = 52$ points and a standard deviation of $\sigma = 11$ points.

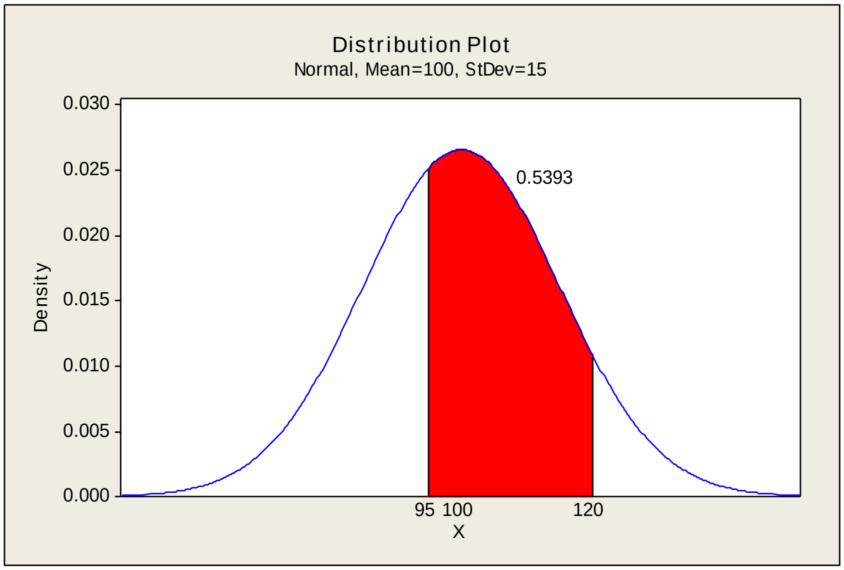
a. About 68% of the y values lie between what two values? What are the respective z-scores?

b. About 95% of the y values lie between what two values? What are the respective z-scores?

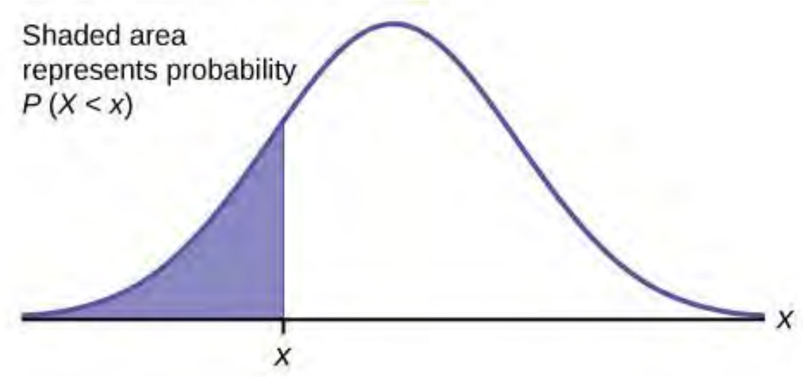
c. About 99.7% of the y values lie between what two values? What are the respective z-scores?

6.2 | Using Normal Distribution

Recall that the probability that the random variable X falls into any interval will be the area under the curve corresponding to that interval. For example, scores for the Stanford-Binet IQ test are normally distributed with a mean of $\mu = 100$ and standard deviation of $\sigma = 15$. Then the probability that a randomly selected individual has an IQ between 95 and 120 would be about 54%, as shown in the graph:



To calculate probabilities like the problem above, we need to work with the *cumulative distribution function*, $P(X < x)$. The shaded area in the graph below shows this probability; it is the area under the normal curve that is to the left of x. To calculate this area manually would require Calculus; but there is a table of values that shows these areas for the standard normal distribution, and many calculators and statistical software packages also will calculate cumulative normal probabilities.



If we are able to calculate the cumulative probability in the figure above, then we can use the **complement** rule to also calculate the probability to the *right* of x:

$$P(X > x) = 1 - P(X < x).$$

Moreover, if $x_1 < x_2$ are two different values of X, then the area *between* x_1 and x_2 is:

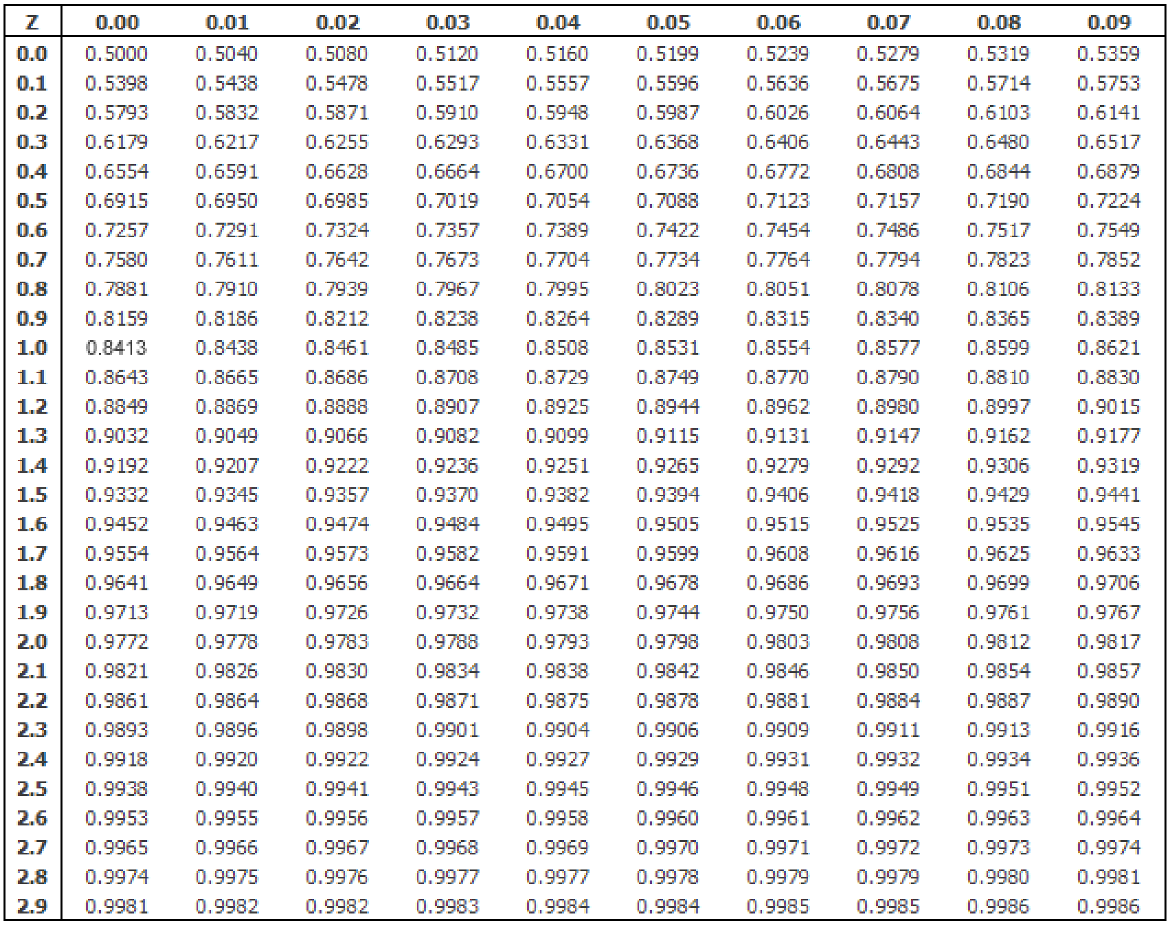
$$P(x_1 < X < x_2) = P(X < x_2) - P(X < x_1).$$

Finally note that $P(X < x)$ is the same as $P(X \le x)$ and $P(X > x)$ is the same as $P(X \ge x)$ for continuous distributions.

Calculating Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators. To calculate normal probabilities without technology we use the normal tables provided in the **Appendix**. The tables include instructions for how to use them; but we will review the steps briefly.

The normal distribution table in the Appendix shows the area under the standardized normal curve to the left of x. I.e. it shows the probability $P(z < x)$. We reproduce a section of the table here:



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

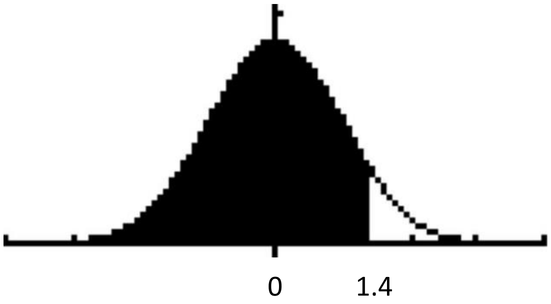
Example 6.7

Use the table to find each of the following:

a. P($z < 1.4$)
b. P($z > 0.74$)
c. P($0 < z < 2.1$)
d. P($1.21 < z < 2.43$)

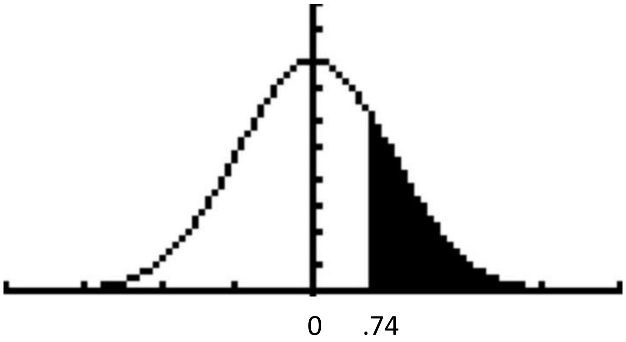
Solution 6.7

a. To find table value corresponding to $z = 1.40$, we look in the row labeled 1.4 and the column labeled as 0.00. This table value is 0.9192, so P($z < 1.4$) = **0.9192.**



b. To find table value corresponding to $z = 0.74$, we look in the row labeled 0.7 and the column labeled as 0.04. This table value is 0.7704, which is the area to the left of $z = 0.74$. We want the area to the *right*, so we use the complement rule:

P($z > 0.74$) = 1 – P($z < 0.74$) = 1 – 0.7704 = **0.2296.**



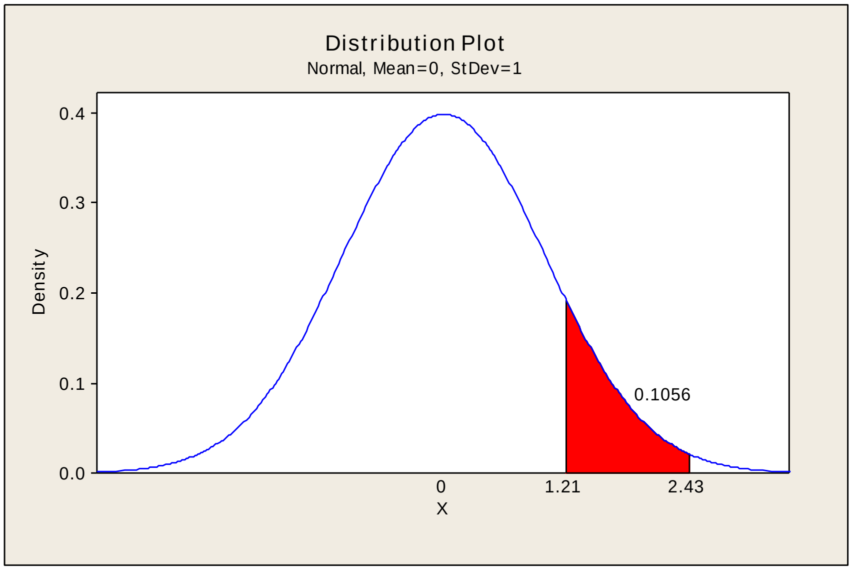
c. The probability P($0 < z < 2.1$) is the area under the normal curve between 0 and 2.1. From the table, the table value for $z = 2.1$ is 0.9821. This is the total area to the left of $z = 2.1$. The area to the left of $z = 0$ is .5000. We can see this in the table, but we also know that the mean and median are both at $z = 0$ due to the perfect symmetry of the normal curve. Thus, the area between z = 0 and z = 2.1 is the difference between these two areas:

P($0 < z < 2.1$) = P($z < 2.1$) – P($z < 0$) = 0.9821 – 0.5000 = **0.4821.**

d. The probability P(1.2 < z < 2.4) is the area under the normal curve between z = 1.2 and z = 2.4. The table value corresponding to z = 1.21 is 0.8869 and the table value of z = 2.43 is 0.9924. Again the area between z = 1.21 and z = 2.43 is the difference between these two areas:

$$P(1.21 < z < 2.43) = P(z < 2.43) - P(z < 1.21) = 0.9924 - 0.8869 = \mathbf{0.1055.}$$

NOTE: For all of these, it is useful to draw a picture of the normal curve, and shade the desired area. E.g., for part d, the graph would look like:



Try It Σ

6.7 Use the tables to find P(z < 0.85), P(z > 1.36) and P(0.45 < z < 1.45)

To use the table for non-standard normally distributed variables, we first rewrite the desired probability in terms of z-scores, and then use the table as demonstrated above. For example, if we have a normally variable $X \sim N(50, 4)$ and wanted to find P(46 < x < 55), we would find the z-scores corresponding to X = 46 and 55, which are z = (46 – 50)/4 = -1 and z = (55 – 50)/4 = 1.25, respectively. From there we could rewrite P(46 < x < 55) = P(-1 < z < 1.25) and use the table.

However, we will usually use technology – either the TI-84 calculator or software – to compute these probabilities:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate normal probabilities

Go to 2nd DISTR, and select item 2: **normalcdf**.
The syntax is: normalcdf (lower bound, upper bound, mean, standard deviation).
I.e., normalcdf(a, b, μ, σ) returns the probability P(a < x < b)

For probabilities of the form P(x < b), substitute -10^99 for the lower bound.
For probabilities of the form P(x > a), substitute 10^99 for the upper bound.

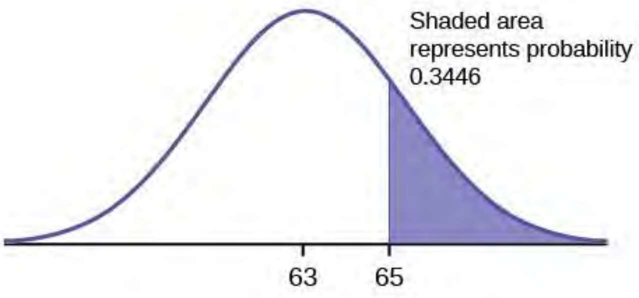
Example 6.8

The final exam scores in a large statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85 on the exam.
c. Find the 90th percentile. That is, find the score that separates the lower 90% of scores from the top 10%.

Solution 6.8

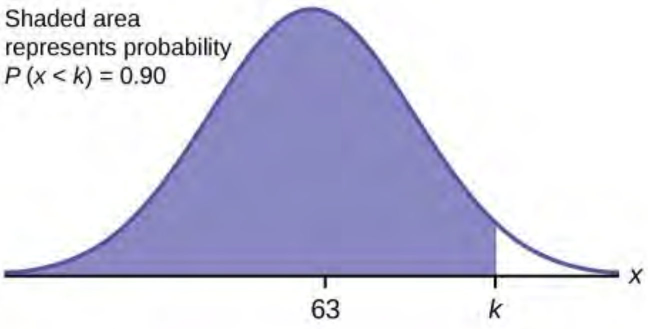
a. Let X = score on the final exam. Then $X \sim N(63, 5)$; i.e. we are given that $\mu = 63$ and $\sigma = 5$.
We start by drawing a graph:



Then, find $P(x > 65)$ = normalcdf(65, 10^99, 63, 5) = **0.3446.**

b. Here we want to find $P(x < 85)$ = normalcdf(-10^99, 85, 63, 5) = **0.999995.**
This tells us that virtually *all* of the students in the class scored below 85.

c. This time we want a value x so that the area to the left of x is exactly 90%.
Again, it is useful to draw a graph:



And there is a function in the calculator that will calculate this value:

.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate percentiles in a normal distribution

Go to 2nd DISTR, and select item 3: **invNorm**.
The syntax is: invNorm(probability to the left, mean, standard deviation).
I.e., $k =$ invNorm(p, μ, σ) value k so that $P(x < k) = p$.

To find the 90th percentile of the exam scores, we enter invNorm(.90, 63, 5) and press ENTER to get 69.4. So 90% of the test scores were at or below 69.4 points.

d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution: 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

Try It Σ

6.8 The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

a. Find the probability that a randomly selected golfer scored less than 65.
b. Find the 80th percentile

Example 6.9

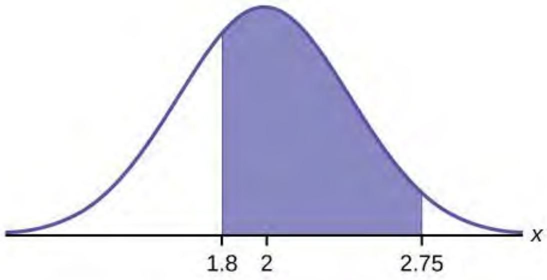
A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times spent on entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution 6.9

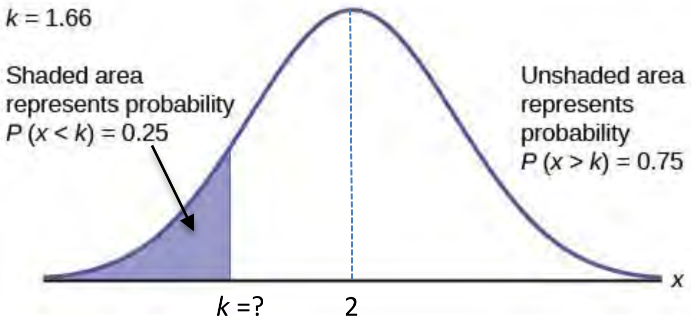
Let X = the amount of time (in hours) a household personal computer is used for entertainment.
We are told that $\mu = 2$ and $\sigma = 1/2$, so $X \sim N(2, 0.5)$

a. We wish to find $P(1.8 < x < 2.75)$. This probability is the area between $x = 1.8$ and $x = 2.75$:



Using the calculator, P(1.8 < *x* < 2.75) = 0.5886 = normalcdf(1.8, 2.75, 2, 0.5) = **0.5886**.

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** *k*. This is the value so that P(*x* < *k*) = 0.25:



Again using the calculator, *k* = invNorm(0.25, 2, 0.5) = **1.66.** So the

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

Try It Σ

6.9 The golf scores for a school team are normally distributed with a mean of 68 and a standard deviation of 3.

a. Find the probability that a randomly selected golfer scores between 66 and 70.
b. Find the score for a golfer that is at the third quartile.

Example 6.10

There are approximately one billion smartphone users in the world today. In the United States the ages of smartphone users are approximately normally distributed with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a randomly selected smartphone user is between 23 and 64.7 years old.
b. Determine the probability that a randomly selected smartphone is at most 50.8 years old.
c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Solution 6.10

Let X = the age of a randomly selected smartphone user. We are told that $\mu = 36.9$ and $\sigma = 13.9$, so $X \sim N(36.9, 13.9)$.

a. $P(23 < x < 64.7)$ = normalcdf(23, 64.7, 36.9, 13.9) = **0.8186**.

b. $P(x \leq 50.8)$ = normalcdf(-10^99, 50.8, 36.9, 13.9) = **0.8413.**

c. The 80^{th} percentile is given by: invNorm(0.80, 36.9, 13.9) = **48.6.**
Thus, 80% of the smartphone users in the U.S. are 48.6 years old or less.

Try It Σ

6.10 Use the information in **Example 6.10** to answer the following questions.

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in is less than 27 years old?

Example 6.11

There are approximately one billion smartphone users in the world today. In the United States the ages of smartphone users approximately follow a normal distribution with mean and standard deviation of 36.9 years and 13.9 years respectively. Use this to answer the following questions, rounding answers to one decimal place:

a. Calculate the interquartile range (IQR) of the distribution.
b. What is the cutoff for the top 40% of ages for smartphone users?

Solution 6.11

a. Recall that $IQR = Q_3 - Q_1$

Since Q_3 is the 75^{th} percentile, we have Q_3 = invNorm(0.75, 36.9, 13.9) = 46.2754.
Similarly, Q_1 is the 25^{th} percentile, so Q_1 = invNorm(0.25,36.9,13.9) = 27.5246.

Thus, $IQR = Q_3 - Q_1 = 46.2754 - 27.5246 = 18.7508$, which we round to **18.8**.

b. We want to find k so that $P(x \geq k) = 0.40$. I.e. 0.40 is the area to the *right* of k, and so the area to the left of k will be $1 - 0.40 = 0.60$. In other words, we are looking for the 60^{th} percentile. This value is given by k = invNorm(0.60, 36.9, 13.9) = 40.4215. Rounding, we get $\boldsymbol{k = 40.4}$.

So, forty percent of the ages are at least 40.4 years.

Try It Σ

6.11 Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean μ = 81 points and standard deviation σ = 15 points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

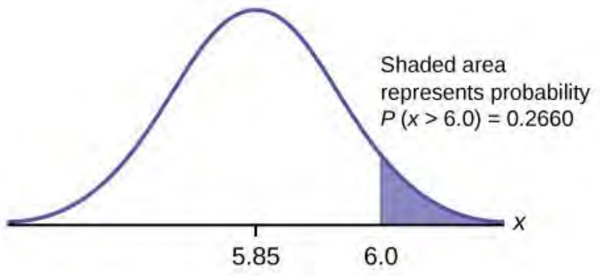
Example 6.12

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b. The middle 20% of mandarin oranges from this farm have diameters between _____and _____.
c. Find the 90th percentile for the diameters of mandarin oranges, and interpret.

Solution 6.12

a. P($x > 6$) = normalcdf(6, 10^99, 5.85, 0.24) = 0.2660.



b. We want to find the cutoffs for the middle 20% of the data. Since 1 – 0.20 = 0.80, the tails of the graph of the normal distribution will each have an area of 0.40. Thus, the cutoffs we want will be the 40th and 60th percentiles

The 40th percentile is: invNorm(0.40, 5.85, 0.24) = 5.79 cm.
The 60th percentile is: invNorm(0.60, 5.85, 0.24) = 5.91 cm

Thus, the middle 20% of diameters are between 5.79 cm and 5.91 cm.

c. The 90th percentile is invNorm(0.90, 5.85, 0.24) = 6.16 cm.
Thus, ninety percent of mandarin oranges have a diameter that is at most 6.15 cm.

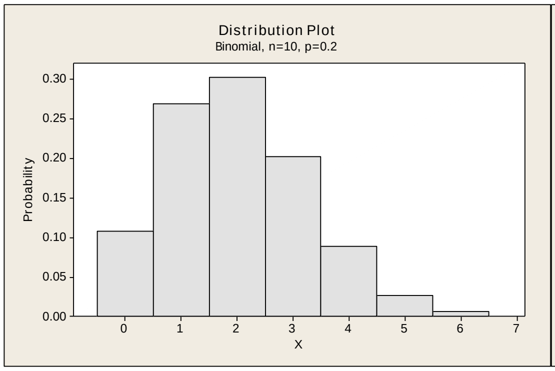
Try It Σ

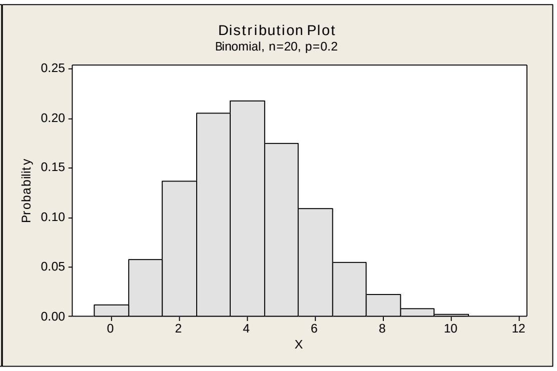
6.12 Using the information from **Example 6.12**, answer the following:

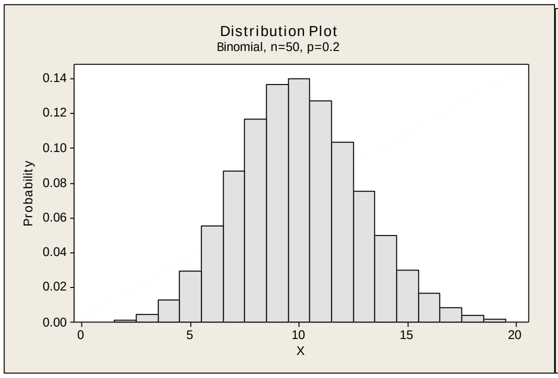
a. The middle 45% of mandarin oranges from this farm are between ___and ___.
b. Find the 16th percentile and interpret it in a complete sentence.

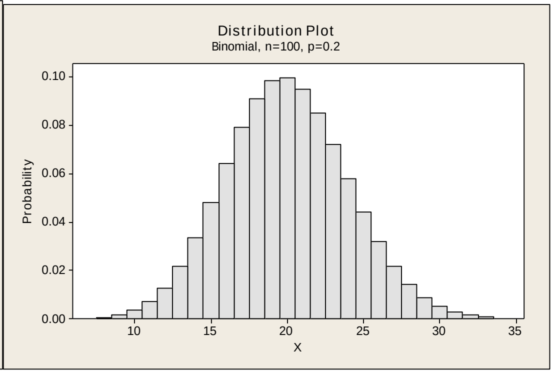
Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was an important application of the normal distribution. Binomial probabilities with a small value for *n* (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of *n*, you had to use the binomial formula, which could be very complicated and time consuming. Using the **normal approximation to the binomial** distribution simplified the process. The graphs below are for four different binomial distributions with varying sample sizes:









We see that as the sample size increases, the histogram for the distribution takes on a distinctive bell shape, which means that a normal distribution will provide a good approximation.

Recall the necessary conditions for a binomial distribution:

- There is a fixed number *n* of identical, independent trials
- The only outcomes for a given 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 we write $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the parameters

n and p should satisfy the inequalities $np > 10$ and $n(1 - p) > 10$. Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$.

In order to get the best approximation, we add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the **continuity correction factor** and is used in the following example.

Example 6.13

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that **at least 150** favor a charter school.
b. Find the probability that **at most 160** favor a charter school.
c. Find the probability that **more than 155** favor a charter school.
d. Find the probability that **fewer than 147** favor a charter school.
e. Find the probability that **exactly 175** favor a charter school.

Solution 6.13

Let X = the number that favor a charter school for grades K through 5. Then X follows a binomial distribution with $n = 300$ and $p = 0.53$; i.e. $X \sim B(300, 0.53)$. Since $np > 10$ and $nq > 10$, we can use the normal approximation to the binomial. The mean and standard deviation of the distribution are $\mu = np = 159$ and $\sigma = \sqrt{300(.53)(.47)} = 8.6447$.

Thus, the random variable for the corresponding normal distribution is $Y \sim N(159, 8.6447)$.

For part a, we want $P(X \geq 150)$, which means that $x = 150$ is included; this means in the histogram, we want the rectangle that is centered at $x = 150$; and the left endpoint of this rectangle is 149.5. So we must find the area to the right of $x = 149.5$, and so we write:

$$P(X \geq 150) \approx P(Y \geq 149.5) = \text{normalcdf}(149.5, 10\text{^}99, 159, 8.6447) = \mathbf{0.8641}.$$

Similarly, for part b, we $P(X \leq 160)$, which means we want to include $X = 160$. In the histogram, the rectangle for this value has a right endpoint of 160.6, and so

$$P(X \leq 160) \approx P(Y \leq 160.5) = \text{normalcdf}(0, 160.5, 159, 8.6447) = \mathbf{0.5689.}$$

For part c, we want to **exclude 155,** so

$$P(X > 155) \approx P(y > 155.5) = \text{normalcdf}(155.5, 10\text{^}99, 159, 8.6447) = \mathbf{0.6572}.$$

For part d, we want to **exclude 147** so

$$P(X < 147) \approx P(Y < 146.5) = \text{normalcdf}(0, 146.5, 159, 8.6447) = \mathbf{0.0741.}$$

For part e, $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5)$
$= \text{normalcdf}(174.5, 175.5, 159, 8.6447) =$ **0.0083**.

Modern calculators and computer software allow us to easily binomial probabilities for any values of n and p, making the normal approximation to the binomial distribution obsolete for calculation purposes. However, in Chapters 7, 8 and 9, we will see that this approximation is important for theoretical reasons.

For the previous example, the probabilities can of course be calculated using the binomcdf and binompdf functions in the TI-84 calculator, with $n = 300$ and $p = 0.53$. Compare the binomial and normal distribution answers:

a. $P(X \geq 150) = 1 - \text{binomcdf}(300, 0.53, 149) = 0.8641$

b. $P(X \leq 160) = \text{binomcdf}(300, 0.53, 160) = 0.5684$

c. $P(X > 155) = 1 - \text{binomcdf}(300, 0.53, 155) = 0.6576$

d. $P(X < 147) = \text{binomcdf}(300, 0.53, 146) = 0.0742$

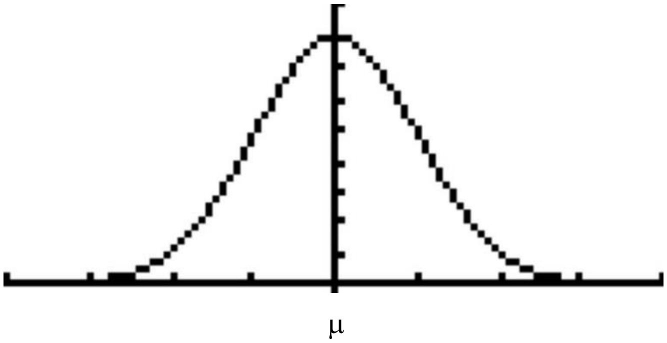
e. $P(X = 175) = \text{binompdf}(300, 0.53, 175) = 0.0083$

Try It Σ

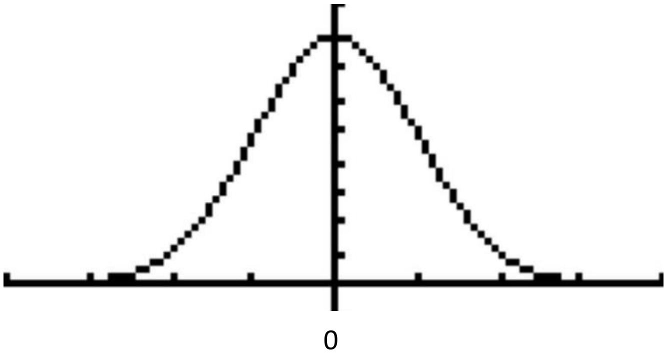
6.13 In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

KEY TERMS

Normal Distribution. A random variable with pdf $f(x) = \dfrac{e^{-(x-\mu)^2/2\sigma}}{\sigma\sqrt{2\pi}}$, where μ is the mean of the distribution and σ is the standard deviation. The graph of this pdf is a bell-shaped curve that is perfectly symmetric about the line $x = \mu$. Notation: $X \sim N(\mu, \sigma)$.



Standard Normal Distribution: A normal distribution with $\mu = 0$ and $\sigma = 1$. I.e., this is a continuous random variable $X \sim N(0, 1)$; since this is the distribution obtained by replacing data values by their respective z-scores, it is often noted as $Z \sim N(0, 1)$.



z-score: If x is any data value, then its z-score is $z = \dfrac{x - \mu}{\sigma}$. The z-score allows us to compare data that are normally distributed but scaled differently. If this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0,1)$.

FORMULA REVIEW

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

For $Z \sim N(0, 1)$:

$P(Z < k)$ = normalcdf(-10^99, k, 0, 1)

$P(Z > k)$ = normalcdf(k, 10^99, 0, 1)

$P(k_1 < Z < k_2)$ = normalcdf(k, k_2, 0, 1)

z = invnorm(probability to left, 0, 1)

For $X \sim N(\mu, \sigma)$:

$P(X < k)$ = normalcdf(-10^99, k, μ, σ)

$P(X > k)$ = normalcdf(k, 10^99, μ, σ)

$P(k_1 < X < k_2)$ = normalcdf(k, k_2, μ, σ)

k = invnorm(probability to left, μ, σ)

EXERCISES FOR CHAPTER 6

1. A bottling plant produces bottles of water whose volumes are normally distributed with a mean of 12.05 fluid ounces and a standard deviation of 0.01 ounces. Describe the random variable X symbolically: $X \sim$ ____.

2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

3. Given that $X \sim N(1, 2)$, $\sigma =$ ____.

4. A company manufactures rubber balls. The diameters of the balls are approximately normally distributed with mean 12 cm and a standard deviation of 0.2 cm. Describe the random variable X symbolically: $X =$ ____.

5. Given that $X \sim N(-4, 1)$, what is the median?

6. Given that $X \sim N(3, 5)$, $\sigma =$ ____.

7. Given that $X \sim N(-2, 1)$, $\mu =$ ____.

8. What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

9. What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

10. Find the z-score of $x = -2$ if X ~ N(0, 1.5)

11. Find the z-score of $x = 7$, if X ~ N(5, 1.2)

12. Suppose $X \sim N(-1, 2)$. What is the z-score of $x = 2$?

13. Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

14. Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

15. Suppose a normal distribution has a mean of 6 and a standard deviation of 1.5.
What is the z-score of $x = 5.5$?

16. Suppose $X \sim N(2, 6)$. What value of x has a z-score of $z = 3$?

17. Suppose $X \sim N(9, 5)$. What value of x has a z-score of -0.5?

18. Suppose $X \sim N(2, 3)$. What value of x has a z-score of -0.67?

19. Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

20. Suppose $X \sim N(4, 2)$. What value of x is 2.8 standard deviations to the right of the mean?

21. Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

22. Suppose $X \sim N(15, 2.5)$. Graph the normal distribution.

23. Suppose $X \sim N(-2, 1)$. Graph the normal distribution

24. Suppose $X \sim N(60, 5.9)$. Graph the normal distribution

25. Suppose $X \sim N(520, 35)$. Graph the normal distribution

26. In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is _____standard deviations to the ______(right or left) of the mean.

27. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ______standard deviations to the ______(right or left) of the mean.

28. In a normal distribution, $x = -2$ and $z = 6$. This tells you that $x = -2$ is ______standard deviations to the ______(right or left) of the mean.

29. In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ___standard deviations to the ______(right or left) of the mean.

30. In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ______standard deviations to the ______(right or left) of the mean.

31. About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

32. About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

33. About what percent of x values lie between the second and third standard deviations (both sides)?

34. Suppose $X \sim N(15, 3)$. Between what x values does the middle 68.26% of the data lie?

35. Suppose $X \sim N(-3, 1)$. Between what x values does the middle 95.44% of the data lie?

36. Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

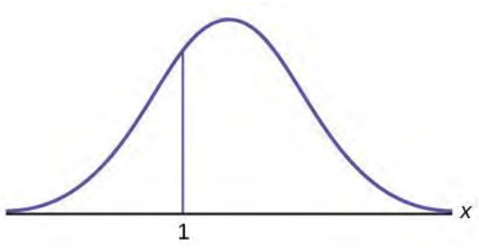
37. About what percent of x values lie between the mean and three standard deviations?

38. About what percent of x values lie between the mean and one standard deviation?

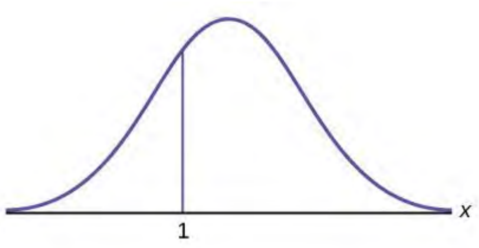
39. About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

40. About what percent of x values lie between the first and third standard deviations (both sides)?

41. How would you represent the area to the left of one as a probability statement?

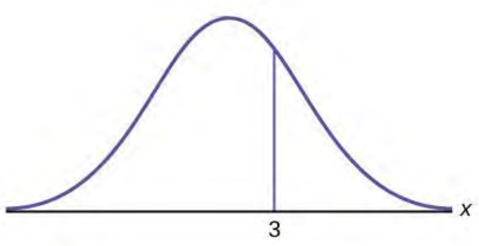


42. Represent the area to the right of $x = 1$ as a probability statement.

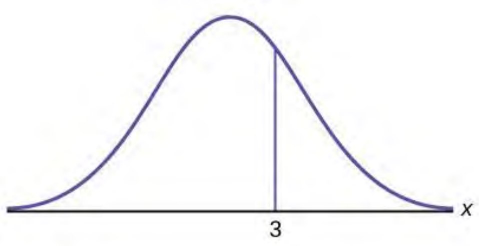


43. In a normal distribution, is $P(x < 1)$ equal to $P(x \leq 1)$? Why or why not?

44. How would you represent the area to the left of $x = 3$ as a probability statement?



45. Represent the area to the right of $x = 3$ as a probability statement.



46. Draw and find $P(z < 2.16)$.

47. Draw and find $P(z < -2.78)$.

48. If the area to the left of *x* in a normal distribution is 0.123, what is the area to the right of *x*?

49. If the area to the right of *x* in a normal distribution is 0.543, what is the area to the left of *x*?

Use the following information to answer the next four exercises: $X \sim N(54, 8)$

50. Find the probability that $x > 56$.

51. Find the probability that $x < 30$.

52. Find the 80th percentile.

53. Find the 60th percentile.

54. Given that $X \sim N(6, 2)$, find the probability that *x* is between 3 and 9.

55. Given that $X \sim N(–3, 4)$, find the probability that *x* is between 1 and 4.

56. Given that $X \sim N(4, 5)$, find the maximum value of *x* for the bottom quartile.

General Word Problems

57. The life span of gaming consoles is normally distributed with mean of 7.5 years and a standard deviation of 1.3 years. We are interested in the length of time a console lasts. The guarantee period is 5 years.

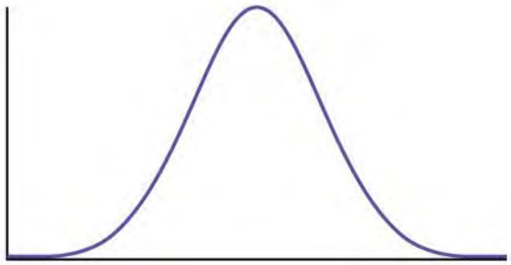
a. Draw the shape of the distribution, label and scale the axes

b. Describe the random variable symbolically: $X \sim$ ____ (____,____)

c. What is the likelihood that a gaming console will break down during the guarantee period?

d. Find the probability that a console last more than 10 years. Is it unusual?

58. The life span of gaming consoles is normally distributed with mean of 7.5 years and a standard deviation of 1.3 years. We are interested in the length of time a console lasts. Find the 70th percentile of the distribution for the time a gaming console lasts.



59. The life span of car tires is normally distributed with a mean of 5 years and a standard deviation of 0.85 years.

a. Draw the shape of the distribution, label and scale the axes

b. Describe the random variable symbolically: $X \sim$ ____ (____,____)

c. What is the likelihood that a car tire will last longer than 7 years?

d. Find the probability that a tire will last between 4 to 5 years.

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

60. What is the median recovery time?

61. What is the *z*-score for a patient who takes ten days to recover?

62. The length of time to find it takes to find a parking space at 9 a.m. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution.

III. The data cannot follow the normal distribution.

a. I only b. II only c. III only d. I, II, and III

63. The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, μ = 79 inches and standard deviation σ = 3.89 inches. For each of the following heights, calculate the *z*-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a *z*-score of 3.5, would you believe him? Explain your answer.

64. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ =125 and standard deviation σ = 14.

a. Calculate the *z*-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

65. Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If X = a systolic blood pressure score then $X \sim N(125, 14)$.

a. Which of the following statements are correct?
   i. Kyle's systolic blood pressure is 175.
   ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

b. Calculate Kyle's blood pressure.

66. Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed, so $X \sim N(10.2, 0.8)$. Calculate and interpret the *z*-scores that correspond to the following weights:

a. $x = 11$ kg b. $x = 7.9$ kg c. $x = 12.2$ kg

67. In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

a. Calculate the *z*-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?

c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Use the following information to answer the next two exercises:

The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

68. What is the probability of spending more than two days in recovery?

69. What is the 90th percentile for recovery times?

Use the following information to answer the next three exercises:
The length of time it takes to find a parking space at 9 am follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

70. Based upon the given information, would you be surprised if it took less than one minute to find a parking space?

a. Yes b. No c. Unable to determine

71. Find the probability that it takes at least eight minutes to find a parking space.

72. Seventy percent of the time, what is the most number of minutes needed to find a parking space?

73. According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let *X* = height of the individual.

a. *X* ~ ___(________, ____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write the probability in terms of the random variable *x.*
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
d. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

74. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let *X* = IQ of an individual.

a. *X* ~ ___(________, ____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

75. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let *X* = percent of fat calories.

a. *X* ~ ___(________, ____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

76. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If X = distance in feet for a fly ball, then X ~ ___(_______,___).
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph, label the horizontal axis X and shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

77. In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. X ~ ___(_______,___)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

78. In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X = number of votes for President Clinton for an election district.

a. State the approximate distribution of X.

b. Is the value 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.

e. Find the third quartile for votes for President Clinton.

79. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable X.
b. X ~ ___(_______,___)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

80. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable *X*.
b. *X* ~ ___(_______,___)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under ____.
e. The middle 80% of her laps are from ___seconds to ____seconds.

81. Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his *z*-score is zero.
c. Anita is in the 70th percentile of students at her college.

82. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let *X* represent the number of defective cars in the sample. What can we say about *X* in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

83. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

a. There is about a 68% chance that the number of heads will be somewhere between _ and __.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

84. A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of
$n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

Business Applications

85. A machine that fills quart bottles with fruit juice is normally distributed with a mean of 31.6 oz per bottle, with a standard deviation of 1.2 oz.

a. What is the probability that the amount of juice in a bottle is less than 1 quart?

b. What is the probability that the amount of juice in a bottle is at least 2 ounces more than a quart?

c. What are the largest and smallest amounts dispensed by the middle 50%?

86. Customers at a certain cosmetic store spend an average of $45.80, with a standard deviation of $1275. Assume the amount spent is normally distributed.

a. What is the probability that a customer spends more than $60?

b. What is the probability that a customer spends between $30 and $40?

c. What are the largest and smallest amounts spend by the middle 60%?

d. Find the 70th percentile.

87. Suppose in a quarter, a group of 25 mutual funds return of 2.8% with a standard deviation of 4.8%. Assume the returns are normally distributed.

a. What percent of funds would you expect to have returns of 7% or more?

b. What percent of funds would you expect to have returns of 2% or less?

c. What percent of funds would you expect to have returns between 5% and 10%?

d. Find the 10th percentile.

88. Human resource departments look at job satisfaction scores. Let's say job satisfaction scores are normally distributed with a mean of 90 and a standard deviation of 12.

a. What percent of scores are less than 70?

b. What percent of scores are between 100 and 120?

c. Human resource departments are concerned with job satisfaction scores that drops below a specific score. What score is considered unusual?

REFERENCES

6.1 The Standard Normal Distribution

"Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

"2012 College-Bound Seniors Total Group Profile Report." CollegeBoard 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

"Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

"List of stadiums by capacity." Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

6.2 Using the Normal Distribution

"Naegele's rule." Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013). "403: NUMMI." Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio- archives/episode/403/nummi (accessed May 14, 2013).

"Scratch-Off Lottery Ticket Playing Tips." WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

"Smart Phone Users, By The Numbers." Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

7 | THE CENTRAL LIMIT THEOREM

Figure 7.1 If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Introduction

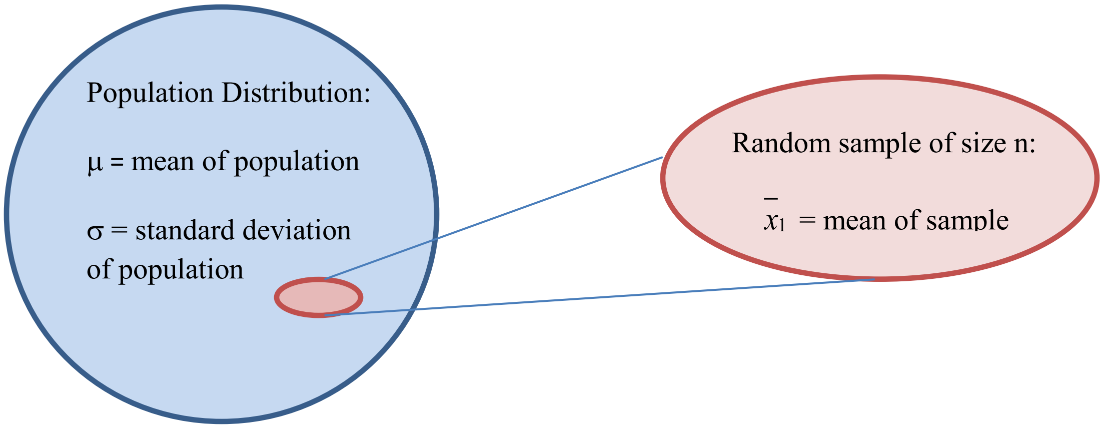
Chapter Objectives

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

- Recognize problems involving the Central Limit Theorem.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for means and proportions.
- Apply and interpret the Central Limit Theorem for sums.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of Statistics. There are two alternative forms of the theorem, and both describe the center, spread and shape of a certain *sampling distribution.* In general, the sampling distribution of a statistic (such as a sample mean or sample proportion) is the distribution of values of that statistic when all possible samples of the same size are taken from the same population. Sampling distributions form the foundation for almost all methods in inferential statistics, and the Central Limit Theorem allows us to explicitly describe the sampling distribution for a sample mean $\bar{x}$ and the sampling distribution for a sample proportion p .

If we select multiple random samples of size n from a population and calculate the mean for each sample, then the sample mean $\bar{x}$ will vary from sample to sample.



That is, $\bar{x}$ will be a random variable, and so it has a probability distribution. If we were able to collect *all* possible random samples of size n from the population and calculate $\bar{x}$ for each sample, then the resulting distribution is called the **sampling distribution for the mean**.

What would this distribution look like? Well, if n is reasonably large, we would expect most of the sample means $\bar{x}$ to be pretty close to the true population mean, μ. Thus, we would expect the mean of all sample means to be equal to μ. There will of course be some variation among the sample means; however, we would expect most of the differences between a sample mean and population mean to be fairly small, with large deviations quite rare. Thus, we would expect the pdf of the distribution to approach zero as we move away from the center. Finally, since a sample mean $\bar{x}$ is as likely to underestimate as it is to underestimate the true population mean, we would expect the positive and negative deviations from the mean to occur with about the same proportion. Thus, we would expect the distribution to be symmetric.

Sampling distribution

Collection of all possible random samples of size n

When we put these three observations together, would expect the sampling distribution for $\bar{x}$ to be a bell-shaped, symmetric distribution – so it is not unreasonable to think that the sampling distribution is a normal distribution. Moreover, the mean of this sampling distribution is the mean, μ of the population from which we are sampling. The Central Limit Theorem validates our intuition; specifically, the CLT states that if we collect samples of "sufficiently large" size *n* then the sampling distribution will be approximately normal. Similarly, the sampling distribution for $\hat{p}$ will be formed and the CLT can be used for this sampling distribution too. Another version of the theorem says that if we again collect samples of size *n* that are "large enough," calculate the *sum* of each sample, then the sampling distribution for the sum will again be approximately normal.

In any case, it does not matter what the distribution of the original population is, or whether we even know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, *n*, that is required in order to be "large enough" depends on the original population from which the samples are drawn. If we are sampling from a normal distribution, the sampling distribution will exhibit a bell-shape even for small *n.* But if we are sampling from a skewed distribution, we will need the sample size to be at least 30.

7.1 | The Central Limit Theorem for Sample Means and Sample Proportions

Suppose X is a random variable with mean μ and standard deviation σ. Suppose that we select random samples of size n and denote the corresponding sample means as $\overline{X}$. Then we denote the mean and standard deviation of the sampling distribution for $\overline{X}$ as $\mu_{\overline{X}}$ and $\sigma_{\overline{X}}$ respectively.

Central Limit Theorem for Sample Means

Suppose X is a random variable with mean μ and standard deviation σ. Suppose that we select random samples of size n. Then the following are true:

i) The mean of the sampling distribution for $\overline{X}$ is $\mu_{\overline{X}} = \mu$.

ii) The standard deviation of the sampling distribution for $\overline{X}$ is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

iii) For $n \geq 30$, the sampling distribution for the sampling distribution for $\overline{X}$ is approximately normal.

The first two parts of the theorem tell us that the sampling distribution has the same mean as the original distribution and a variance that equals the original variance divided by the square root of the sample size. To state the third part more precisely, if we draw random samples of size n, the distribution of the random variable $\overline{X}$ approaches a normal distribution as the sample size n increases. We can summarize all three parts of the theorem succinctly by saying:

$$X \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Note that the standard deviation for the sampling distribution, $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, is often called the **standard error of the mean**. Note also that as n grows larger, the standard error gets smaller. That is as the sample size gets larger there is less variation among the sample means – this reinforces our intuition that larger samples give more reliable results when we use a sample mean $\overline{x}$ to estimate μ.

Also recall the **Law of Large numbers,** which says that if you take samples of larger and larger size from any population, then the sample mean $\overline{x}$ tends to get closer and closer to the population mean μ. We now see that this is a direct result of the Central Limit Theorem: we know that as n gets larger and larger, the sample means follow a normal distribution with mean μ. Moreover, as the sample size increases, the standard deviation for the sampling distribution decreases. So n becomes large, the sample means $\overline{x}$ really will get closer and closer to the population mean μ.

The fact that the sampling distribution of $\overline{x}$ is approximately normal means that we can use the techniques learned in Chapter 6 for the calculations. However, we must be careful to use the

standard deviation for the *sampling distribution*. For example, if we want to calculate the *z*-score for a particular sample mean $\bar{x}$, then we would use the formula:

$$z = \frac{\bar{x} - \mu_{\bar{X}}}{\left(\sigma_{\bar{X}}\right)} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

And when we calculate probabilities, we will make sure that we will always write the probability statement in terms of the correct random variable.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the probabilities involving the sampling distribution for $\bar{x}$, we go to 2nd DISTR and use the **normalcdf** function:

$$P(a < \bar{x} < b) = \text{normalcdf}(a,\ b,\ \mu,\ \frac{\sigma}{\sqrt{n}})$$

Here, μ = the mean of the distribution we are sampling, σ = the standard deviation, and *n* is the sample size.

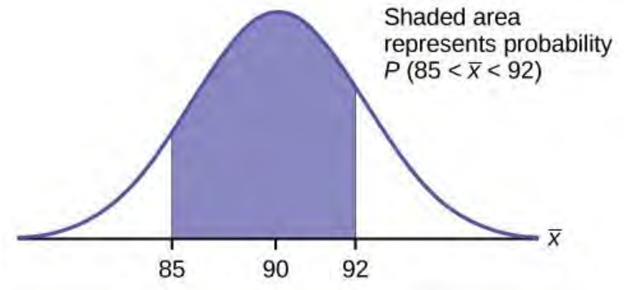
Example 7.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 36$ are drawn randomly from the population.

a. Find the probability that a randomly selected sample mean is between 85 and 92.
b. Find the value that is two standard deviations above the expected value of the sample mean.

Solution 7.1

a. This question asks you to find a probability involving the **sample mean**: $P(85 < \bar{x} < 92)$.
We first draw a graph:



We from the CLT that the distribution of $\bar{x}$ is approximately normal, with mean and standard deviation $\mu_{\bar{X}} = \mu = 90$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5$. Since the distribution is normal, we can use the **normalcdf** function in the calculator to find the probability:

$P(85 < \bar{x} < 92) = \text{normalcdf}(85, 92, 90, 2.5) =$ **0.7654.**

b. To find the value of $\bar{x}$ that is two standard deviations above the expected value 90, use the formula:

$$\bar{x} = \mu_{\overline{X}} + 2\sigma_{\overline{X}} = 90 + 2(2.5) = \mathbf{95}.$$

The value of $\bar{x}$ that is two standard deviations above the expected value is 95.

Try It Σ

7.1 An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size $n = 30$ are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

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 two hours and a standard deviation of 0.5 hours. A sample of size $n = 50$ is drawn randomly from the population. Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

Solution 7.2

Let X = the time, in hours, it takes to play one soccer match.
We are asked to find a probability involving the **sample mean** time (in hours) needed to play one soccer match.

We are told that $\mu = 2$, $\sigma = .50$ and $n = 50$. From the Central Limit Theorem, the sampling distribution of $\bar{x}$ is approximately normal, with mean $\mu_{\overline{X}} = \mu = 2$ and $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{.5}{\sqrt{50}} \approx 0.0707$.

I.e. $X \sim N(2, 0.0645)$. Thus, $P(1.8 < \bar{x} < 2.3)$ = normalcdf(1.8, 2.3, 2, 0.0707) = **0.9977**.

Try It Σ

7.2 The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

We can also calculate percentiles in the sampling distribution using the techniques from Chapter 6:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate percentiles in the sampling distribution for $\bar{x}$, go to 2nd DISTR and select the **invNorm** function:

The p-th percentile is $k =$ invNorm(p, μ, $\frac{\sigma}{\sqrt{n}}$)

Here, μ = the mean of the distribution we are sampling, σ = the standard deviation, and n is the sample size.

Example 7.3

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

Solution 7.3

a. The mean of the sampling distribution is $\mu_{\bar{X}} = \mu = 34$ and the standard deviation of the sampling distribution is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$.

b. From the Central Limit Theorem, we would expect that the sampling distribution will be approximately normal.

c. The probability that the sample mean age is more than 30 is given by P($X > 30$) = normalcdf(30, 10^99, 34, 1.5) = **0.9962**.

d. Let k = the 95th percentile; then k = invNorm(0.95, 34, 1.5) = 36.467, or about 36.5 years old.

Try It Σ

7.3 In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Example 7.4

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. We select a random sample of 60 tablet users and measure the mean time spent on each app.

a. What are the mean and standard deviation for the sample mean number of minutes for app engagement by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution 7.4

a. The mean of the sampling distribution is $\mu_{\overline{X}} = \mu = 8.2$ minutes and the standard deviation of the sampling distribution is $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.1291$.

b. Recall that the standard error of the mean is just $\sigma_{\overline{X}}$, the standard deviation of the sampling distribution for $\overline{x}$. If we select many samples of size 60, this statistic describes the spread of the sample means about the population mean.

c. The 90th percentile is given by invNorm(.90, 82, .1291) = **8.37 minutes**.
This indicates that 90% of average app engagement times are less than 8.37 minutes.

d. $P(8 < \overline{x} < 8.5) =$ normalcdf(8, 8.5, 8.2, 0.1291) = **0.9293**.

Try It Σ

7.4

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34$, $\overline{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Suppose X is a random variable with population proportion p. Suppose that we select random samples of size n and denote the corresponding sample proportion as $\hat{p}$. Then we denote the mean and standard deviation of the sampling distribution for $\hat{p}$ as $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ respectively.

Central Limit Theorem for Sample Proportions

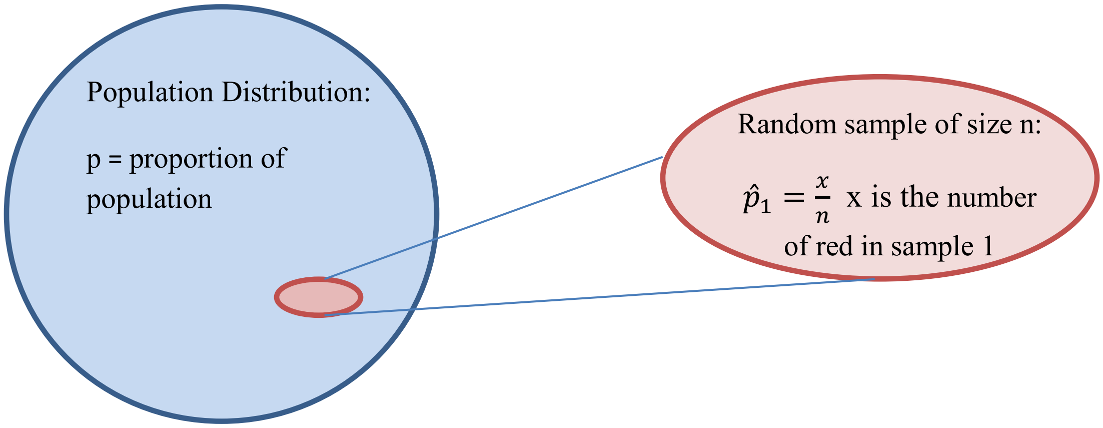
Suppose X is a random variable with population proportion p. Suppose that we select random samples of size n. Then the following are true:

i) The mean of the sampling distribution for $\hat{p}$ is $\mu_{\hat{p}} = p$.

ii) The standard deviation of the sampling distribution for $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$.

iii) The sampling distribution for the sampling distribution for $\hat{p}$ is approximately normal when n is large and p is not too near 0 or 1. $X \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$

Let's look at how CLT works for a sampling distribution for sample proportion. We open up a regular bag of colored candy and notice that there is 23 red candies out of 100 pieces. Therefore $p = \frac{23}{100} = 0.23$. Now we will take samples of size 9, which the notation is n = 9. Note the properties are approximate when the population is finite and no more than 10% of the population is included in the sample. In each sample we find the proportion of red candies, $\hat{p} = \frac{x}{9}$. x is the count of red candies in that sample of 9 candies.

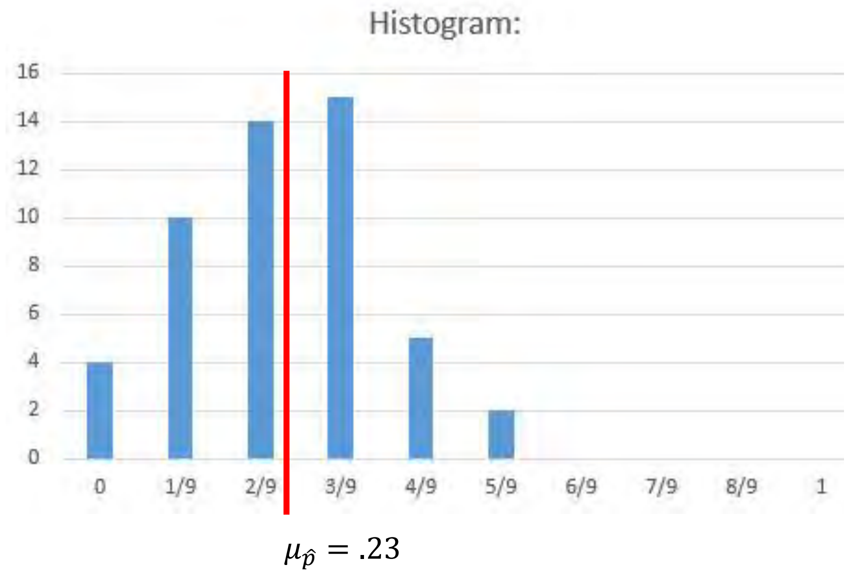

Population Distribution:
p = proportion of population
Random sample of size n:
$\hat{p}_1 = \frac{x}{n}$ x is the number of red in sample 1

The list of 50 sample proportions of size 9:

0	$^{1}/_{9}$	$^{3}/_{9}$	$^{3}/_{9}$	$^{2}/_{9}$	$^{2}/_{9}$	$^{1}/_{9}$	$^{4}/_{9}$	$^{2}/_{9}$	$^{3}/_{9}$
$^{2}/_{9}$	$^{3}/_{9}$	$^{2}/_{9}$	0	$^{1}/_{9}$	$^{3}/_{9}$	$^{1}/_{9}$	$^{3}/_{9}$	$^{4}/_{9}$	$^{2}/_{9}$
$^{3}/_{9}$	$^{2}/_{9}$	$^{1}/_{9}$	$^{5}/_{9}$	$^{2}/_{9}$	$^{3}/_{9}$	0	$^{4}/_{9}$	$^{2}/_{9}$	$^{3}/_{9}$
$^{2}/_{9}$	$^{1}/_{9}$	$^{3}/_{9}$	$^{4}/_{9}$	$^{1}/_{9}$	$^{4}/_{9}$	$^{3}/_{9}$	$^{2}/_{9}$	$^{1}/_{9}$	$^{1}/_{9}$
$^{5}/_{9}$	$^{3}/_{9}$	$^{2}/_{9}$	$^{1}/_{9}$	0	$^{3}/_{9}$	$^{2}/_{9}$	$^{3}/_{9}$	$^{3}/_{9}$	$^{2}/_{9}$

If we look at the frequency distribution of this situation, we can see that the histogram is symmetric around p = .23. We take the mean of all the $\hat{p}$ drawn which we denote $\mu_{\hat{p}}$. CLT states that $\mu_{\hat{p}}$ = p and $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$.

$\hat{p}$	f
0	4
$\frac{1}{9}$	10
$\frac{2}{9}$	14
$\frac{3}{9}$	15
$\frac{4}{9}$	5
$\frac{5}{9}$	2
$\frac{6}{9}$	0
$\frac{7}{9}$	0
$\frac{8}{9}$	0
1	0



Example 7.5

Suppose that 68% of students like donuts. If you randomly survey 40 students, what is the probability your survey says at least 80% like donuts?

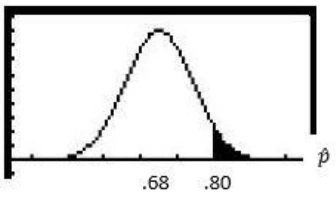
Solution 7.5

P($\hat{p} \geq .80$) = ? Since we are asking the probability of a sample proportion, we are using central limit theorem for proportions.

$\mu_{\hat{p}} = p = .68$

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.68 \cdot .32}{40}} = 0.0738$

P($\hat{p} \geq .80$) = normalcdf(.80, 1E99, .68, .0738) = 0.052



Example 7.6

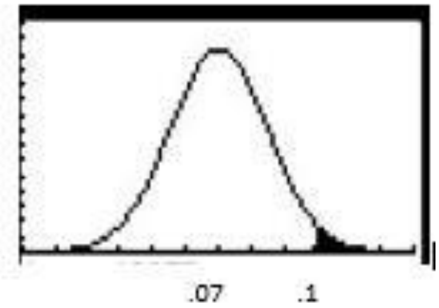
Based on past experience, a bank claims that 7% of the people who receive student loans will not make the payments on time. The bank recently approved 300 student loans.

a. What is the mean of the proportion of students in this group who may not make timely payments?
b. What is the standard deviation of the proportion of students in this group who may not make timely payments?
c. What's the probability that over 10% of these students will not make the payments on time?

Solution 7.6

P($\hat{p}$ > .10) = ? Since we are asking the probability of a sample proportion, we are using central limit theorem for proportions.

a. $\mu_{\hat{p}} = p = .07$
b. $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.07 \cdot .93}{300}} = 0.015$
c. P($\hat{p} \geq .10$) = normalcdf(.10, 1E99, .07, .015) = 0.0228



7.2 | 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 that we select samples of size n. Recall that ΣX represents the sum of all data values from the sample. Again, ΣX is a random variable, as this sum will vary from sample to sample; thus we can discuss the sampling distribution for ΣX as well. We denote the standard deviation of this sampling distribution as $\mu_{\Sigma X}$ and $\sigma_{\Sigma X}$ respectively. Then we have the following variant of the CLT:

Central Limit Theorem for Sample Sums

Suppose X is a random variable with mean μ and standard deviation σ. Suppose that we select random samples of size n. Then the following are true:

i) The mean of the sampling distribution for ΣX is $\mu_{\sum X} = n\mu$.

ii) The standard deviation of the sampling distribution for ΣX is $\sigma_{\sum X} = \sigma\sqrt{n}$.

iii) For $n \geq 30$, the sampling distribution for the sampling distribution for ΣX is approximately normal.

In other words, if we draw random samples of size n, the random variable ΣX consisting of sums tends to be **normally distributed** and $\Sigma X \sim N(n\mu, \sqrt{n}\,\sigma)$. Again this means that we can apply the techniques from Chapter 6 for calculating z-scores and probabilities involving sample sums.

Example 7.7

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.

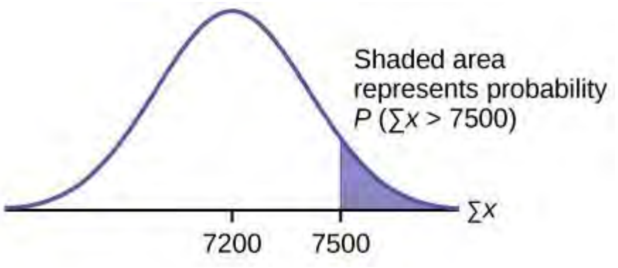
a. Find the probability that the sum of the 80 values is more than 7,500.
b. Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution 7.7

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 of 80 values. Since $\mu = 90$, $\sigma = 15$, and $n = 80$, $\Sigma\, X \sim N((80)(90),\ \sqrt{80}\,(15))$.
That is, $\Sigma\, X \sim N(7200, 134.16)$.

a. Find $P(\Sigma x > 7{,}500)$. It is helpful to draw a graph:



To calculate the probability we use the normalcdf function:

$$P(\Sigma x > 7{,}500) = \text{normalcdf}(7500,\ 10\text{^}99,\ 7200,\ 134.16) = \mathbf{0.0127.}$$

b. Find Σx where $z = 1.5$. This sum is 1.5 standard deviations above the mean, so

$$\Sigma x = 7200 + 1.5(134.16) = \mathbf{7401.2.}$$

Try It Σ

7.5 An unknown distribution has a mean of 45 and a standard deviation of eight. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

Example 7.8

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. The sample of size is 50.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?

b. Find the probability that the sum of the ages is between 1,500 and 1,800 years.

c. Find the 80th percentile for the sum of the 50 ages.

Solution 7.8

a. $\mu_{\Sigma x} = n\mu = 50(34) = 1{,}700$ and $\sigma_{\Sigma x} = \sqrt{n}\ \sigma = \sqrt{50}\ (15) = 106.01$

The distribution is normal for sums by the central limit theorem.

b. $P(1500 < \Sigma x < 1800) = \text{normalcdf}\,(1500, 1800, 1700, 106.01 = 0.7974$

c. Let k = the 80th percentile. Then $k = \text{invNorm}(0.80,\ 1700,\ 106.01) = 1{,}789.3$

Try It Σ

7.6 In a recent study reported Oct.29, 2012 on the Flurry Blog, the mean age of tablet users is 35 years. Suppose the standard deviation is ten years. The sample size is 39.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?

b. Find the probability that the sum of the ages is between 1,400 and 1,500 years.

c. Find the 90th percentile for the sum of the 39 ages.

Example 7.9

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Suppose that we select a sample of size 70.

a. What are the mean and standard deviation for the sums?

b. Find the 95th percentile for the sum of the sample. Interpret this value in a complete sentence.

c. Find the probability that the total time for the sample is at least ten hours.

Solution 7.9

a. $\mu_{\Sigma x} = n\mu = 70(8.2) = 574$ and $\sigma_{\Sigma x} = \sqrt{n}\,\sigma = \sqrt{70}\ (1) = 8.37$ minutes

b. Let k = the 95th percentile. Then k = invNorm (0.95, 574, 8.37) = 587.76 minutes
Ninety five percent of the app engagement times are at most 587.76 minutes.

c. First note that 10 hours = 600 minutes.
$P(\Sigma x \geq 600)$ = normalcdf(600, 10^99, 574, 8.37) = **0.0009**.

Try It Σ

7.7 The mean number of minutes for app engagement by a table use is 8.2 minutes. Suppose the standard deviation is one minute. We select a random sample size of 70.

a. What is the probability that the total time for the sample is between seven hours and ten hours? What does this mean in context of the problem?

b. Find the 84th and 16th percentiles for the sum of the sample. Interpret these values in context.

7.3 | Using the Central Limit Theorem

It is important to understand when to use the **central limit theorem**. If you are being asked to find the probability of the mean, use the CLT for the mean. 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 means and sums.

NOTE

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.**

Example 7.10

A study involving stress is conducted among the students on a college campus. The stress scores follow a uniform distribution with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

a. The probability that the **mean stress score** for the 75 students is less than two.
b. The 90th percentile for the **mean stress score** for the 75 students.
c. The probability that the **total of the 75 stress scores** is less than 200.
d. The 90th percentile for the **total stress score** for the 75 students. Let X = one stress score.

Solution 7.10

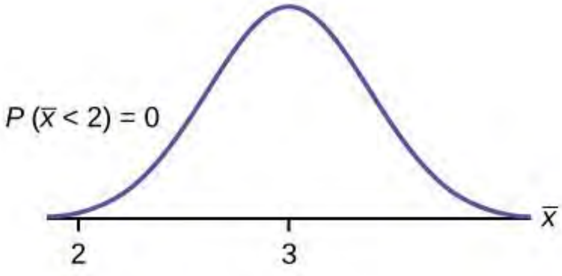
Problems a and b ask you to find a probability or a percentile for a **mean**.
Problems c and d ask you to find a probability or a percentile for a **total or sum**.
The sample size is $n = 75$.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$; thus the mean and standard deviation are $\mu = (1 + 4)/2 = 3$ and $\sigma = (5 - 1)/\sqrt{12} \approx 1.155$. (See Ch. 5).

Since $\mu =$ 3 and $\sigma = \sqrt{12} \approx 1.155$, $\bar{X} \sim N[3,\ 1.155 / \sqrt{75}\,]$; that is, $\bar{X} \sim N(3, 0.1334)$.

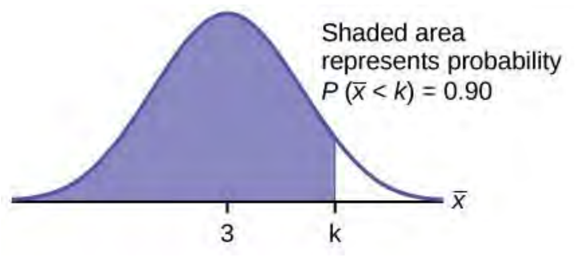
And $\Sigma X \sim N[(75)(3),\ \sqrt{75}\ (1.155)]$; that is, $\Sigma X \sim N(225, 10.002)$.

a. We want to find $P(\bar{x} < 2)$. It is useful to draw a graph.



Since $\overline{X} \sim N(3, 0.1334)$, $P(\bar{x} < 2) =$ normalcdf(-10^99, 2, 3, 1.155/$\sqrt{75}$) = 0.
The probability that the mean stress score is less than 2 is virtually zero.

b. Let k = the 90th percentile; that is, $P(\bar{x} < k) = 0.90$. Again, a graph is helpful:



Using the calculator, and using the parameters from part a, we see that the 90th percentile for the mean of 75 scores is k = invNorm(.90, 3, 0.13334), which is about 3.2. This tells us that, among all samples of size 75, 90% of the mean stress scores are at most 3.2, and that 10% are at least 3.2.

c. Since $\Sigma X \sim N(225, 10.002)$, $P(\Sigma x < 200)$ = normalcdf(-10^99, 200, 225, 10.002) = **0.0062**.

d. Let k = the 90th percentile; that is, $P(\bar{x} < k) = 0.90$.
Using the calculator, and using the parameters from part c, we see that the 90th percentile for the mean of 75 scores is k = invNorm(.90, 225, 10.002), which is about 237.8. This tells us that, among all samples of size 75, 90% of the totals of the stress scores are at most 237.8, and 10% are at 237.8 or more.

Try It Σ

7.8 Use the information in Example 7.8, but use a sample size of 55 to answer the following questions:

a. Find $P(\bar{x} < 7)$.
b. Find $P(\Sigma x > 170)$.
c. Find the 80th percentile for the mean of 55 scores.
d. Find the 85th percentile for the sum of 55 scores.

Example 7.11

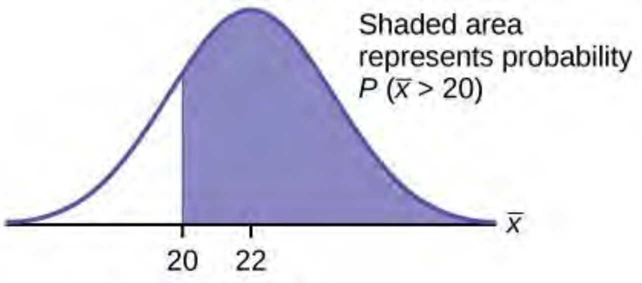
Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows 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, and find the following probabilities:

a. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes.
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 parts a and b are different.
d. Find the 95th percentile for the mean excess time for samples of 80 customers who exceed their basic contract time allowances

Solution 7.11

a. This is asking us to find $P(\bar{x} > 20)$. Since the mean of the exponential distribution is $\mu = 22$, the standard deviation is also $\sigma = 22$. By the CLT, $\bar{X} \sim N(22, \frac{22}{\sqrt{80}})$.



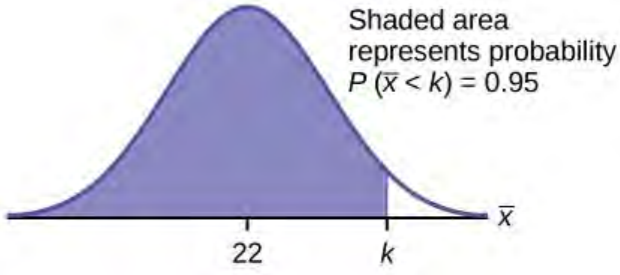
Using the calculator, $P(\bar{x} > 20)$ = normalcdf(20, 10^99, 22, $\frac{22}{\sqrt{80}}$) = **0.7919**.

b. Find $P(x > 20)$. Remember to use the exponential distribution for an **individual**.

And $X \sim \text{Exp}\left(\frac{1}{22}\right)$; thus $P(x > 20) = e^{-\frac{1}{22}(20)} = 0.4029$.

c. The probabilities are not equal because they involve different random variables, and hence use different distributions.

d. To find the 95th percentile for the mean excess time for samples of 80 customers, we again use the CLT, and the fact that $\bar{X} \sim N(22, \frac{22}{\sqrt{80}})$. Let k be the 95th percentile; we first draw a graph.



This value is given by k = invNorm(.95, 22, $\frac{22}{\sqrt{80}}$) = 26.05. Thus, about ninety five percent of such samples would have a mean under 26 minutes; only five percent of such samples would have a mean above 26 minutes.

Try It

7.9 Use the information in **Example 7.9**, but change the sample size to 144.

a. Find P(20 < $\bar{x}$ < 30).

b. Find P(Σx $\geq$ 3,000).

c. Find the 75th percentile for the sample mean excess time of 144 customers.

d. Find the 85th percentile for the sum of 144 excess times used by customers.

Example 7.12

A study was done about violence against prostitutes and the symptoms of the posttraumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

a. In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?

b. Is it likely that the mean age of the sample group could be more than 50 years? Explain.

c. Find the 95th percentile for the mean age of 65 randomly selected prostitutes and interpret.

Solution 7.12

Since the mean of the distribution is $\mu = 30.9$ and the standard deviation is $\sigma = 9$, we use the CLT to get $\bar{X} \sim N(30.9, \frac{9}{\sqrt{25}}) = N(30.9,\ 1.8)$.

a. P($\bar{x}$ < 35) = normalcdf(-10^99, 35, 30.9, 1.8) = **0.9886**.

b. Calculate P($\bar{x}$ > 50) = normalcdf(50, 10^99, 30.9, 1.8) $1.35 \text{ x } 10^{-26} \approx 0$. For this sample group, it is virtually impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.

c. The 95th percentile = invNorm(0.95, 30.9, $\frac{9}{\sqrt{65}}$) = 32.7. This indicates that in 95% all samples of size 65, the average age of the prostitutes in the sample was less 32.7 years.

Try It

7.11 According to Boeing data, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69.0 inches and a standard deviation of 2.8 inches.

a. What mean doorway height would allow 95% of men to enter the aircraft without bending?

b. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?

c. For engineers designing the 757, which result is more relevant: the height from part a or part b?

KEY TERMS

Central Limit Theorem for Sample Means: Given a random variable X with known mean μ and standard deviation σ, we select random samples of size n and examine the sampling distribution of the sample mean $\bar{x}$. This sampling distribution has the following characteristics:

i) The mean of the sampling distribution for $\bar{X}$ is $\mu_{\bar{X}} = \mu$.

ii) The standard deviation of the sampling distribution for $\bar{X}$ is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

iii) For $n \geq 30$, the sampling distribution for the sampling distribution for $\bar{X}$ is approximately normal. That is, $\bar{X}$ ~N(μ, $\frac{\sigma}{\sqrt{n}}$).

Central Limit Theorem for Sample Proportions: Suppose X is a random variable with population proportion p. Suppose that we select random samples of size n. The sampling distribution has the following characteristics:

i) The mean of the sampling distribution for $\hat{p}$ is $\mu_{\hat{p}} = p$.

ii) The standard deviation of the sampling distribution for $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$.

iii) The sampling distribution for the sampling distribution for $\hat{p}$ is approximately normal. $X \sim N(\mu_{\hat{p}}, \sigma_{\hat{p}})$

Central Limit Theorem for Sample Sums: Given a random variable X with known mean μ and standard deviation σ, we select random samples of size n and examine the sampling distribution of the sample sum, ΣX. This sampling distribution has the following characteristics:

i) The mean of the sampling distribution for ΣX is $\mu_{\Sigma X} = n\mu$.

ii) The standard deviation of the sampling distribution for ΣX is $\sigma_{\Sigma X} = \sigma\sqrt{n}$.

iii) For $n \geq 30$, the sampling distribution for the sampling distribution for ΣX is approximately normal. That is, $\Sigma X \sim N(n\mu, \sqrt{n}\,\sigma)$.

Sampling Distribution: The sampling distribution of a statistic (such as a sample mean or sample proportion) is the distribution of values of that statistic when all possible samples of the same size are taken from the same population.

Standard Error of the Mean: The standard deviation of the sampling distribution for $\bar{x}$: $\frac{\sigma}{\sqrt{n}}$.

FORMULA REVIEW

z-score: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$

For $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

$P(\bar{X} < k)$ = normalcdf(-10^99, k, μ, $\frac{\sigma}{\sqrt{n}}$)

$P(\bar{X} > k)$ = normalcdf(k, 10^99, μ, $\frac{\sigma}{\sqrt{n}}$)

$P(k_1 < \bar{X} < k_2)$ = normalcdf(k, k_2, μ, $\frac{\sigma}{\sqrt{n}}$)

k = invnorm(probability to left, μ, $\frac{\sigma}{\sqrt{n}}$)

For $\hat{p} \sim N(p, \sqrt{\frac{pq}{n}})$

$P(\hat{p} < k)$ = normalcdf(-10^99, k, p, $\sqrt{\frac{pq}{n}}$)

$P(\hat{p} > k)$ = normalcdf(k, 10^99, p, $\sqrt{\frac{pq}{n}}$)

$P(k_1 < \hat{p} < k_2)$ = normalcdf(k, k_2, p, $\sqrt{\frac{pq}{n}}$)

k = invnorm(probability to left, p, $\sqrt{\frac{pq}{n}}$)

Exercises for Chapter 7

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review; assume that X is normally distributed. Let X-bar be the random variable representing the mean time to complete 16 reviews. Assume that the 16 reviews represent a random set of reviews.

1. What is the mean, standard deviation, and sample size?

2. Describe the distributions:

a. $\overline{X} \sim$ ___(___,___)   b. $\Sigma X \sim$ ___(___,___)

3. Find the probability that a single review will require between 3.5 and 4.25 hours, and sketch the graph.

4. Find the probability that the mean time for a month's reviews will require between 3.5 and 4.25 hours, and sketch the graph.

5. What causes the probabilities in Exercise 7.3 and Exercise 7.4 to be different?

6. Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.

Use the following information to answer the next four exercises: An unknown distribution has a mean of 80 and a standard deviation of 12. A sample size of 95 is drawn randomly from the population.

7. Find the probability that the sum of the 95 values is greater than 7,650.

8. Find the probability that the sum of the 95 values is less than 7,400.

9. Find the sum that is two standard deviations above the mean of the sums.

10. Find the sum that is 1.5 standard deviations below the mean of the sums.

Use the following information to answer the next five exercises: The distribution of results from a cholesterol test has a mean of 180 and a standard deviation of 20. A sample size of 40 is drawn randomly.

11. Find the probability that the sum of the 40 values is greater than 7,500.

12. Find the probability that the sum of the 40 values is less than 7,000.

13. Find the sum that is one standard deviation above the mean of the sums.

14. Find the sum that is 1.5 standard deviations below the mean of the sums.

15. Find the percentage of sums between 1.5 standard deviations below the mean of the sums and one standard deviation above the mean of the sums.

Use the following information to answer the next six exercises: A researcher measures the amount of sugar in several cans of the same soda. The mean is 39.01 with a standard deviation of 0.5. The researcher randomly selects a sample of 100.

16. Find the probability that the sum of the 100 values is greater than 3,910.

17. Find the probability that the sum of the 100 values is less than 3,900.

18. Find the probability that the sum of the 100 values falls between 3900 and 3910.

19. Find the sum that has a z-score of -2.5.

20. Find the sum that has a z–score of 0.5.

21. Find the probability that the sums will fall between the z = -2 and z = 1.

Use the following information to answer the next four exercises: An unknown distribution has a mean 12 and a standard deviation of one. A sample size of 25 is taken. Let X = the object of interest.

22. What is the mean of ΣX?

23. What is the standard deviation of ΣX?

24. What is P(Σx = 290)?

25. What is P(Σx > 290)?

26. True or False: Only the sums of normal distributions are also normal distributions.

27. In order for the sums of a distribution to approach a normal distribution, what must be true?

28. What three things must you know about a distribution to find the probability of sums?

29. An unknown distribution has a mean of 25 and a standard deviation of six. Let X = one object from this distribution. What is the sample size if the standard deviation of ΣX is 42?

30. An unknown distribution has a mean of 19 and a standard deviation of 20. Let X = the object of interest. What is the sample size if the mean of ΣX is 15,200?

Use the following information to answer the next three exercises. A market researcher analyzes how many electronics devices customers buy in a single purchase. The distribution has a mean of three with a standard deviation of 0.7. She samples 400 customers.

31. What is the z-score for Σx = 840?

32. What is the z-score for Σx = 1,186?

33. What is P(Σx < 1,186)?

Use the following information to answer the next three exercises: An unknown distribution has a mean of 100, a standard deviation of 100, and a sample size of 100. Let X = one object of interest.

34. What is the mean of ΣX?

35. What is the standard deviation of ΣX?

36. What is P($\Sigma x > 9{,}000$)?

Use the following information to answer the next ten exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is selected.

37. a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deviation?
b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

38. Draw the graph for Exercise 7.37c.

39. Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

40. Draw the graph for Exercise 7.39.

41. Find the 90th percentile for the mean weight for the 100 weights.

42. Draw the graph from Exercise 7.41.

43. a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?
b. Find P($\Sigma x < 2{,}450$).

44. Draw the graph for Exercise 7.43.

45. Find the 90th percentile for the total weight of the 100 weights.

46. Draw the graph for Exercise 7.45.

Use the following information to answer the next eight exercises: The length of time a particular smartphone's battery lasts follows an exponential distribution with a mean of ten months. A sample of 64 of these smartphones is taken.

47. a. What is the standard deviation for the population distribution?
b. What is the parameter μ?

48. What is the distribution for the length of time that one battery lasts?

49. Suppose that a sample of 64 batteries is selected. What is the distribution for the mean length of time that the batteries last?

50. What is the distribution for the total length of time that 64 batteries last?

51. Find the probability that the sample mean for a sample of 64 batteries will be between 7 months and 11 months.

52. Find the 80th percentile for the total length of time that 64 batteries will last.

53. Find the *IQR* for the mean amount of time for a sample of 64 batteries.

54. Find the middle 80% for the total amount of time that 64 batteries will last.

Use the following information to answer the next six exercises:
A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is selected.

55. Find $P(\Sigma x > 420)$.

56. Find the 90th percentile for the sums.

57. Find the 15th percentile for the sums.

58. Find the first quartile for the sums.

59. Find the third quartile for the sums.

60. Find the 80th percentile for the sums.

61. Previously, De Anza Statistics students estimated that the amount of change daytime Statistics students carry is exponentially distributed with a mean of $0.88. Suppose that we randomly pick 25 daytime statistics students.

a. In words, $X =$ _____
b. $X \sim$ ___(____,___)
c. In words, $\overline{X}$ = ___
d. $\overline{X} \sim$ ___(____,___)
e. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
f. Find the probability that the average of the 25 students was between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
g. Explain why there is a difference in part e and part f.

62. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. Describe the distribution for $\overline{x}$, the average distance in feet, for 49 fly balls.
b. What is the probability that the 49 balls travel a mean distance of less than 240 feet?
c. Find the 80th percentile of the distribution of the average of 49 fly balls.

63. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X= _____

b. In words, $\overline{X}$ = ___

c. $\overline{X}$ ~ ___(___,___)

d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.

e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

64. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. We select a random sample of 49 races, and let x-bar represent the average time for the 49 running times.

a. $\overline{X}$ ~ ___(___,___)

b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.

c. Find the 80th percentile for the average of these 49 marathons.

d. Find the median of the average running times.

65. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X= _____

b. X~ _______

c. In words, $\overline{X}$ = _____

d. Describe the distribution: $\overline{X}$ ~ ___(___,___)

e. Find the first quartile for the average song length.

f. The IQR(interquartile range) for the average song length is from ____–______.

66. In 1940 the average size of a U.S. farm was 174 acres, with a standard deviation of 55 acres. Suppose we randomly survey 38 farms from 1940. What is the probability that the average size was more than 200 acres?

67. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, $\mu_{\overline{X}}$ is approximately equal to μ.

b. When the sample size is large, $\overline{X}$ is approximately normally distributed.

c. When the sample size is large, the standard deviation of the distribution for $\overline{X}$ is approximately the same as the standard deviation for X.

68. The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation 10. Suppose that 16 individuals are randomly chosen.

a. Find the probability that the mean percent of fat calories for this group of 16 exceeds 40.

b. Find the first quartile for the average percent of fat calories.

69. The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.

a. In words, $X =$ _____

b. In words, $\overline{X}$ = ___

c. $\overline{X} \sim$ __(__,___)

d. How is it possible for the standard deviation to be greater than the average?

e. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to$2,200?

70. Which of the following is NOT TRUE about the distribution of sample means?

a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the *x*-axis.
d. The curve is skewed to the right.

71. 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. The distribution to use for the average cost of gasoline for the 16 gas stations is:

a. $\overline{X} \sim N(4.59, 0.10)$

b. $\overline{X} \sim N(4.59, \frac{0.10}{\sqrt{16}})$

c. $\overline{X} \sim N(4.59, \frac{0.10}{16})$

d. $\overline{X} \sim N(4.59, \frac{16}{\sqrt{0.10}})$

72. The attention span of a two-year-old is exponentially distributed with a mean of about eight minutes. Suppose we randomly survey 60 two-year-olds.

a. In words, $X =$ _____

b. $X \sim$ ____(___,___)

c. In words, $\overline{X}$ = ___

d. $\overline{X} \sim$ __(___,___)

e. Before doing any calculations, which do you think will be higher? Explain why.
   i. The probability that an individual attention span is less than ten minutes.
   ii. The probability that the average attention span for the 60 children is less than ten minutes?

f. Calculate the probabilities in part e.

Use the following information to answer the next two exercises: The time to wait for a particular rural bus is distributed uniformly from zero to 75 minutes. One hundred riders are randomly sampled to learn how long they waited.

73. The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

a. 315.0 minutes b. 40.3 minutes c. 38.5 minutes d. 65.2 minutes.

74. Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

a. yes b. no c. There is not enough information to answer this.

Use the following to answer the next two exercises: 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.

75. What's the approximate probability that the average price for 16 gas stations is over $4.69?

a. almost zero b. 0.1587 c. 0.0943 d. unknown

76. Find the probability that the average price for 30 gas stations is less than $4.55.

a. 0.6554 b. 0.3446 c. 0.0142 d. 0.9858 e. 0

77. Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through five. A simple random sample of 300 is surveyed. Calculate following using the normal approximation to the binomial distribution.

a. Find the probability that less than 100 favor a charter school for grades K through 5.
b. Find the probability that 170 or more favor a charter school for grades K through 5.
c. Find the probability that no more than 140 favor a charter school for grades K through 5.
d. Find the probability that there are fewer than 130 that favor a charter school for grades K through 5.
e. Find the probability that exactly 150 favor a charter school for grades K through 5.

78. Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities.

a. Find the probability that Janice is the driver at most 20 days.
b. Find the probability that Roberta is the driver more than 16 days.
c. Find the probability that Barbara drives exactly 24 of those 96 days.

79. Salaries for teachers in a particular elementary school district are normally distributed with a mean of $44,000 and a standard deviation of $6,500. We randomly survey ten teachers from that district.

a. Find the 90th percentile for an individual teacher's salary.
b. Find the 90th percentile for the average teacher's salary.

80. The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

a. In words, X = ______
b. In words, $\overline{X}$ = ____
c. $\overline{X}$ ~ ___(___,___)
d. Is it likely that an individual stayed more than five days in the hospital? Why or why not?
e. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
h. Which is more likely:
   i. An individual stayed more than five days.
   ii. the average stay of 80 women was more than five days.

81. NeverReady Batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

82. Men have an average weight of 172 pounds with a standard deviation of 29 pounds.

a. Find the probability that 20 randomly selected men will have a total weight exceeding 3600 lbs.
b. If 20 men have a sum weight greater than 3500 lbs, then their total weight exceeds the safety limits for water taxis. Based on (a), is this a safety concern? Explain.

83. Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

84. A typical adult has an average IQ score of 105 with a standard deviation of 20. If 20 randomly selected adults are given an IQ test, what is the probability that the sample mean scores will be between 85 and 125 points?

85. Certain coins have an average weight of 5.201 grams with a standard deviation of 0.065 g. If a vending machine is designed to accept coins whose weights range from 5.111 g to 5.291 g, what is the expected number of rejected coins when 280 randomly selected coins are inserted into the machine?

86. Suppose that 54% of students register two weeks before the start of the semester. If you randomly survey 70 students, what is the probability your survey says at most 20% register two weeks before the start of the semester?

87. According to USA Today article, 56% of 18 to 24 year olds voted for Hilary Clinton in the 2016 election. If we randomly survey thirty 18 to 24 year olds, what is the probability that those selected at least 60% voted for Hilary Clinton?

88. A truckload of watermelons arrive at a packing facility. A random sample of 200 is selected and examined for bruises and other defects. The truckload will be rejected if more than 5% of the sample has defects. Suppose that in fact 8% of the watermelons on the truck have defects. What's the probability that the shipment will not be rejected?

REFERENCES

7.1 The Central Limit Theorem for Sample Means (Averages)

Baran, Daya. "20 Percent of Americans Have Never Used Email."WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013). Data from the United States Department of Agriculture.

7.2 The Central Limit Theorem for Sums

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

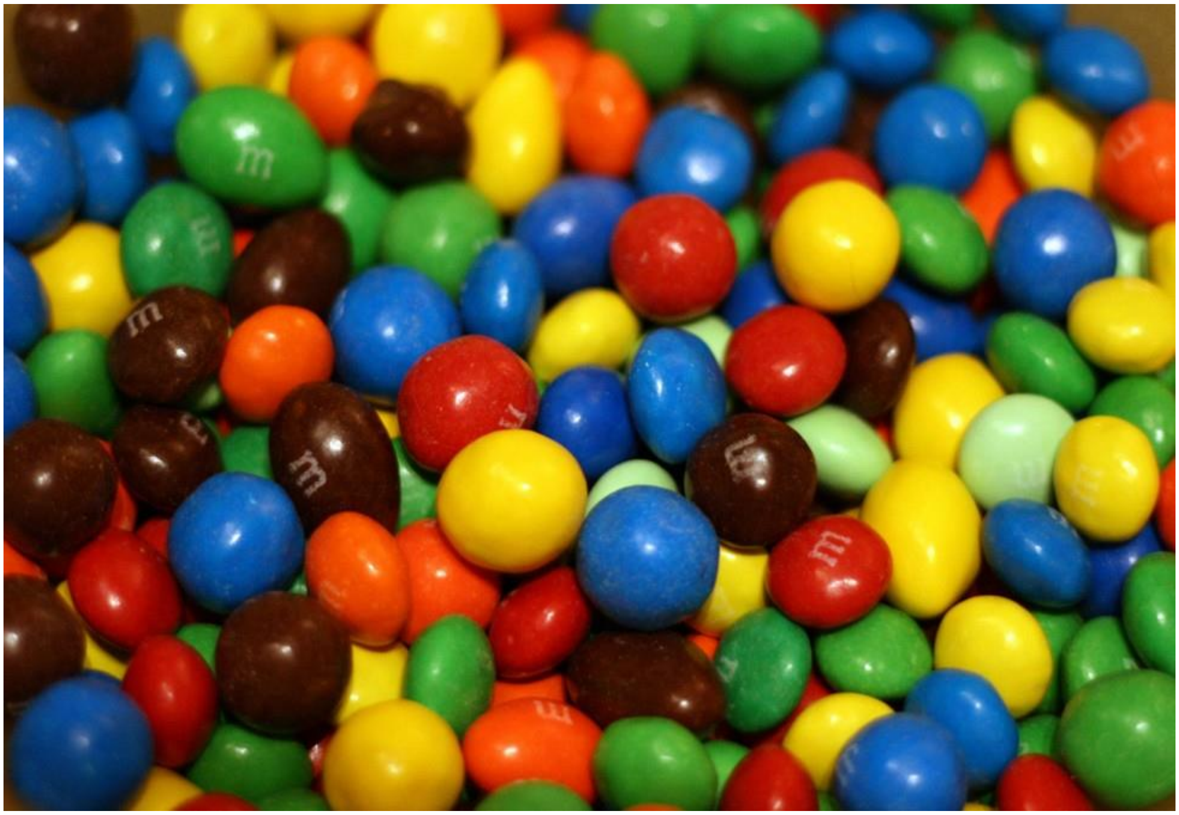
7.3 Using the Central Limit Theorem

Data from the Wall Street Journal.

"National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Castillo, Walbert, et al. "How We Voted - by Age, Education, Race and Sexual Orientation." *USA Today*, Gannett Satellite Information Network, 9 Nov. 2016, college.usatoday.com/2016/11/09/how-we-voted-by-age-education-race-and-sexual-orientation/ (accessed July 17, 2018).

8 | CONFIDENCE INTERVALS



Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (Credit: comedy_nose/flickr)

Chapter Objectives

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

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's t distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a **point estimate** of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a **point estimate** for the true proportion.

8.1 INTRODUCTION

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. Therefore, **point estimate** is a sample statistic which is a starting point to estimate the population parameter.

Example 8.1

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x} = 4$ songs, and the sample standard deviation, $s = 0.1$. You would use to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$ is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

After calculating point estimates, we construct interval estimates, called **confidence intervals**. Intervals in mathematics represent a range of values that the variable can be. Recall in Algebra, a variable that is between two values can be written in two ways:

Inequality notation: $3 < x < 10$, here we are saying x is between 3 and 10.

Interval notation: (3, 10)

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The **confidence interval** is likely to include an unknown population parameter between a lower bound value and upper bound value with a certain level of confidence.

Mean:	lower bound value < $\boldsymbol{\mu}$ < upper bound value
Proportion:	lower bound value < **p** < upper bound value
Standard Deviation:	lower bound value < $\boldsymbol{\sigma}$ < upper bound value

NOTE: For mean and proportion lower bound value is the point estimate – margin of error. The upper bound value is the point estimate + margin of error. The interval notation of this form is **(point estimate – margin of error, point estimate + margin of error)**

The **empirical rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. From **Example 8.1**, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ. Therefore, 0.2 is considered the **margin of error.**

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower bound value is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper bound value is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the **Example 8.1**, the unknown population mean μ is between $\bar{x} - 0.2 = 4 - 0.2 = 3.8$ (lower bound value) and $\bar{x} + 0.2 = 4 + 0.2 = 4.2$ (upper bound value).

$$3.8 \text{ songs} < \mu < 4.2 \text{ songs}$$

We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 3.8 and 4.2. The 95% confidence interval is (3.8, 4.2).

Example 8.2

Find the point estimate for the population mean, μ, for a 90% confidence interval. Assume the distribution is normally distributed with a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Solution 8.2

Point estimate for the population mean is the sample mean, $\bar{x}$. Therefore $\bar{x}$ =36 minutes is the point estimate.

Example 8.3

Find the point estimate for the population proportion, p, for a 98% confidence interval. In a survey of 11,605 parents, 4912 think that the government should subsidize the costs of computers for lower-income families (Adapted from DisneyFamily.com)

Solution 8.3

Point estimate for the population proportion is the sample proportion, $\hat{p}$. Therefore $\hat{p} = \frac{4912}{11{,}605}$ is the point estimate. $\hat{p} = \frac{4912}{11{,}605}$ can be written as a decimal or a percent.

Try It Σ

Find the point estimate for the population mean and population standard deviation, μ and σ. Assume the distribution is approximately normal. In a random sample of 8 people, the mean commute to work was 35.5 minutes and the sample standard deviation was 7.2 minutes.

8.2 | Confidence Interval for a Single Population Mean when σ known

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; $\bar{x}$~N(μ_x, $\frac{\sigma}{\sqrt{n}}$) Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for this parameter has the form:

(point estimate – margin of error, point estimate + margin of error)

Or

$$\left(\bar{x}-E, \bar{x}+E\right)$$

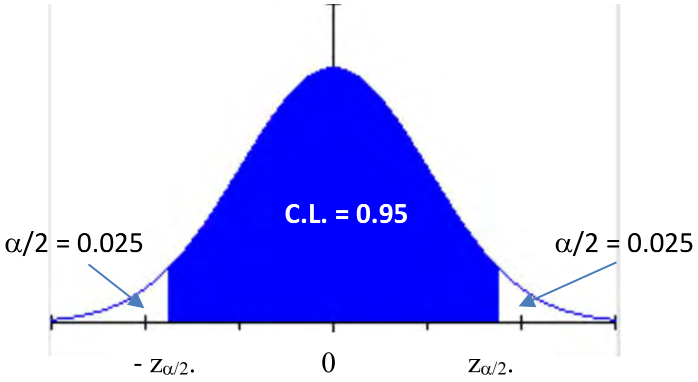
The **margin of error**, E, depends on the **confidence level**, CL, or percentage of confidence and the standard error of the mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The value that represents the level of confidence for a certain distribution is known as a **critical value.** The critical value for the Z-distribution has the notation of $z_{\alpha/2}$. Therefore, $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Critical Value, $Z_{\alpha/2}$.

The confidence level, CL, is the area in the middle of the Z-distribution (standard normal distribution). The **confidence level** is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

CL = 1 – α, so α is the area that is not in the center and is split equally between the two tails. Therefore, **α** is the probability that the interval will not contain the true population parameter. Each of the tails contains an area equal to α/2. The z-score that has an area to the right of α/2 is denoted by $z_{\alpha/2}$.

For example, when CL = 0.95, α = 0.05 and α/2 = 0.025; we write $z_{\alpha/2} = z_{0.025}$.

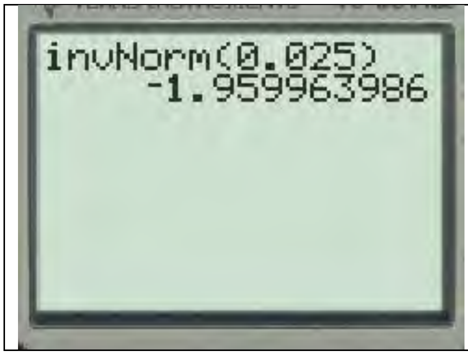


Recall in chapter 6 to find the value of z when given the area to the left, we can use the calculator to find this value.

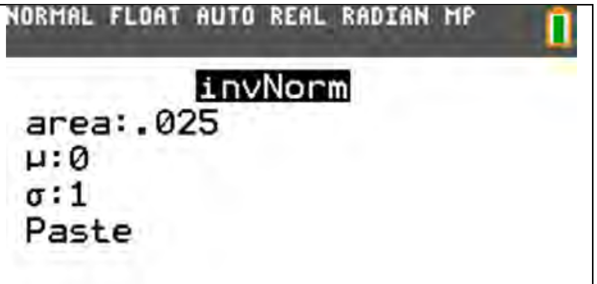
Using the TI-83, 83+, 84, 84+ Calculator

To calculate

Press 2nd Vars, choose #3 invnorm



```
invNorm(0.025)
        -1.959963986
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
           invNorm
 area:.025
 μ:0
 σ:1
 Paste
```

The left critical value is negative since it is on the left side of zero. The right critical value is the symmetric opposite of the left critical value. Round critical values to two decimal places.

Left: $-z_{\alpha/2} = -z_{0.025} = -1.96$

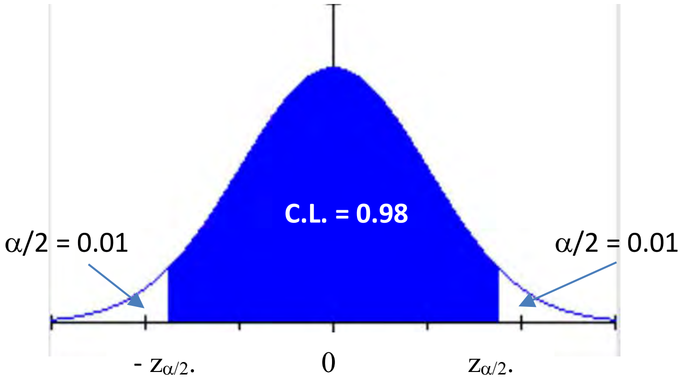
Right: $z_{\alpha/2} = z_{0.025} = 1.96$

Example 8.4

Find the critical value for the z-distribution for 98% confidence level.

Solution 8.4

C.L = .98, $\alpha = .02$, $\alpha/2 = .01$ we write $z_{\alpha/2} = z_{0.01}$.



Left: $-z_{\alpha/2} = -z_{0.01}$ = invnorm(.01) = -2.33

Right: $z_{\alpha/2} = z_{0.01} = 2.33$

Calculating the Confidence Interval for mean when σ is known

Point estimate – margin of error < μ < point estimate + margin of error

$$\left(\bar{x} - E, \bar{x} + E\right)$$

where $E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the critical value, $z_{\alpha/2}$ that corresponds to the confidence level.
- Calculate the margin of error, E.
- 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.)

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a population mean), and state the confidence interval (both endpoints). "We estimate with % confidence that the true population mean (include the context of the problem) is between and (include appropriate units)."

Example 8.5

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution 8.5

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

8.5 Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the margin of error, E.

$\bar{x} = 68$

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\sigma = 3$; $n = 36$; the confidence level is 90% (CL = 0.90) so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$\alpha = 0.05$ $z_{\alpha/2} = z_{0.05}$ = the positive value of invnorm(.05) = 1.645

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.645\left(\frac{3}{\sqrt{36}}\right) = .8225$$

$\bar{x}$ - E = 68 - 0.8225 = 67.1775

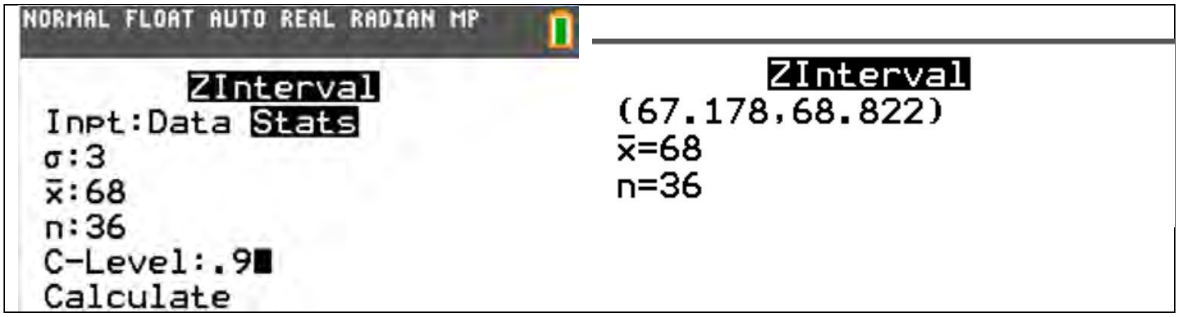
$\bar{x}$ + E = 68 + 0.83 = 68.8225

67.1775 < μ < 68.8225

The 90% confidence interval for the population mean is (67.1775, 68.8225).

8.5 Solution B

Press STAT, TESTS. #7:ZInterval.



```
NORMAL FLOAT AUTO REAL RADIAN MP
        ZInterval
 Inpt:Data Stats
 σ:3
 x̄:68
 n:36
 C-Level:.9
 Calculate
```

```
ZInterval
(67.178,68.822)
x̄=68
n=36
```

The confidence interval for the population mean is (67.178, 68.822).

Interpretation:

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level:

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Try It Σ

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes. Find a 90% confidence interval estimate for the population mean delivery time.

Example 8.6

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table 8.1 shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

Phone Model	**SAR**	**Phone Model**	**SAR**	**Phone Model**	**SAR**
iPhone 4s	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsugn Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU 575	1.26	Samsung Messenger III	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Table 8.1

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

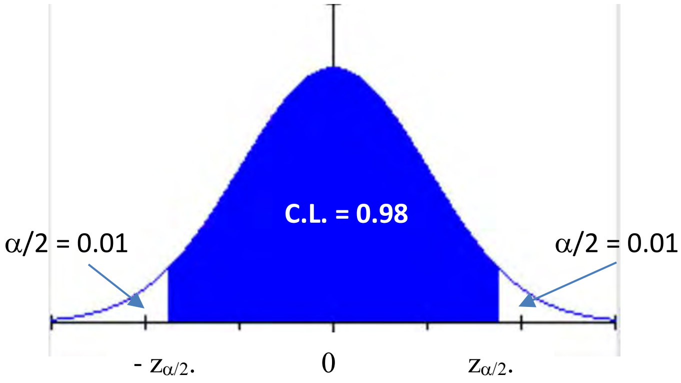
Solution 8.6 Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$\bar{x}$ can be found using 1-var Stat on the calculator (Chapter 2)

$\bar{x} = 1.024$

Next, find the margin of error, E. Because you are creating a 98% confidence interval, CL = 0.98.



Left: - $z_{\alpha/2}$ = -$z_{0.01}$ = invnorm(.01) = -2.33

Right: $z_{\alpha/2} = z_{0.01} = 2.33$

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$E = 2.33\frac{0.337}{\sqrt{30}} = 0.1434$$

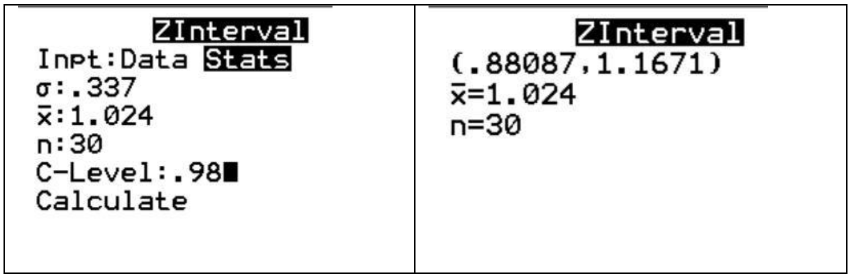
To find the 98% confidence interval, find the lower and upper bounds of the interval.

Lower bound: $\bar{x}$ – E = 1.024 – 0.1434 = 0.8806

Upper bound: $\bar{x}$ + E = 1.024 + 0.1434 = 1.1674

Solution 8.6 Solution B

Press STAT, TESTS 7: ZInterval.



```
    ZInterval
Inpt:Data Stats
σ:.337
x̄:1.024
n:30
C-Level:.98
Calculate
```

```
    ZInterval
(.88087,1.1671)
x̄=1.024
n=30
```

The confidence interval is 0.881< μ < 1.167.

Interpretation: We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8806 and 1.1674 watts per kilogram.

Try It Σ

Table 8.2 shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is σ = 0.337.

Phone Model	SAR	Phone Model	SAR
BlackBerry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optiimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Table 8.2

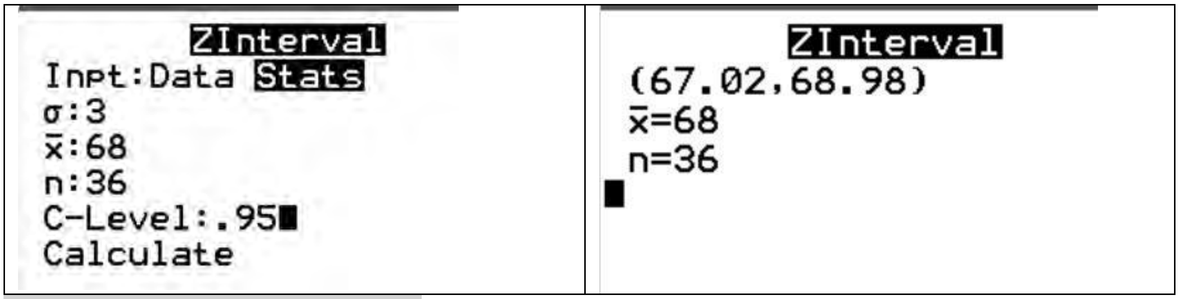
Changing the Confidence Level or Sample Size

Example 8.7

Suppose we change the original problem in **Example 8.5** by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution 8.7 using Calculator

To find the confidence interval about the mean with the population standard deviation known, we use Zinterval.



```
ZInterval
Inpt:Data Stats
σ:3
x̄:68
n:36
C-Level:.95
Calculate
```

```
ZInterval
(67.02,68.98)
x̄=68
n=36
```

E = upper bound – point estimate

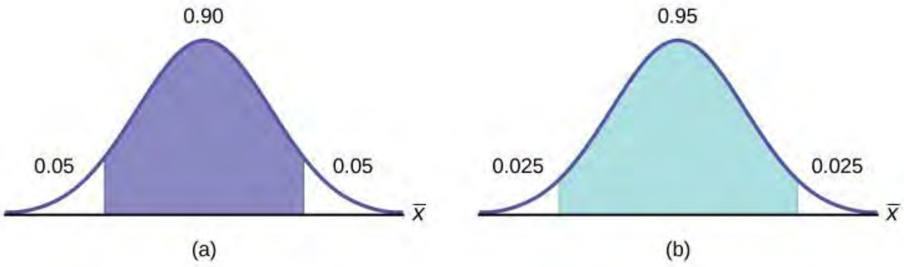
E = 68.98 – 68 = .98

Notice that the E is larger for a 95% confidence level than the 90% confidence of Example 8.5

90% Confidence Level	95% Confidence Level
E = .8225 $67.18 < \mu < 68.82$	E = .98 $67.02 < \mu < 68.98$

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the margin of error, making the confidence interval wider.
- Decreasing the confidence level decreases the margin of error, making the confidence interval narrower.

Try It Σ

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Example 8.8

Suppose we change the original problem in **Example 8.5** to see what happens to the margin of error if the sample size is changed.

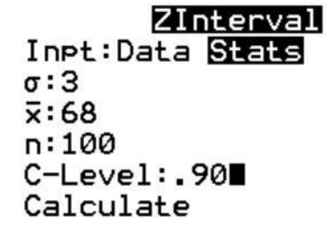
Leave everything the same except the sample size. Use the original 90% confidence level.

a. What happens to the margin of error and the confidence interval if we increase the sample size and use n = 100 instead of n = 36?

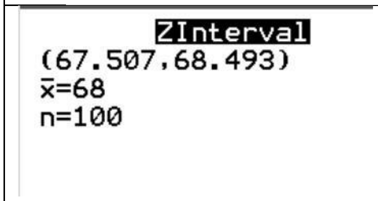
b. What happens if we decrease the sample size to n = 25 instead of n = 36?

Solution 8.8 by Calculator

a.) Zinterval

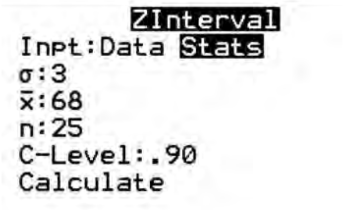


```
ZInterval
Inpt:Data Stats
σ:3
x̄:68
n:100
C-Level:.90
Calculate
```

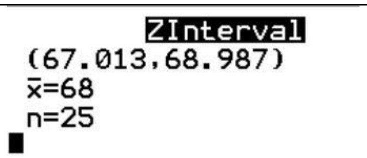


```
ZInterval
(67.507,68.493)
x̄=68
n=100
```

b.) Zinterval



```
ZInterval
Inpt:Data Stats
σ:3
x̄:68
n:25
C-Level:.90
Calculate
```



```
ZInterval
(67.013,68.987)
x̄=68
n=25
```

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the margin of error to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the margin of error to increase, making the confidence interval wider.

Try It Σ

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Calculating the Sample Size, n

If researchers desire a specific margin of error, then they can use the margin of error formula to calculate the required sample size. The margin of error formula for a population mean when the population standard deviation is known has the following formula:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

When we solve for n, the formula for sample size is as follows:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

NOTE: clue word for margin of error is the word **"within".**

Example 8.9

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

Solution 8.9

From the problem, we know that σ = 15 and E= 2.

$z = z_{0.025} = 1.96$, because the confidence level is 95%.

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2 = \left(\frac{1.96 \cdot 15}{2} \right)^2 = 216.09$$

Use n = 217: Always round the answer **UP** to the next higher integer to ensure that the sample size is large enough. Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Try It Σ

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

8.3 | Confidence Interval for a Single Population Mean when σ unknown

In practice, we rarely know the population standard deviation, σ. 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. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's t-distribution. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the **Student's t-distribution** whenever s is used as an estimate for σ.

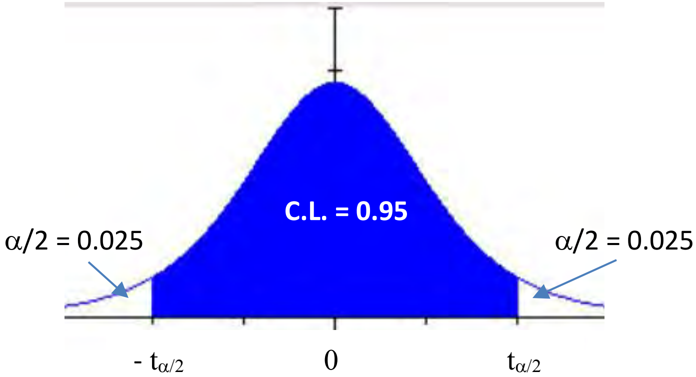
Characteristics/Properties of Student's t-distribution:

- The graph for the Student's t-distribution is similar to the standard normal curve; however, it has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- t-score, $t = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, has the same interpretation as the z-score which is the number of standard deviations $\bar{x}$ is from the mean.
- Degrees of Freedom, **df = n – 1**, come from the calculation of the sample standard deviation s. In Appendix H, we used n deviations (x – $\bar{x}$ values) to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other n – 1 deviations.
- Notation for t-distribution where T is the random variable: T~ t_{df}
- Critical value notation for t-distribution, $t_{\alpha/2}$

Critical Value, $t_{\alpha/2}$.

Recall, confidence level (CL) is the area in the middle of the distribution. The **confidence level** is often considered the probability that the calculated confidence interval estimate will contain the true population parameter.

For example, when CL = 0.95, $\alpha = 0.05$ and $\alpha/2 = 0.025$; we write $t_{\alpha/2} = t_{0.025}$.



There are two ways to find the critical value for the t-distribution. Both ways need the area in the tail ($\alpha/2$) and the degree of freedom ($n - 1$).

1.) For those who have, TI-83, 83+, 84 to find $t_{\alpha/2}$ you must use the t-distribution chart

t-distribution						
α						
	0.005	0.01	0.025	0.05	0.10	0.25
Degrees	(one tail)	(one tail)	(one tail)	(one tail)	(one tail)	(one tail)
of	0.01	0.02	0.05	0.10	0.2	0.5
freedom	(two tails)	(two tails)	(two tails)	(two tails)	(two tails)	(two tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	0.816
3	5.841	4.541	3.182	2.353	1.638	0.765
4	4.604	3.747	2.776	2.132	1.533	0.741
5	4.032	3.365	2.571	2.015	1.476	0.727
6	3.707	3.143	2.447	1.943	1.440	0.718
7	3.499	2.998	2.365	1.895	1.415	0.711
8	3.355	2.896	2.306	1.860	1.397	0.706
9	3.250	2.821	2.262	1.833	1.383	0.703
10	3.169	2.764	2.228	1.812	1.372	0.700
11	3.106	2.718	2.201	1.796	1.363	0.697
12	3.055	2.681	2.179	1.782	1.356	0.695
13	3.012	2.650	2.160	1.771	1.350	0.694
14	2.977	2.624	2.145	1.761	1.345	0.692
15	2.947	2.602	2.131	1.753	1.341	0.691
16	2.921	2.583	2.120	1.746	1.337	0.690
17	2.898	2.567	2.110	1.740	1.333	0.689
18	2.878	2.552	2.101	1.734	1.330	0.688
19	2.861	2.539	2.093	1.729	1.328	0.688
20	2.845	2.528	2.086	1.725	1.325	0.687
21	2.831	2.518	2.080	1.721	1.323	0.686
22	2.819	2.508	2.074	1.717	1.321	0.686
23	2.807	2.500	2.069	1.714	1.319	0.685
24	2.797	2.492	2.064	1.711	1.318	0.685
25	2.787	2.485	2.060	1.708	1.316	0.684
26	2.779	2.479	2.056	1.706	1.315	0.684
27	2.771	2.473	2.052	1.703	1.314	0.684
28	2.763	2.467	2.048	1.701	1.313	0.683
29	2.756	2.462	2.045	1.699	1.311	0.683
Large (z)	2.575	2.327	1.96	1.645	1.282	0.675

2.) For those who have the newer TI 84 + calculators

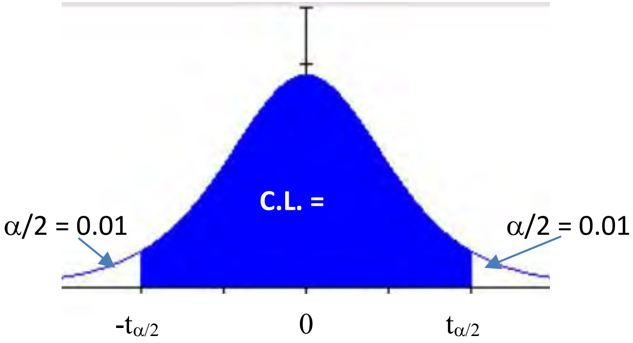
Press 2nd Vars, choose #4 invT(area in the tail, df)

Example 8.10

Find the critical value for the t-distribution for 98% confidence level where n = 12.

Solution 8.10 Solution A

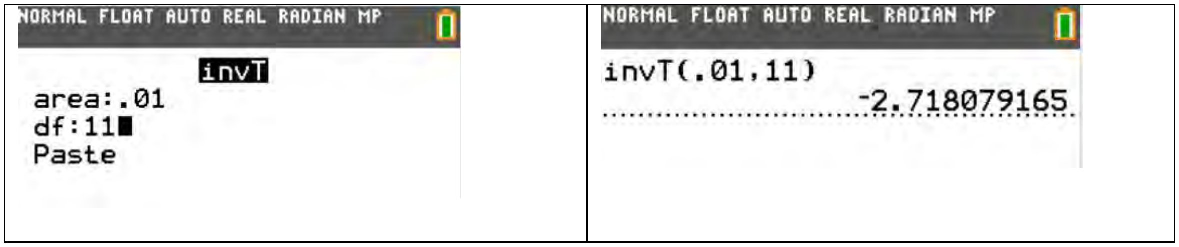
C.L = .98, α = .02, $\alpha/2$ = 0.01 we write $t_{\alpha/2} = t_{0.01}$.d.f. = n – 1 = 11



t-distribution						
	α					
Degrees of freedom	0.005 (one tail) 0.01 (two tails)	0.01 (one tail) 0.02 (two tails)	0.025 (one tail) 0.05 (two tails)	0.05 (one tail) 0.10 (two tails)	0.10 (one tail) 0.2 (two tails)	0.25 (one tail) 0.5 (two tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	0.816
3	5.841	4.541	3.182	2.353	1.638	0.765
4	4.604	3.747	2.776	2.132	1.533	0.741
5	4.032	3.365	2.571	2.015	1.476	0.727
6	3.707	3.143	2.447	1.943	1.440	0.718
7	3.499	2.998	2.365	1.895	1.415	0.711
8	3.355	2.896	2.306	1.860	1.397	0.706
9	3.250	2.821	2.262	1.833	1.383	0.703
10	3.169	2.764	2.228	1.812	1.372	0.700
11	3.106	2.718	2.201	1.796	1.363	0.697
12	3.055	2.681	2.179	1.782	1.356	0.695
13	3.012	2.650	2.160	1.771	1.350	0.694
14	2.977	2.624	2.145	1.761	1.345	0.692
15	2.947	2.602	2.131	1.753	1.341	0.691
16	2.921	2.583	2.120	1.746	1.337	0.690
17	2.898	2.567	2.110	1.740	1.333	0.689
18	2.878	2.552	2.101	1.734	1.330	0.688
19	2.861	2.539	2.093	1.729	1.328	0.688
20	2.845	2.528	2.086	1.725	1.325	0.687
21	2.831	2.518	2.080	1.721	1.323	0.686
22	2.819	2.508	2.074	1.717	1.321	0.686
23	2.807	2.500	2.069	1.714	1.319	0.685
24	2.797	2.492	2.064	1.711	1.318	0.685
25	2.787	2.485	2.060	1.708	1.316	0.684
26	2.779	2.479	2.056	1.706	1.315	0.684
27	2.771	2.473	2.052	1.703	1.314	0.684
28	2.763	2.467	2.048	1.701	1.313	0.683
29	2.756	2.462	2.045	1.699	1.311	0.683
Large (z)	2.575	2.327	1.96	1.645	1.282	0.675

Left: $-t_{\alpha/2} = -t_{0.01} = -2.718$ Right: $t_{\alpha/2} = t_{0.01} = 2.718$

Solution 8.10 Solution B (TI 84+ Calculator)


NORMAL FLOAT AUTO REAL RADIAN MP
NORMAL FLOAT AUTO REAL RADIAN MP

```
invT
area:.01
df:11
Paste
```

```
invT(.01,11)
                 -2.718079165
```

Left: - $t_{\alpha/2}$ = -$t_{0.01}$ = invT(.01, 11) = -2.718

Right: $t_{\alpha/2}$ = $t_{0.01}$ = - invT(.01, 11) = 2.718

Calculating the Confidence Interval for mean when σ is unknown

Point estimate – margin of error < μ < point estimate + margin of error

$$\left(\bar{x}-E,\bar{x}+E\right)$$

where $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$.

The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ and s from the sample data.
- Find the critical value, $t_{\alpha/2}$ that corresponds to the confidence level.
- Calculate the margin of error, E.
- 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.)

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a population mean), and state the confidence interval (both endpoints). "We estimate with % confidence that the true population mean (include the context of the problem) is between and (include appropriate units)."

Example 8.11

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

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

Solution 8.11

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution 8.11 Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, sample standard deviation, s, and the margin of error, E.

$\bar{x}$ and s can be found using 1-Var Stat on the calculator

$\bar{x} = 8.2267$, $s = 1.6722$, $n = 15 \rightarrow$ d.f. = 14

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

The confidence level is 95% (CL = 0.95)

CL = 0.95 so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\alpha/2 = 0.025$ $t_{\alpha/2} = t_{0.025}$ = the positive value of invT(.025,14) = 2.14

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.14\left(\frac{1.6722}{\sqrt{15}}\right) = .924$$

Lower bound: $\bar{x}$ - E = 8.2267 - 0.924 = 7.3

Upper bound: $\bar{x}$ + E = 8.2267 + 0.924 = 9.15

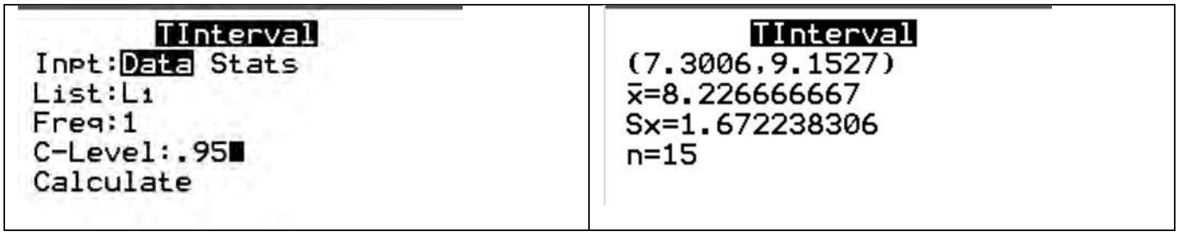
$7.3 < \mu < 9.15$

The 95% confidence interval for the population mean is (7.3, 9.15).

Interpretation: We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution 8.11 Solution B

Press STAT, TESTS. #8: Tinterval.



```
TInterval
Inpt:Data Stats
List:L1
Freq:1
C-Level:.95
Calculate
```

```
TInterval
(7.3006,9.1527)
x̄=8.226666667
Sx=1.672238306
n=15
```

Try It Σ

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Example 8.12

Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP (Human Toxome Project) tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. The table below shows how many of the targeted chemicals were found in each infant's cord blood. Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Solution 8.12 Solution A

From the sample, you can calculate $\bar{x}$ = 127.45 and s = 25.965. n = 20, and df = 20 – 1 = 19.

You are asked to calculate a 90% confidence interval:

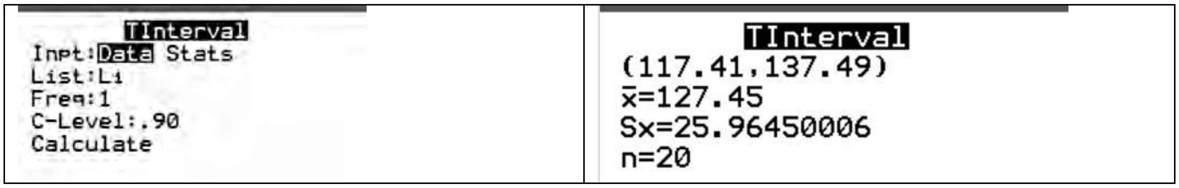
CL = 0.90, so α = 1 – CL = 1 – 0.90 = 0.10 then α/2 = 0.05, $t_{\alpha/2} = t_{0.05}$ = -invT(.05, 19) = 1.729.

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}} = 1.729\left(\frac{25.965}{\sqrt{20}}\right) = 10.038$$

$$\left.\begin{array}{l} \text{Lower bound: } \bar{x} - E = 127.45 - 10.038 = 117.412 \\ \\ \text{Upper bound: } \bar{x} + E = 127.45 + 10.038 = 137.488 \end{array}\right\} \qquad 117.412 < \mu < 137.488$$

Solution 8.12 Solution B:

Press STAT, TESTS. #8: Tinterval.



```
TInterval
Inpt:Data Stats
List:L1
Freq:1
C-Level:.90
Calculate
```

```
TInterval
(117.41,137.49)
x̄=127.45
Sx=25.96450006
n=20
```

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

8.4 | Confidence Interval for a Single Population Proportion, p

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03, 0.40 + 0.03).

How do you know you are dealing with a proportion problem? First, the underlying distribution is a binomial distribution. (There is no mention of a mean or average.) If X is a binomial random variable, then X ~ B(n, p) where n is the number of trials and p is the probability of a success. To form a proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable $\hat{P}$ (read "P hat") is that proportion,

$$\hat{P} = \frac{X}{n}$$

When n is large and p is not close to zero or one, we can use the normal distribution to approximate the binomial.

$\hat{P} \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$ Recall that for Binomial $\mu = np$ and $\sigma = \sqrt{npq}$

$\hat{P} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ is the reduced version

Calculating the Confidence Interval for proportion

Point estimate – margin of error < μ < point estimate + margin of error

$$(\hat{p} - E, \hat{p} + E)$$

where $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ and $\hat{q} = 1 - \hat{p}$

NOTE: the confidence interval can be used only if the number of success $n\hat{p}$ and number of failures $n\hat{q}$ are both greater than ten.

The steps to construct and interpret the confidence interval are:

- Calculate the sample proportion, $\hat{p} = \frac{x}{n}$, where x is the frequency, from the sample data.
- Determine if the number of success $n\hat{p}$ and number of failures $n\hat{q}$ are both greater than ten.
- Find the critical value, $z_{\alpha/2}$ that corresponds to the confidence level.
- Calculate the margin of error, E.
- 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.)

Example 8.13

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution 8.13

- 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 8.13 Solution A

To calculate the confidence interval, you must first find $\hat{p}$ (point estimate).

x = number of success

$\hat{p} = \frac{x}{n} = \frac{421}{500}$=0.842 therefore $\hat{q} = 1 - \hat{p}$ = 0.158

Second, determine if $n\hat{p}$ and $n\hat{q}$ are both greater than ten.

$n\hat{p}$ = (500)(0.842) = 421 and $n\hat{q}$ = (500)(0.158) = 79

Third find margin of error, E where the confidence level is 95% (CL = 0.95)

so α = 1 – CL = 1 – 0.95 = 0.05 and α/2 = 0.025

$z_{\alpha/2} = z_{0.025}$ = the positive value of invnorm(.025) = 1.96

$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{0.842*0.158}{500}}$=0.032

NOTE: the units of E are the same units as the point estimate; therefore in this section E is a percentage. E = 3.2%

Lower Bound: $\hat{p} - E$ = 0.842 – 0.032 = 0.81

Upper Bound: $\hat{p} + E$ = 0.842 + 0.032 = 0.874

$0.81 < p < 0.874$

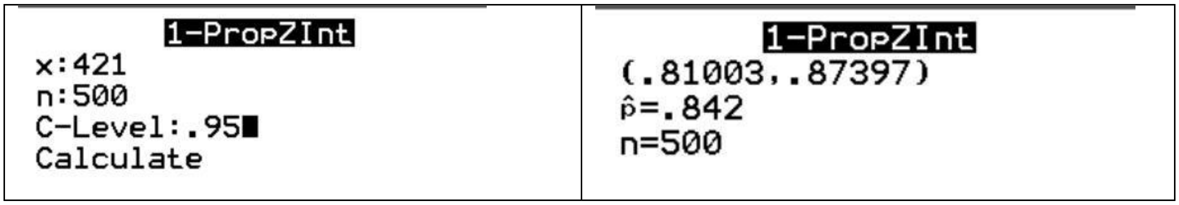
$81\% < p < 87.4\%$

Solution 8.13 Solution B

Using the TI-83, 83+, 84, 84+ Calculator

To calculate

STAT, TESTS, A: 1-PropZint



```
1-PropZInt
x:421
n:500
C-Level:.95
Calculate
```

```
1-PropZInt
(.81003,.87397)
p̂=.842
n=500
```

Interpretation:

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level:

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Try It Σ

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Example 8.14

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 59.7% are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution 8.14

- 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 8.14 Solution A

$\hat{p} = 0.597$; $\hat{q} = 1 - \hat{p} = 0.403$

The confidence level is 90% (CL = 0.90) so α = 1 – CL = 1 – 0.90 = 0.10

$\alpha/2 = 0.05$ $z_{\alpha/2} = z_{0.05}$ = the positive value of invnorm(.05) = 1.645

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.645\sqrt{\frac{0.597*0.403}{500}}=0.036$$

Lower Bound: $\hat{p} - E$ = 0.597 – 0.036 = 0.81

Upper Bound: $\hat{p} + E$ = 0.597 + 0.036 = 0.874

$0.561 < p < 0.633$

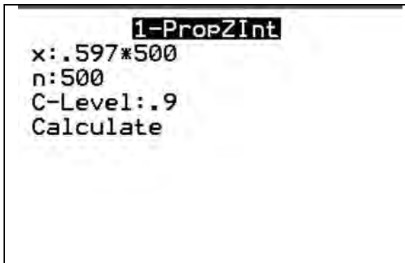
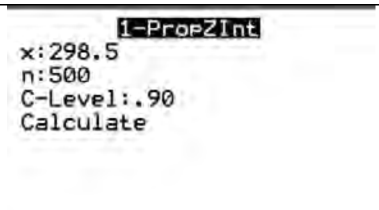
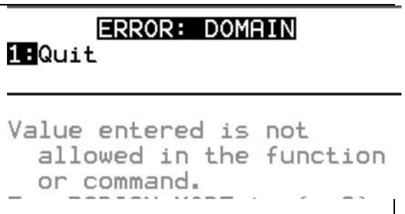
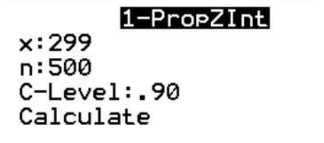
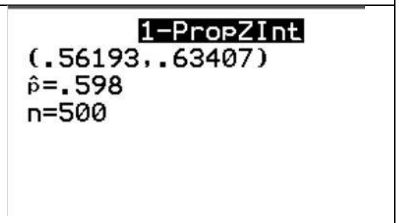
$56.1\% < p < 63.3\%$

Solution 8.14 Solution B

Using the TI-83, 83+, 84, 84+ Calculator

To calculate

STAT, TESTS, A: 1-PropZint

 1-PropZInt x:.597*500 n:500 C-Level:.9 Calculate	 1-PropZInt x:298.5 n:500 C-Level:.90 Calculate	 ERROR: DOMAIN 1:Quit Value entered is not allowed in the function or command.
NOTE: x on the calculator must be in whole number form; therefore the calculator solution will be slightly off from Solution A.	 1-PropZInt x:299 n:500 C-Level:.90 Calculate	 1-PropZInt (.56193,.63407) p̂=.598 n=500

$0.562 < p < 0.634$

$56.2\% < p < 63.4\%$

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.2% and 63.4%.

- Alternate Wording: We estimate with 90% confidence that between 56.2% and 63.4% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Try It Σ

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 79.8% are against the new legislation. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

“Plus Four” Confidence Interval for p

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is n + 4, and the new count of successes is x + 2.

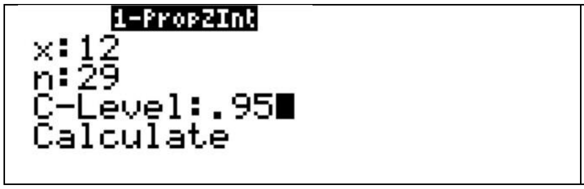
Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example 8.15

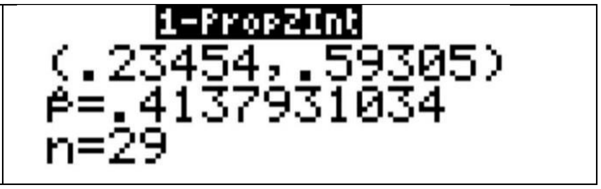
A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Ten students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution 8.15

Ten students out of 25 reported smoking within the past week, so x = 10 and n = 25. Because we are using the “plus-four” method, we will use x = 10 + 2 = 12 and n = 25 + 4 = 29.



```
1-PropZInt
x:12
n:29
C-Level:.95
Calculate
```



```
1-PropZInt
(.23454,.59305)
p̂=.4137931034
n=29
```

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 23.5% and 59.3%.

Try It Σ

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus- four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the margin of error formula to calculate the required sample size.

The margin of error formula for a population proportion is $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ where $\hat{q} = 1 - \hat{p}$

Solving for n gives you an equation for the sample size:

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Recall a clue word for E is "within". NOTE: if the estimated sample proportion is not given, then use $\hat{p}$ =.5

Example 8.16

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is **within** three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones?

Solution 8.16

$E = 0.03$ $z_{\alpha/2}$ = -invnorm(.05) = 1.645 because C.L. = 0.90

$\hat{p}$ =.5 since an estimated sample proportion is not given

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

$$n = .5(.5)\left(\frac{1.645}{.03}\right)^2$$

$$n = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Try It Σ

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be

90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

KEY TERMS

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.

Degrees of Freedom (df) the number of objects in a sample that are free to vary

Margin of error depends on the confidence level, sample size, and known or estimated population standard deviation.

Inferential Statistics also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter a numerical characteristic of a population

Point Estimate a single number computed from a sample and used to estimate a population parameter

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

Student's t-Distribution investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a "family of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

Formula Review

Calculating the Confidence Interval for mean when σ is known

$$\left(\bar{x} - E, \bar{x} + E\right) \text{ where } E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Calculator function: Zinterval

Calculating the Confidence Interval for mean when σ is unknown

$$\left(\bar{x} - E, \bar{x} + E\right) \text{ where } E = t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Calculator function: Tinterval

Calculating the Confidence Interval for proportion

$$(\hat{p} - E, \hat{p} + E) \text{ where } E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ and } \hat{q} = 1 - \hat{p}$$

NOTE: the confidence interval can be used only if the number of success $n\hat{p}$ and number of failures $n\hat{q}$ are both greater than ten.

Calculator function: 1propZint

Confidence Level	C = 90%	C = 95%	C = 99%
$Z_{\alpha/2}$	$z_{.05} = 1.645$	$z_{.025} = 1.96$	$z_{.005} = 2.576$

Calculating the Sample Size for a population mean

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Calculating the Sample Size for a population proportion

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

NOTE: if the estimated sample proportion is not given, then use $\hat{p}$ =.5

EXERCISES FOR CHAPTER 8

1. Find the critical value, $z_{\alpha/2}$, for the following different confidence levels
   a. 92%
   b. 94%
   c. 97%

2. A random sample of 49 students has a grade point average with a population standard deviation of 0.78. Find the margin of error if the confidence level is 98%.

3. A random sample of 19 students has a grade point average with a standard deviation of 0.78. Find the margin of error if the confidence level is 98%.

4. The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

   a. Identify the following:
      i. $\bar{x}$ =
      ii. σ =
      iii. n =
      iv. s =
      v. Determine which of the above values the point estimate is.
   b. Construct a 95% confidence interval for the population mean weight of newborn elephants.
      i. State the confidence interval,
      ii. Find the critical values and sketch the graph associated with the problem,
      iii. Calculate the margin of error.
   c. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

5. The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. The population standard deviation is 2.2 minutes. The population distribution is assumed to be normal.

   a. Identify the following:
      i. $\bar{x}$ =
      ii. σ =
      iii. n =
   b. Construct a 90% confidence interval for the population mean time to complete the forms.
      i. State the confidence interval,
      ii. Find the critical values.
      iii. Calculate the margin of error.

c. If the Census wants to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make?
d. If the Census did another survey, kept the margin of error the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
e. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

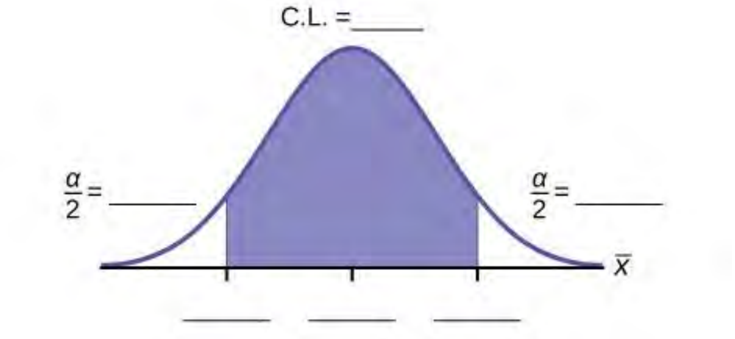
6. A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

a. Identify the following:
   i. $\bar{x} =$
   ii. $\sigma =$
   iii. $n =$
   iv. $s =$
   v. Determine which of the above values the point estimate for estimating population mean is.
b. Construct a 90% confidence interval for the population mean weight of the heads of lettuce.
c. Construct a 95% confidence interval for the population mean weight of the heads of lettuce.
d. In complete sentences, explain why the confidence interval in part b is larger than in part c.
e. What would happen if 40 heads of lettuce were sampled instead of 20, and the margin of error remained the same?
f. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

7. The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

a. Identify the following:
   i. $\bar{x} =$
   ii. _____ = 15
   iii. $n =$
b. Construct a 95% confidence interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.
c. What is $\bar{x}$ estimating?
d. How much area is in both tails (combined)? α
e. Identify the margin of error.
f. In one complete sentence, explain what the interval means.
g. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the margin of error become larger or smaller? How do you know?
h. Using the same mean, standard deviation, and sample size, how would the margin of error change if the confidence level were reduced to 90%? Why?

i. Fill in the blanks on the graph with the area, upper and lower bounds of the confidence interval, and the sample mean



8. Find the critical value, $t_{\alpha/2}$, for the following different confidence levels

d. 90% when n = 25
e. 95% when n = 10
f. 98% when n = 36

9. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

a. Construct a 95% confidence interval for the population mean time spent waiting.
   i. State the confidence interval,
   ii. Find the critical values and sketch the graph,
   iii. Calculate the margin of error.
b. Explain in complete sentences what the confidence interval means.

10. One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

a. Define the random variable X in words.
b. Define the random variable $\overline{X}$ in words.
c. Which distribution should you use for this problem? X ~ ______________
d. Construct a 99% confidence interval for the population mean hours spent watching television per month.
   i. State the confidence interval,
   ii. Find the critical value
   iii. Calculate the margin of error.
e. Why would the margin of error change if the confidence level were lowered to 95%?

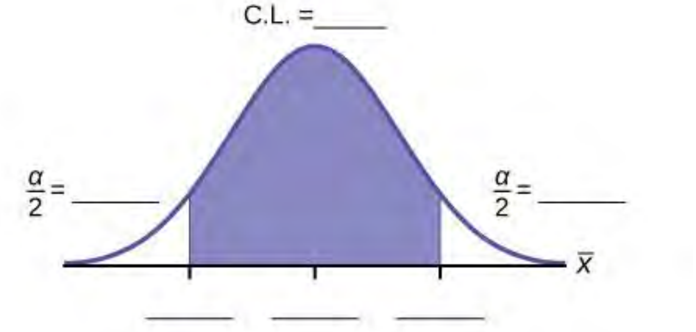
11. The data in the following table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence

interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq
1	1
2	7
3	18
4	7
5	6

a. Calculate the following:
   i. $\bar{x}$ =
   ii. s_x =
   iii. n =
b. Define the random variable $\bar{X}$ in words.
c. Construct a 95% confidence interval for the true mean number of colors on national flags.
d. How much area is in both tails (combined)?

e. Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.



f. In one complete sentence, explain what the interval means.
g. Using the same $\bar{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the margin of error become larger or smaller? How do you know?
h. Using the same $\bar{x}$, s_x, and n = 39, how would the margin of error change if the confidence level were reduced to 90%? Why?

12. Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?
b. If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

13. Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

a. Identify the following:
    i. x =
    ii. n =
    iii. $\hat{p}$ =
b. Define the random variables X and $\hat{p}$ in words.
c. Which distribution should you use for this problem? X ~ _______________
d. Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions.
    i. State the confidence interval,
    ii. Sketch the graph
    iii. Calculate the margin of error.
e. List two difficulties the company might have in obtaining random results, if this survey were done by email.

14. Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid- level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

a. Construct a 95% confidence interval for the percent of executives who prefer trucks.
b. Calculate the margin of error.
c. Suppose we want to lower the sampling error. What is one way to accomplish that?
d. The sampling error given in the survey is ±2%. Explain what the ±2% means.

15. A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

a. Find a 90% confidence interval, and state the confidence interval and the margin of error.
b. What would happen to the confidence interval if the level of confidence were 95%?

16. The Ice Chalet offers dozens of different beginning ice- skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

a. Calculate the following:
    i. x =
    ii. n =
    iii. $\hat{p}$ =

b. State the estimated distribution of X. X~ ___________

c. What is $\hat{p}$ estimating?

d. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

e. How much area is in each tail?

f. Calculate margin of error. .

g. In one complete sentence, explain what the interval means.

h. Using the same $\hat{p}$ and level of confidence, suppose that n were increased to 100. Would the margin of error become larger or smaller? How do you know?

i. Using the same $\hat{p}$ and n = 80, how would the margin of error change if the confidence level were increased to 98%? Why?

j. If you decreased the allowable margin of error, why would the minimum sample size increase (keeping the same level of confidence)?

17. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a. Construct a 95% confidence interval for the population mean height of male Swedes.

b. Find the critical values associated with this problem.

c. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

18. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal. Construct a 95% confidence interval for the population mean length of engineering conferences.

19. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. Which distribution should you use for this problem? Explain your choice.

b. Construct a 90% confidence interval for the population mean time to complete the tax forms.

c. Find the critical values.

d. Calculate the margin of error.

e. If the firm wished to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make?

f. If the firm did another survey, kept the margin of error the same, and only surveyed 49 people, what would happen to the level of confidence? Why?

g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

20. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight

was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. Which distribution should you use for this problem? Explain your choice.
b. Construct a 90% confidence interval for the population mean weight of the candies.
c. Calculate the margin of error.
d. Construct a 98% confidence interval for the population mean weight of the candies.
e. In complete sentences, explain why the confidence interval in part b is smaller than the confidence interval in part d.
f. In complete sentences, give an interpretation of what the interval in part d means.

21. A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

a. Construct a 90% confidence interval for the population mean number of letters campers send home.
b. What will happen to the margin of error and confidence interval if 500 campers are surveyed? Why?

22. MULTIPLE CHOICE: What is meant by the term "90% confident" when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

23. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. The table below shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$2,309,200	$7,400	$391,000	$13,300	$468,700	$353,900	$3,714,500	$5,800
$1,243,900	$2,900	$467,400	$56,800	$733,200	$986,100	$1,109,300	$3,800
$10,900	$400	$632,500	$15,800	$953,800	$88,600	$1,113,500	$6,600
$385,200	$5,800	$512,900	$75,200	$745,100	$378,200	$3,072,100	$9,500
$581,500	$3,600	$405,200	$41,000	$202,400	$13,200	$1,626,700	$8,000

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the margin of error.
c. Create a 95% confidence interval for the mean total individual contributions.

d. Interpret the confidence interval in the context of the problem.

24. The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the margin of error for mean U.S. household income.

25. The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

26. In six packages of "The Flintstones® Real Fruit Snacks" there were five Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

a. Calculate the point estimate.
b. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
c. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

27. A random survey of enrollment at 35 community colleges across the United States yielded the following figures:

6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
b. Calculate the margin of error.
c. State the point estimate.
d. State the critical value.
e. What will happen to the margin of error and confidence interval if 500 community colleges were surveyed? Why?

28. Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. Construct a 95% confidence interval for the population mean time wasted.
b. Find the critical values.
c. Explain in a complete sentence what the confidence interval means.

29. A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample

of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. Which distribution should you use for this problem? Explain your choice.
b. Construct a 95% confidence interval for the population mean length of time.
c. Find the critical values.
d. What does it mean to be "95% confident" in this problem?

30. Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. Construct a 99% confidence interval for the population mean length of time using training wheels.
b. Calculate the margin of error.
c. Why would the margin of error change if the confidence level were lowered to 90%?

31. The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00	$2,555,363.20
$29,050.00	$2,000.00	$18,000.00	$181,557.20	$63,520.00	$1,287,933.80
$19,500.00	$0	$61,810.20	$149,970.80	$708,258.90	$219,148.30
$12,025.00	$6,500.00	$76,530.80	$409,000.00	$135,810.00	$502,578.00
$60,521.70	$0	$31,500.00	$119,459.20	$2000.00	$705,061.10

a. Find $\bar{x} = \$$___________
b. Find sample standard deviation.
c. Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

32. Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The Table 8.13 shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56	59
74	63	53	50	59	60

60	57	46	55	63	57
47	55	57	43	61	62
49	67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms.

33. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a. Identify the following
   i. $\bar{x}$ =
   ii. s =
   iii. n =
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

34. In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Construct a 95% confidence interval for the population mean cost of a used car.
b. Calculate the margin of error.
c. Explain what a "95% confidence interval" means for this study.

35. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
b. Find the critical value.
c. Calculate the margin of error.
d. If you wanted a smaller margin of error while keeping the same level of confidence, what should have been changed in the study before it was done?

36. A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a. Find
   i. $\bar{x}$ =
   ii. s =
   iii. n =

b. Construct a 95% confidence interval for the population mean worth of coupons.
c. Find the critical values
d. Calculate the margin of error.
e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

37. A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

a. Find the 95% Confidence Interval for the true population mean for the amount of soda served.
b. What is the margin of error?

38. Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

39. Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. Identify
   i. x =
   ii. n =
   iii. $\hat{p}$ =
b. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
c. Calculate the margin of error.
d. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

40. According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Which distribution should you use for this problem? X ~ __________________
b. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
c. Calculate the margin of error.

41. An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites.

In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families.

a. Which distribution should you use for this problem? X ~ __________________
b. Construct a 95% confidence interval.
c. Calculate the margin of error.

42. Refer to the information in Exercise 41

a. Construct three 95% confidence intervals.
    i. percentage of all Asians who would welcome a white person into their families.
    ii. percentage of all Asians who would welcome a Latino into their families.
    iii. percentage of all Asians who would welcome a black person into their families.
b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

43. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and $\hat{p}$ in words.
b. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
c. Explain what a "97% confidence interval" means for this study.

44. A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was "What is the main problem facing the country?" Twenty percent answered "crime." We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and $\hat{p}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
d. Calculate the margin of error.
e. Suppose we want to lower the sampling error. What is one way to accomplish that?
f. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

45. Refer to Exercise 44 Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the

population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
b. Calculate the margin of error.
c. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

46. According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that "education and our schools" is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

a. Determine the point estimate for the true population proportion.
b. Find the 92% confidence interval for the population proportion. .
c. Find the critical value.
d. Find the margin of error.

47. 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.

a. Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.
b. Find the point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness.

48. On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify CL and α.
d. Calculate the margin of error based on the information provided.
e. Compare the margin of error in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

49. A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the margin of error for this confidence interval.

b. Can we conclude with 95% confidence that more than half of all American adults believe this?

c. Use the point estimate from part a and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.

d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

50. Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.

b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The margin of error of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.

c. Describe how the confidence interval would change if the CL changed from 99% to 90%.

51. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

52. In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as "likely" or "very likely." Use the "plus four" method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

53. In a March 2018 poll by CNN, 57% say that things are going well in the United States today. A random sample of 1015 adults by telephone/cellphone was conducted.

a.) Find the 90% confidence interval for the true proportion of adults that says things are going well in the United States today.

b.) According the article, the full sample has a margin of error of plus/minus 3.6 percentage points. If we want a margin of error of 3 percentage points with 95% confidence, how large of a sample should be used?

54. According to a survey conducted by Bankrate, which surveyed 2,194 adults, including 1,330 homeowners. It was seen that 29 percent of respondents with a mortgage either didn't know their rate or wouldn't say.

a. State the point estimate as a fraction.

b. Find the 98% confidence of the true proportion of adults with a mortgage either don't know their rate or wouldn't say.

c. interpret the confidence in sentence form

REFERENCES

8.1 Confidence interval about a single population mean when σ known

"American Fact Finder." U.S. Census Bureau. Available online at *http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).*

"Disclosure Data Catalog: Candidate Summary Report 2012." U.S. Federal Election Commission. Available online at *http://www.fec.gov/data/index.jsp (accessed July 2, 2013).*

"Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall." Foothill De Anza Community College District. Available online at *http://research.fhda.edu/factbook/FH_Demo_Trends/ FoothillDemographicTrends.htm (accessed September 30, 2013).*

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at *http://www.cdc.gov/growthcharts/*

La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at *http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).*

"Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at *http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).*

"Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at *http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml (accessed July 2, 2013).*

"National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at *http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).*

8.2 Confidence interval about a single population mean when σ unknown

"America's Best Small Companies." Forbes, 2013. Available online at *http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).*

Data from Microsoft Bookshelf.

Data from http://www.businessweek.com/. Data from *http://www.forbes.com/.*

"Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at *http://www.fec.gov/data/index.jsp (accessed July 2,2013).*

"Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at *http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).*

"Metadata Description of Leadership PAC List." Federal Election Commission. Available online at *http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml (accessed July 2, 2013).*

8.3 Confidence interval about a single population proportion

Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at *http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).*

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy." PewInternet, 2013. Available online at *http://www.pewinternet.org/Reports/2013/Teens-Social- Media-And-Privacy.aspx (accessed July 2, 2013).*

Prince Survey Research Associates International. "2013 Teen and Privacy Management Survey." Pew Research Center: Internet and American Life Project. Available online at *http://www.pewinternet.org/~/media//Files/Questionnaire/2013/ Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).*

Saad, Lydia. "Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at *http://www.gallup.com/poll/162758/three-four- workers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).*

The Field Poll. Available online at *http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).*

Zogby. "New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security." Zogby Analytics, 2013. Available online at *http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor- prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).*

"52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at *http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).*

Agiesta, Jennifer. "CNN Poll: Trump Approval Steady amid Rising Outlook for the Country." *CNN*, Cable News Network, 7 May 2018, www.cnn.com/2018/05/07/politics/cnn-poll-trump-steady-right-direction-rises/index.html. (accessed July 17, 2018)

Martin, Emmie. "30% Of Homeowners Are Making a Mistake That Could Cost Them Thousands." *CNBC*, CNBC, 1 May 2018, www.cnbc.com/2018/05/01/a-third-of-homeowners-dont-know-their-mortgage-rate.html?__source=twitter%7Cmain. (accessed July 17, 2018)

9 | HYPOTHESIS TESTING WITH ONE SAMPLE

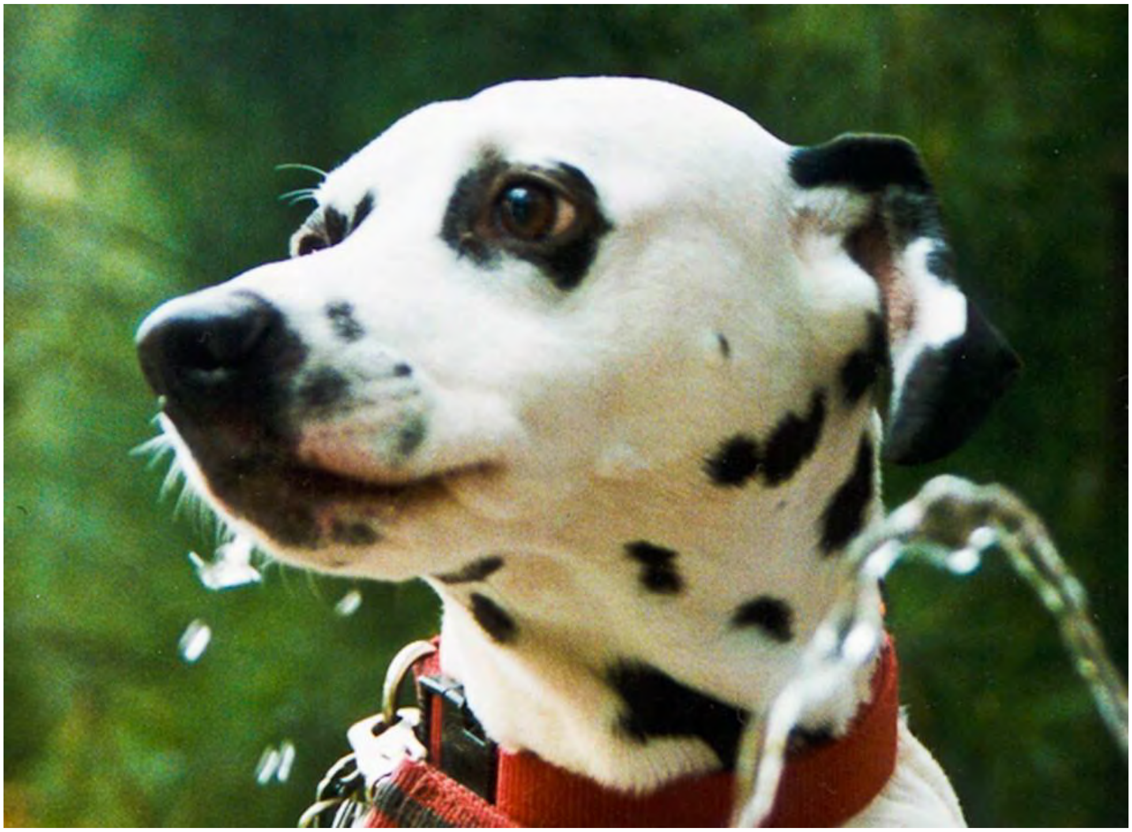


Figure 9.1 You can use a hypothesis test to decide if a dog breeder's claim that the average Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

Chapter Objectives

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

- 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.
- Differentiate between Type I and Type II Errors

Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. In the last chapter, we used confidence intervals to estimate an unknown population parameter. In this chapter we will use statistical inference to make a decision about a parameter; that is, given two competing claims about an unknown parameter, we will decide which of them is more plausible. The method for doing this is called *hypothesis testing*. We already have all of the tools needed to implement this method – and we will shortly distill the process into a straightforward five-step method. We will illustrate the basic ideas with an example:

The ACME chemical corporation makes a fabric cleaner; the company claims that this product successfully removes 90% of all stains. A consumer organization questions this claim, and decides to test the product. They select a random sample of 100 stained garments, and apply the product to each. They find that the stain remover successfully cleaned only 78 of them. Does this sample data provide evidence that the company is overstating the effectiveness of their product? That is, does the sample data provide evidence that the success rate for the fabric cleaner is *less than* 90%?

We let p denote the success rate for the cleaner – that is, this is the proportion of *all* stains in the population that will be removed by the product. Now we see that we have two competing claims:

- ACME claims that $p = 0.90$
- The consumer organization claims that $p < 0.90$.

Which one of these claims is true? Of course, there is no way to determine p explicitly, but we could use the methods of Chapter 8, along with the given sample data, to compute a confidence interval for p. Doing this gives us a 95% confidence interval of **(0.699, 0.861)**. That is, we are 95% confident that $69.9\% < p < 86.1\%$. Since the upper bound of the interval is well under 90%, the sample data does indeed provide evidence that p is less than 90%.

But there is another way we could go about this. Note that if the company's claim is true, then we would expect to get about 90 successes in a sample of 100. But the number of successes in 100 trials will change from sample to sample, so the fact that we got fewer than 90 does not in itself provide evidence that the success rate is less than 90%. ACME could claim that the discrepancy between their claimed success rate and the sample data is simply due to sample variation. This is where probability comes in. Let's give the company the benefit of the doubt: let's assume that the success rate really is 90%, and calculate the probability of getting a sample of 100 garments with a success rate of 78% or less (a sample proportion less than 78% would also support the consumer organization's claim). In Chapter 7, we saw that if $n = 100$ and $p = .90$, then the sampling proportion for $\hat{p}$ is approximately normal, with mean $\mu_{\hat{p}} = p = .90$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.90(.10)}{100}} = 0.03$. Using the calculator, we see that the probability we want is:

$P(\hat{p} \leq 0.78)$ = normalcdf(-10^99, .78, .90, .03) = 0.000031.

So if ACME's claim really is true, then getting 78 or fewer successes in a sample of 100 garments would be highly unlikely (only about 3 times for every 100,000 samples).

However, the consumer group really *did* get a sample of 100 with only 78 successes! This means that there are two possibilities: The success rate really is 90%, and a very rare event occurred, or else the assumption that p = .90 is not correct. Given how small the probability we calculated was, the second option seems to be more plausible. In other words, the sample data is not compatible with ACME's claim about the success rate of their product; so we would reject the assumption that the success rate is 90%. Thus, this sample data provides evidence that ACME was overstating the success rate of their product.

The line of reasoning we used here is known as **hypothesis testing**, and it is a widely used method of statistical inference. In general, a hypothesis test will always consist of two contradictory hypotheses or statements, a decision based on sample data, and a conclusion. Note that in all of our problems, we will be given either the sample data or the summary statistics. Reviewing our example, we see that a hypothesis test will involve the following five steps:

1. Set up two contradictory hypotheses, which we call the *null hypothesis* and the *alternative hypothesis.*
2. Determine the correct sampling distribution to perform the calculations.
3. Assuming that the null hypothesis is true, we calculate the probability of getting sample data like that we have actually observed.
4. If this probability is sufficiently small, we reject the null hypothesis.
5. Interpret the decision to write a meaningful conclusion; i.e. interpret the decision to answer the original question.

These five steps comprise what is called the **Five Step Method** of hypothesis testing.

In this chapter, we develop hypothesis tests on single means and single proportions. We will also learn about the errors associated with these tests. In subsequent chapters, we learn many other tests to apply these ideas to other situations – but *every* test we learn will follow the Five Step Method.

9.1 | Elements of a Hypothesis Test

Null and Alternative Hypotheses

Every hypothesis test begins by considering two hypotheses. These are called *the null hypothesis* and *the alternative hypothesis*. These hypotheses contain opposing viewpoints, and almost always are stated in terms of one or more unknown population parameters.

The null hypothesis is denoted by H_0. It is a statement about the parameter that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt. Thus all of our calculations will be done using the assumption that the null hypothesis is true.

The alternative hypothesis is denoted by H_a. It is a claim about the parameter that is contradictory to H_0. The alternative H_a is also sometimes called the "research hypothesis".

Since the hypotheses are contradictory, only one of them can be true; thus, if our sample data leads us to reject H_0, then we will have statistical evidence that H_a is true. And we only will reject H_0 when the sample data provides compelling evidence that it is false (this will be discussed in greater detail in the next section).

Based on calculations using sample data, we will make a decision about H_0. There are only two possible options for a decision: They are:

- Reject H_0
- Do not reject H_0

If the decision is to reject H_0, then this means that the sample data is incompatible with the assumption that H_0 is true. When the decision is to reject H_0 then the sample data provides evidence that H_a is true. On the other hand, if the decision is "do reject H_0", then the sample data simply does not provide enough evidence to reject the null hypothesis.

By its very nature, H_0 asserts that the value of the parameter is known; thus, H_0 will *always* include an equal sign; so the null hypothesis will always be a statement involving either "=", "≤" or "≥". On the other hand, H_a will never have a symbol with an equal in it; so H_a will be a statement involving either "≠", ">" or "<". The choice of symbol depends on the claim being tested. (Note that some practitioners write H_0 as a simple statement of equality, even with > or < as the symbol in the alternative hypothesis. This practice is also acceptable.)

Basic Rules for Setting up the Hypotheses

1. The null hypothesis must include an equal sign.
2. One of the hypotheses must represent the claim being tested.
3. The hypotheses should be contradictory statements.

Example 9.1

We wish to test the claim that more than 30% of registered voters in a certain county voted in the primary election. Find the hypotheses for this test.

Solution 9.1

The claim here concerns a population proportion p, the proportion of all registered voters in the county that voted in the primary. The claim is that *more than* 30% voted, or $p > 0.30$. This claim must be represented by either H_0 or H_a; since the claim does not include an equal sign, it will be H_a. Remember that H_0 and H_a are opposite of each other. Thus we get the hypotheses:

$$H_0:\ p \le 0.30\ ,\quad H_a:\ p > 0.30$$

Try It Σ

9.1 A medical trial is conducted to test whether or not a new medicine reduces cholesterol by at least 25%. State the null and alternative hypotheses for the test.

Example 9.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). State the hypotheses for this test.

Solution 9.2

Here the claim being tested involves a population mean, μ; and the claim states that the mean GPA is *different from* 2.0. So the claim is that $\mu \neq 2.0$. Since this is a "not equal" statement, it must be represented by H_a, and H_0 will be the opposing statement. So the hypotheses are:

$$H_0:\ \mu = 2.0 \quad \text{vs.} \quad H_a:\ \mu \neq 2.0$$

Try It Σ

9.2 We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses:

a. H_0: μ _____ 66 b. H_a: μ _____ 66

Example 9.3

Suppose that we want to test whether the average time needed to complete a bachelor's degree in the U.S. is less than five years.

a) State the hypotheses for the test.

b) Suppose that sample data has been collected and analyzed, and the decision is to reject H_0. Interpret this decision to write a meaningful conclusion for this test.

Solution 9.3

a) The claim is that the population mean μ is *less* than five years; i.e. the claim is $\mu < 5$. This must be represented by the alternative hypothesis, and so the null and alternative hypotheses are:

$$H_0\text{: } \mu \geq 5 \quad \text{vs.} \quad H_a\text{: } \mu < 5.$$

b) Since the decision is to reject H_0, there is evidence that H_0 is false. That is, there is evidence that H_a is true. Thus, there is significant evidence that the average time needed to complete a bachelor's degree in the U.S. is less than five years.

Try It Σ

9.3 We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ ____ 45   b. H_a: μ _____ 45

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 of those, a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Suppose that a test is conducted to test whether the percentage of U.S. students who take advanced placement exams is more than 6.6%.

a. State the null and alternative hypotheses for the test.

b. Suppose that sample data has been collected and analyzed, and the decision is to *not* reject H_0. Interpret this decision to write a meaningful conclusion for this test.

Solution 9.4

a. The claim is that more than 6.6% pass; i.e. that $p > 0.066$. This is a strict inequality, so it will be represented by H_a, and we get the hypotheses: H_0: $p \leq 0.066$ vs. H_a: $p > 0.066$

b. Since the decision is to not reject H_0, this means that there is not enough evidence to support H_a. Thus, there is not enough evidence to conclude that the percentage of U.S. students who take advanced placement exams is more than 6.6%.

Try It Σ

9.4 On a state driver's test, about 40% pass the test on the first try. We want to test the claim that more than 40% pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses:

a. H_0: p _____ 0.40   b. Ha: p _____ 0.40

Type I and Type II Errors

When we perform a hypothesis test, the decision to reject H_0 depends on randomly selected sample data – so there is always the possibility of an error. No matter how careful we are, there is always a small probability that we will just get an unusual sample which leads us to the wrong decision. Since there are two possible decisions, and each of these can be either correct or incorrect, there are a total of four possible outcomes for a test, which are summarized in the following table:

	H_0 is true	H_0 is false
Reject H_0	**Type I Error**	Correct Decision
Fail to reject H_0	Correct Decision	**Type II Error**

As we see, there are two outcomes where the correct decision is made, and two outcomes where an error occurs.

A **Type I error** occurs when the decision is to reject H_0 when in fact H_0 is true. The probability of a Type I error is denoted by the Greek letter α.

A **Type II error** occurs when the decision is to *not* reject H_0 when in fact H_0 is false. The probability of a Type II error is denoted by the Greek letter β.

Obviously, both α and β should be as small as possible because they are probabilities of errors.

There is one more probability related to this chart; the *Power of the Test* is the probability of rejecting the null hypothesis when H_0 is false. Thus, the Power of the test is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example 9.5

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
TypeII error: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe.
β = probability that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error; if Frank thinks his rock climbing equipment is safe, he will go ahead and use it.

Try It Σ

9.5 Suppose the null hypothesis, H_0, is: The blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Example 9.6

Suppose the null hypothesis, H_0, is: The victim of an automobile accident arriving at the emergency room of a hospital is infected with HIV.

a. Which type of error has a greater consequence, a Type I or Type II error?
b. Describe the probability α in words.

Solution 9.6

Type I error: The emergency crew thinks that the victim does not have HIV when, in reality the victim is infected.
Type II error: The emergency crew thinks that the victim does have HIV, when in reality the patient is *not* infected.

a. The error with the greater consequence is the Type I error. If the emergency crew mistakenly believes that the victim is uninfected with HIV, then they may not take proper precautions to avoid infection while treating him.

b. Recall that α is the probability of a Type I error; that is, the probability that the emergency room staff *thinks* the patient does not have HIV when in fact he/she does have the virus.

Try It Σ

9.6 Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Example 9.7

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious, a Type I or Type II error?

Solution 9.7

The claim is that the cure rate is at least 75%. So the claim is $p \geq 0.75$; this statement includes an equal sign, and so it will be the null hypothesis H_0. Thus, the errors are:

Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Try It Σ

9.7 Determine the Type I and Type II errors for the following scenario:
Assume a null hypothesis, H_0, states that the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

The *p*-value and Significance Level for a Test

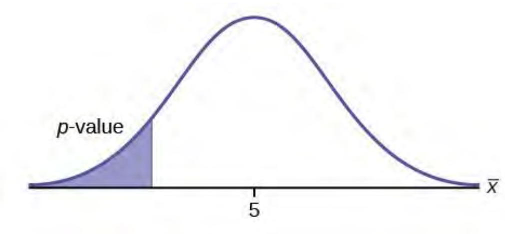
Recall the Five Step Method for Hypothesis Testing:

1. Set up the null hypothesis and the alternative hypothesis.
2. Select the correct test; that is, determine the correct sampling distribution to perform the calculations.
3. Assume that the null hypothesis is true, and calculate the probability of getting a sample statistic that differs at least as much from the hypothesized mean as the statistic computed from our sample data.
4. If this probability is sufficiently small, we reject the null hypothesis.
5. Interpret the decision to write a meaningful conclusion; i.e. interpret the decision to answer the original question.

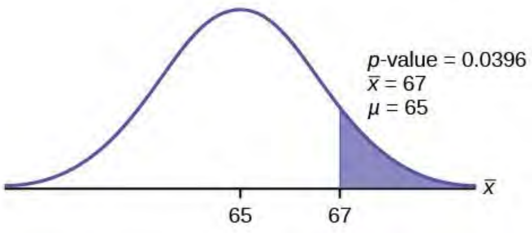
We have discussed at some length Steps 1 and 5 of the process; now we focus on Steps 3 and 4. The probability calculated in Step 3 is called the ***p-value*** for the test. To compute this value, we assume that the null hypothesis is true, and then find the probability of getting a sample statistic that is at least as extreme as the statistic from our sample. Thus, the smaller the *p*-value is, the stronger the evidence against H_0. For any test, the *p*-value will be calculated in an appropriate sampling distribution. For tests about a single population mean or population proportion, we can use the techniques learned in Chapter 7. For example, suppose that we want to test the claim that the average time spent to complete a bachelor's degree is less than five years. As we saw in Example 9.3, the hypotheses for this test are:

$$H_0\text{: } \mu \geq 5 \quad \text{vs.} \quad H_a\text{: } \mu < 5$$

Further suppose that we collected a sample and found that $\bar{x} = 3.5$. Then the *p*-value would be the probability *p-value* = P($\bar{x} \leq 5$). Graphically, this probability is the area under the normal curve that is to the *left* of $\bar{x} = 3.5$. For this reason, a test in which the alternative hypothesis is a "less than" statement is called a *left-tailed test.*



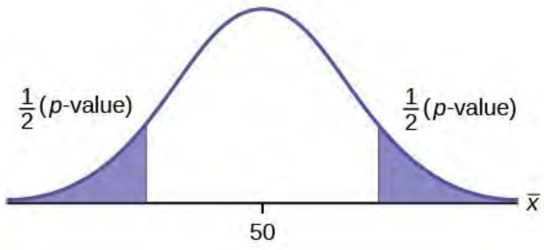
Similarly, if we were testing the hypotheses H_0: $\mu \leq 65$ vs. H_a: $\mu > 65$, and the sample data gave us $\bar{x} = 67$, then the *p*-value would be the probability p = P($\bar{x} \geq 67$). Graphically, this probability is the area under the normal curve that is to the *right* of $\bar{x} = 67$. So a test in which the alternative hypothesis is a "greater than" statement is called a *right-tailed test.*



Finally, if the alternative hypothesis is a "not equal" statement, then the p-value will be the area in *both* tails. And if the sampling distribution is symmetric, then the p-value will be split evenly between the two tails. E.g. suppose we test the hypotheses:

H_0: $\mu = 50$ H_a: $\mu \neq 50$

Then the graphical representation of the *p*-value would be as follows:



Tests in which the alternative hypothesis is a "not equal" statement are called *two-tailed tests*.

Once we have calculated the *p*-value, we still need to use it to decide whether or not to reject the null hypothesis. According to Step 4 of the Five Step Method, we will reject the null hypothesis H_0 only if p is "sufficiently small"; and the criterion for this is provided by a numerical value called the ***significance level*** of the test. Before conducting a hypothesis test, the researcher will decide on a maximum allowable probability for a Type I error, which we denote as α. This value is a threshold value: **if the *p*-value is less than α, then we will reject H_0.** The significance level chosen will depend on the consequences of making a Type I error. For example, in the social sciences, it is not uncommon to use a significance level of α = .10, whereas in the pharmaceutical industry the standard level of significance for drug testing is α = .01. The most widely used level is α = .05. The significance level will always be given to us.

Rejection Rule Using the *p*-value

- We reject H_0 at significance level α if $p < \alpha$.

- If $p \geq \alpha$, then we do not reject H_0.

NOTE: we do not use the phrase "accept H_0" if the conclusion is "do not reject H_0".

Example 9.9

It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. We let p be the proportion of male babies born to It's A Boy clients. Suppose that we test the hypotheses

$$H_0: p = 0.50, \qquad H_a: p > 0.50$$

using a significance level of $\alpha = 0.01$.

a. Is this test left-tailed, right-tailed or two-tailed? Explain.
b. Suppose further that the sample data yields a p-value = 0.025. What would be the decision for H_0?
c. Interpret this decision to state a conclusion in simple, non-technical terms.

Solution 9.9

a. Because the alternative hypothesis is a ">" statement, this is a **right-tailed** test.

b. Since p – value = .025 > .01, we **do not reject H_0**.

c. The alternative hypothesis represents the claim that the company improves a couple's chances of having a boy. Since we do not reject H_0, there is not enough evidence to conclude that H_a is true. Thus, **there is not enough evidence to support the company's claim that their procedures improve the chances of a boy being born.**

Non-technical terms (interpretation) of the conclusion

If claim is H_0 and Reject H_0, then "there is enough evidence at ___ % level of significance to reject the claim that …

If claim is H_a and Reject H_0, then "there is enough evidence at ____% level of significance to support the claim that …

Try It Σ

9.9 A certain computer component is known to have an average time of 520 hours between failures. A modification is made to the component that is supposed to extend the time between failures. To test that the modification is successful, a researcher samples 100 of the components and uses the data to test the hypotheses H_0: $\mu \leq 520$, H_a: $\mu > 520$ using a 0.05 significance level.

a. Is this test left-tailed, right-tailed or two-tailed? Explain.
b. Suppose further that the sample data yields a p-value of p =0.0141. What would be the decision for H_0?
c. Interpret this decision to state a conclusion in simple, non-technical terms.

9.2 | The Z-Test for a Single Population Mean

Suppose that we wish to test a claim about an unknown population mean, μ. The appropriate point estimate for μ is the sample mean, $\bar{x}$. From the Central Limit Theorem, we know that the sampling distribution for $\bar{x}$ will be approximately normal provided either the population distribution is normal, or if the sample size is $n \geq 30$. Moreover, if we assume that the population standard deviation σ is known, then the mean and standard deviation of the sampling distribution are:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

When these conditions are met, we can use the normal distribution to find the *p*-value for the test. And such a test is referred to as a **Z**-Test. We will illustrate this method with an example:

Example 9.10

A baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. The baker knows from baking hundreds of loaves of bread that the standard deviation for the height is 1.5 cm. and the distribution of heights is normal. He bakes 30 loaves of bread and finds that the mean height for the sample is 17 cm. Using a .05 significance level, does this sample data provide evidence to that the mean bread height is more than 15 cm?

Solution 9.10

The claim being tested is that the mean is more than 15 cm, or $\mu > 15$. Since this is a strict inequality, this claim will be represented by H_a. The null hypothesis must contradict the alternate hypothesis, so we get the hypotheses:

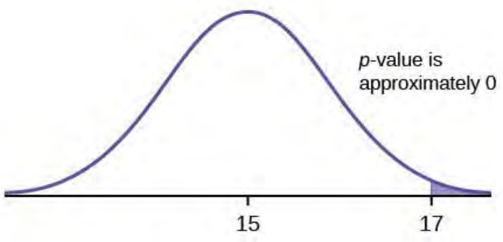
$$H_0: \mu \leq 15 \quad \text{and} \quad H_a: \mu > 15$$

We are sampling from a normal distribution and the population SD is known (σ = 1.5 cm.), so the sampling distribution for $\bar{x}$ is normal with mean $\mu_{\bar{x}} = 15$ and $\sigma_{\bar{x}} = \frac{1.5}{\sqrt{30}} = 0.27386$.

Next we calculate the *p*-value for the test. Suppose the null hypothesis is true: that is, assume that the mean height of the loaves is no more than 15 cm. Using this assumption, we find the probability of observing a sample mean that is greater than or equal to 17; since the sampling distribution is normal, we can use any of the methods from Chapter 6 to do the calculation. For example, we can use the normalcdf function in the TI-84 calculator:

$$p = P(\bar{x} \geq 17) = \text{normalcdf}(17, 10\text{^}99, 15, \frac{1.5}{\sqrt{30}}) = 1.42 \times 10^{-13}$$

In other words, if the population mean really is μ = 15, then the probability of selecting a sample of 10 with $\bar{x} \geq 17$ is $p = 0.00000000000014$, which is virtually zero.



Since p < .05 we reject H_0. In fact, with such a small p-value, we would have rejected the null hypothesis at *any* level of significance. A p-value of approximately zero tells us that, if the population mean height really been 15 cm, it would be virtually impossible to get a sample mean of 17 cm. purely by *chance*. And because the outcome of 17 cm. is so unlikely, its occurrence provides strong evidence against the null hypothesis. There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

NOTE: We could also gauge the evidence against the null hypothesis by computing the z-score corresponding to $\bar{x} = 17$:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{17 - 15}{1.5 / \sqrt{30}} \approx 7.30$$

This z-value is called the **test statistic** or **test value** for the hypothesis test. In this case, we see that a sample mean of $\bar{x} = 17$ is more than 7 standard deviations (in the sampling distribution) from the expected value – this again provides clear evidence that the null hypothesis is incorrect. At the end of this section we will explore an optional way to make the decision about H_0 using the only the test statistic.

Try It Σ

9.10 A machine at a bottling plant is programmed to fill bottles to an average volume of 12.2 oz. The volumes are normally distributed with a standard deviation of σ = 0.6 oz. The line manager suspects that the machine may be overfilling the bottles, and so he will test the claim that the mean is greater than 12.2. A sample of 36 is selected; the sample mean is 12.5 oz, and the p-value is calculated to be $p = 0.0013$. What can we conclude from this test?

Note that in our example, we followed the **Five Step Method**:

Five Step Method for Hypothesis Testing

1. Set up the hypotheses H_0 and Ha.
2. Select the correct test and identify the significance level α.
3. Calculate the test statistic and p-value.
4. Use these results to make the decision about H_0.
   In particular, if $p < \alpha$, then we reject H_0.
5. Interpret the decision about H_0 in the context of the given problem to state a conclusion.

The Five Step Method streamlines and organizes our work by providing a template for performing a hypothesis test. Note that of the five steps involved, only Step 3 requires any calculation; and these calculations are programmed into the TI-83 and TI-84 calculators:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the test statistic and *p*-value for a Z-Test, go to the STAT menu, and then scroll over to the TESTS menu.

- If we are given summary statistics σ, $\bar{x}$ and n, then we select the Stats option.
  Enter the mean hypothesized in H_0 for μ_0; enter the values for σ, $\bar{x}$ and n.
  In the row marked as "μ: ", scroll over to the option that matches the alternative H_a.
  Put the cursor on **Calculate**, and press Enter. The test statistic z and *p*-value will be displayed.
  If you instead put the cursor on **Draw**, the calculator will show a graph of the normal distribution, with the area corresponding to the *p*-value shaded; this option also shows the test statistic z and *p*-value.

- If we are given raw data, go to the EDIT menu and enter the data into a list.
  Enter the hypothesized mean in H_0 for μ_0, the value for σ, and the list name (e.g. press 2nd 1 for list L1). In the row marked as "μ: ", scroll over to the option that matches the alternative H_a. Press **Calculate** and Enter to see the test statistic and *p*-value.

Example 9.11

Jeffrey, as an eight-year old, established a mean time of 16.43 seconds for swimming the 25-yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a significance level of $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution 9.11

We will follow the Five Step Method:

1. Set up the hypotheses: This is a test of a single population mean; and claim is that Jeffrey swims faster with the new goggles. For Jeffrey to swim faster, his time will be less than 16.43 seconds; the claim is that $\mu < 16.43$. This claim will be the alternative H_a, and we have:

$$H_0: \mu \geq 16.43 \qquad H_a: \mu < 16.43 \text{ (claim)}$$

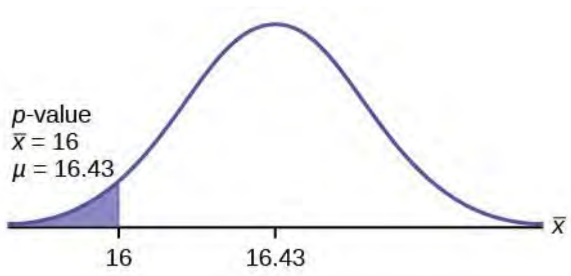
The "<" in the alternative means that this is a left-tailed test.

2. Since the swim times are normally distributed, and the population standard deviation is known ($\sigma = 0.8$), we use a **Z-Test**.

3. In the TI-84, go to the STAT menu and then the TESTS menu; select Z-Test (item 1). Enter the following:

   μ_0: 16.43 (this is the mean hypothesized in H_0)
   σ: 0.8
   $\bar{x}$: 16
   n : 15
   μ : < μ_0

   Put the cursor on Calculate and press Enter to get the p-value = 0.0187.
   The graph looks like:



4. Since *p-value* < α, we **reject *H_0***.
   Note that if H_0 is true, there is a 0.0187 probability that Jeffrey's mean time to swim the 25-yard freestyle will 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

5. 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.

Try It Σ

9.11 The mean throwing distance of a football for Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Test the coach's claim using a significance level of $\alpha = 0.05$. Assume that the throwing distances for footballs are normally distributed, with a standard deviation of $\sigma = 10$ yards.

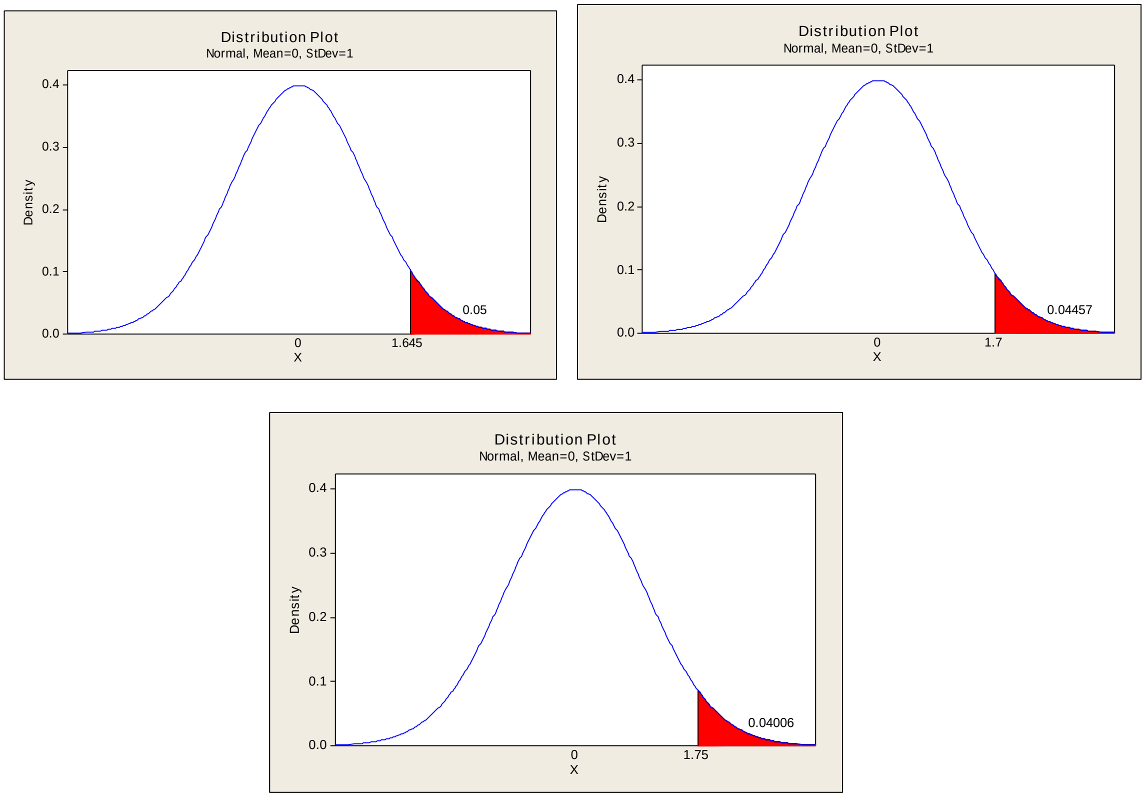
Critical Values

Before statistical software and graphing calculators became widely available, probabilities with the normal distribution had to be calculated by converting to z-scores and and using the normal table. This is time consuming, so statistical practitioners instead used the concept of a *critical value* to make the decision whether or not to reject H_0. We encountered critical values in Chapter 8; recall that the value z_α is the z-value so that the area to the *right* is exactly equal to α.

For example, suppose that we are performing a right-tailed z-test for a population mean, using a .05 significance level. Note that we can find $z_{.05}$ using the invNorm function in the calculator:

$$z_{.05} = \text{invNorm}(.95) \approx 1.645$$

We can restate this by saying $P(z > 1.645) = 0.05$. Note also that as the z value increases, the area to the right of z decreases; this is easy to see using graphs:



It follows that that the area to the right of z is less than 0.05 if and only if $z > 1.645$. Thus: **the *p*-value for the test is less than .05 if and only if the test statistic *z* is greater than .05**. We would then refer to the area to the right of $z_{.05} = 1.645$ as the ***rejection region*** for the test, since we would reject H_0 if and only if the test statistic z is in this region.

Of course, we could find the critical value and rejection region for any type of test, and for any significance level. But the most tests use the significance levels α = .10, .05 or .01. The rejection region for z-tests using these significance levels is summarized in the table below:

	Rejection Regions for Z-Test		
Significance Level	**Left-Tailed**	**Right Tailed**	**Two-Tailed**
$\alpha = .10$	$z < -1.28$	$z > 1.28$	$z < -1.645$ or $z > 1.645$
$\alpha = .05$	$z < -1.645$	$z > 1.645$	$z < -1.96$ or $z > 1.96$
$\alpha = .01$	$z < -2.33$	$z > 2.33$	$z < -2.576$ or $z > 2.576$

Example 9.12

According to a recent study, the mean cost of a heart-bypass operation is slightly less than \$26,100, and approximately 230,000 operations are performed annually. A sample of 36 bypass operations showed a mean cost of \$25,000. Assume the population standard deviation is \$2400. Develop and test an appropriate hypothesis to see whether or not the mean bypass operation is less than \$26,100 using $\alpha = 0.05$.

Solution 9.12

The claim being tested is $\mu < 26{,}100$, which will be the alternative hypothesis, H_a. So the hypotheses are: H_0: $\mu \geq 26{,}100$ vs. H_a: $\mu < 26{,}100$.

We are testing a claim about a mean μ, and the population SD σ is known, so we use a Z-Test.

The test statistic is $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{25000 - 26100}{2400/\sqrt{36}} = -2.75$. This is a left-tailed test, so the critical value will be the z-score that has an area of .05 to the *left*; i.e. the critical value is $z_{.05} = -1.645$. So we will reject H_0 if $z < -1.645$.

Since $z = -2.75 < -1.645$, we **reject H_0.** There is sufficient evidence to conclude that $\mu < 26{,}100$. That is, there is significant evidence that the mean cost of a bypass operation is less than \$26,100.

Summary of Z-Test

Used to test a claim about a population mean μ.

Assumptions: Data comes from a random sample
Sampling from a normal distribution, or sample size $n \geq 30$
The population SD, σ, is known.

Test statistic: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ Calculator function: **Z-Test**

P-value method versus critical value method

The first few steps are the same for both methods in hypothesis testing. After finding the test value, the difference comes between the two methods. P-value is the probability of the test value happening. Critical value is the value from the distribution that separates the rejection region from the non-rejection region. Here is a recap of the rejection rules for both methods:

P-value method: Reject Ho if p-value < α

Critical Value method: Reject Ho if test value is in rejection region

Example 9.13

According to an article in the Wall Street Journal, they claim that the mean time it takes a taxi from Manhattan to LaGuardia is now more than 35 minutes. Suppose new study using 40 taxis has an average of 42 minutes travel time from Manhattan to LaGuardia. Assume the population standard deviation is 5.1 minutes. Develop and test an appropriate hypothesis to see whether or not the mean travel time is more than 35 minutes using $\alpha = 0.05$.

Solution 9.13

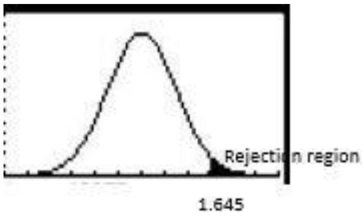
H_0: $\mu \leq 35$

H_a: $\mu > 35$ (claim)

$\alpha = 0.05$; right – tailed

Test Value: Z-test

μ_0: 35 (this is the mean hypothesized in H_0)
σ: 5.1
$\bar{x}$: 42
n : 40
$\mu : > \boldsymbol{\mu_0}$ $Z = 8.68$

p-value = 1.99E -18 which is 0.0000000000000000199 Reject Ho because p-value < .05	$Z_{.05}$ = -invnorm(.05,0,1) = 1.645 (critical value)  Reject Ho because 8.68 is in the rejection region

There is enough evidence at 5% level of significance to support the claim that the mean travel time from Manhattan to LaGuardia is now more than 35 minutes.

9.3 | T-Test for a Single Population Mean

In the preceding section, we tested claims involving a population mean μ. Throughout that section we assumed that we knew the population SD, σ. But in actual applications, it is often the case that σ is unknown; and in these cases it is necessary to estimate σ by the sample standard deviation, *s*.

If we are sampling from a normal distribution, the sampling distribution for $\bar{x}$ will be approximately normal. But for a small sample, the sample standard deviation may be a poor estimate for σ. To compensate for this potential error, we use the *t*-distribution introduced in Chapter 8 to calculate our *p*-values and critical values for our hypothesis tests. Recall that this distribution is bell-shaped like the normal distribution, but with a larger variance. Moreover, there is a different *t*-distribution for every sample size, and the smaller the sample size the larger the variance. Finally, as the sample size gets very large, the *t*-distribution approaches the normal distribution.

A test for a population mean μ that is performed under these conditions (i.e. in which σ is unknown) is called a **T-Test**. We start with an example:

Example 9.14

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. To test his claim, he selects a random sample of ten statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He uses this data to conduct a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution. Does this sample data provide evidence that the mean score for *all* students is more than 65?

Solution 9.14

This is a test of a single population mean. Since the instructor thinks the average score is *more* than 65, his claim is represented by the alternative hypothesis, and we have:

$$H_0\text{: } \mu \leq 65 \qquad H_a\text{: } \mu > 65$$

A 5% level of significance means that $\alpha = 0.05$. 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 we must use a *t*-test.

Therefore, the distribution for the test is t_9 where $n = 10$ and so df = 10 - 1 = 9. The mean of the distribution will be $\mu_{\bar{x}} = \mu = 65$. We do not have an explicit formula for the standard deviation of the sampling distribution as we did for normal distributions. But the test statistic and *p*-value can easily be calculated using the TI-84:

Using the TI-83, 83+, 84, 84+ Calculator

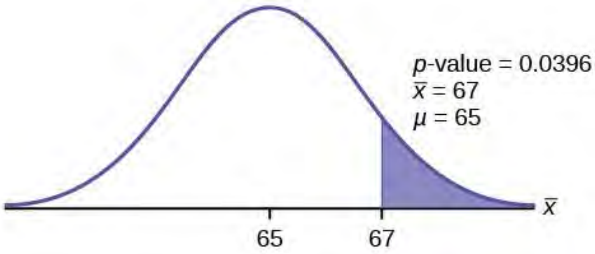
To calculate the p-value and test statistic for a T-Test, follow the same basic instructions as for a Z-Test, but in the TESTS menu we will select **T-Test**.

Go the STAT menu, and then to the TESTS menu; select T-Test (option 2). In this problem, we are given raw data, so we first go to the STAT menu, select the Edit menu, and press ENTER. Next we enter the 10 data values into a list, say L1.

Press STAT, arrow over to the TESTS menu, and select **T-Test**. For Inpt:, select the Data option, and press ENTER. Then, enter the following:

μ_0 : 65 (This is the value of μ that is hypothesized in H_0)
List: L1 (the name of the list where the data is entered)
Freq: 1 (this will be 1 unless we are given the data as a frequency distribution)
μ : $> \mu_0$ (this is the alternative hypothesis)

Arrow down to Calculate and press ENTER.
The display screen will show the test statistic, $t = 1.978$ and the *p*-value, $p = 0.0396$.
It is useful to have a graphical representation of this *p*-value; this is a right-tailed test, so the *p*-value is P($\bar{x}$ > 67), the area to the *right* of the observed sample mean:



Next we make a decision about H_0:
Since *p* –value = 0.0396 < 0.05 = α*,* the decision is to **reject *H*₀**.

And since we reject μ = 65, we have sufficient evidence to conclude that *H*a is true.
At a 5% level of significance, the sample data provide sufficient evidence that the mean test score is more than 65.

NOTES:

- If we select Draw instead of Calculate, the calculator will graph the area corresponding to the *p*-value; the test statistic and *p*-value will still appear on this screen.
- The output screen also shows the alternative hypothesis, μ > 65 at the top. It is always a good idea to double check that the correct alternative hypothesis was selected.
- The output screen also displays the sample mean , sample standard deviation and sample size: $\bar{x}$ = 67, *s* = 3.1972, and *n* = 10, respectively. One of the advantages of using the data option is that the calculator will automatically calculate the key sample statistics for us. We just need to enter the data!

Before starting another example we will take this opportunity to show another way to calculate the *p*-value. First, we can quickly use 1-VarStats to calculate the sample mean and sample standard deviation. Then we calculate the *t*-statistic using the formula:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{67 - 65}{3.1972 / \sqrt{10}} = 1.978 .$$

This is a right-tailed test, so the p-value = P($\bar{x}$ > 67); and this probability can be rewritten as *p-value* = P($\bar{x}$ > 67) = P(t > 1.978).

The latter probability can be calculated using the tcdf function in the TI-84 calculator: Press 2nd VARS to access the DISTR menu, and select the **tcdf** function; this function is simila1r to the normalcdf function that we have used many times. However the tcdf function needs only three inputs – the left endpoint, right endpoint and degrees of freedom. For this problem, we have

$$p = \text{P}(\bar{x} > 67) = \text{P}(t > 1.98) = \text{tcdf}(1.978,\ 10\text{^}99,\ 9) \approx 0.0396.$$

Try It Σ

9.13 It is believed that a stock price for a particular company will grow, on average, at a rate of $5 per week with a standard deviation of $1. An investor claims that the stock won't grow as quickly. The changes in stock price are recorded for ten weeks and are as follows:

$4, $3, $2, $3, $1, $7, $2, $1, $1, $2.

Use this data, along with a 5% level of significance, to test the claim that the mean

Example 9.15

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Using α = .05, does this data provide evidence that the average conductivity of this type of glass is greater than one? Assume the population is normal.

Solution 9.15

We follow the Five Step Method:

1. The claim being tested is that the average conductivity of the selected glass is greater than one; i.e. the claim is that μ > 1, which will be the alternative hypothesis. So the hypotheses will be:

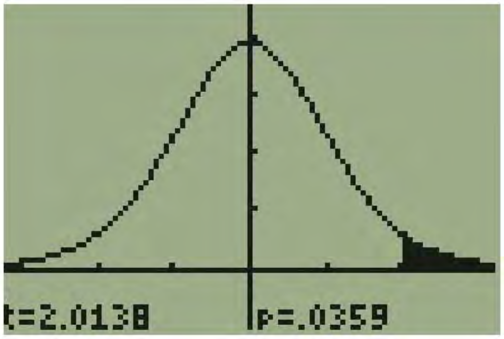
H_0: $\mu \le 1$ vs. H_a: $\mu > 1$

2. We are testing a sample mean and the population standard deviation is unknown. Therefore, we need to use a Student's-*t* distribution. (Assume the underlying population is normal)

3. Do the calculations: In the TI-84, go to the STAT menu, then to the Edit menu; enter the data into one of the lists. Go to the STAT menu, then to the TESTS menu, and select T-Test.
Select the data option and enter the following:

μ_0 : 1 (This is the value of μ that is hypothesized in H_0)
List: Enter the name of the list where the data is entered)
Freq: 1
μ : $> \mu_0$ (this is a right-tailed test)

Arrow down to Calculate, and press Enter to get $t = 2.014$ and *p-value* = 0.0359.
Alternatively, arrow down to Draw, and press enter to see a graphical representation of the area corresponding to the *p*-value; it is the area in the right tail:



t=2.0138 p=.0359

4. Since *p-value* = .0359 < .05, we reject H_0 at the .05 level of significance.

5. Thus the results are statistically significant; there is sufficient evidence to conclude that the mean conductivity for this type of glass is greater than 1.

In our last example for this section, we demonstrate how critical values can be used for a *t*-test.

Example 9.16

An engineer at Duracell has designed a new battery for industrial use. The old style battery had an average life of 29.2 hours. The engineer claims the new battery is an improvement over the old one since the new design has a longer life. A sample of 20 new batteries finds a mean life of 32.3 with a standard deviation of 3.9 hours. Test the engineer's claim using $\alpha = 0.01$.

Solution 9.16
We again use the Five Step Method.

The engineer’s claim is that the new battery is an improvement, meaning that it will have a longer average life. So the claim is that $\mu > 29.2$ hours, and the hypotheses are:

$$H_0: \mu \leq 29.2 \quad \text{vs.} \quad H_a: \mu > 29.2$$

We are testing a claim about a population mean μ and the population standard deviation is unknown, so we use a *t*-test.

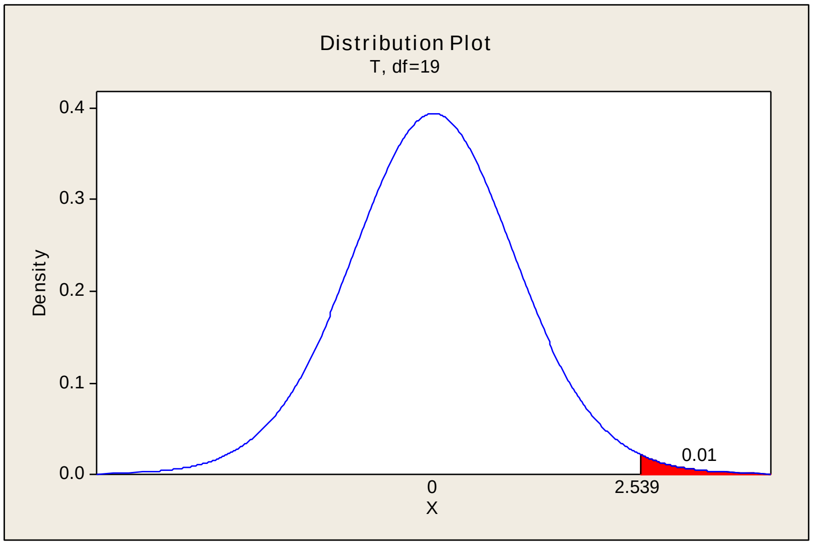
Using the given sample statistics, the test value is $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = \dfrac{32.3 - 29.2}{3.9/\sqrt{20}} = 3.55$.

Next we find the critical value; the critical value will be $t_{.01}$, the *t* value so that the area to the right of it is exactly 0.01. This value can be found using the invT function; this function works just like the invNorm function that we have used many times. The inputs will be the cumulative probability and the degrees of freedom. In this problem, *n* = 20, so df = 19. And since the area in the right tail will be .01, the area to the left – the cumulative probability – will be .99.

Press 2nd VARS to access the DISTR menu, and select the invT function. The critical value is:

$t_{.01}$ = invT(.99, 19) = 2.539.

So we will reject H_0 if and only if $t > 2.539$; the ***rejection region*** is shown below:



Since $t = 3.55 > 2.539$, we **reject H_0** at the 0.01 significance level.

Thus, the data provides sufficient evidence that μ > 29.2 hours. There is sufficient statistical evidence to support the engineer's claim.

Summary of T-Test

Used to test a claim about a population mean μ.

Assumptions: Data comes from a random sample
Sampling from a normal distribution
The population SD, σ, is not known.

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ Calculator function: **T-test**

9.4 | Hypothesis Test for a Single Population Proportion

In this section we will discuss tests for a single population proportion p. If we are testing a claim about a population proportion, then the appropriate random variable will be $\hat{p}$, the sample proportion. From Chapter 8, we know that if we select random samples of size n, where both

$$np > 10 \quad \text{and} \quad n(1-p) > 10,$$

then the sampling distribution for $\hat{p}$ will be approximately normal. Moreover the mean and standard deviation for this sampling distribution are $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, respectively.

Using this information and the Five Step Method, we can test claims like the one in the following example:

Example 9.17

A consumer group asserts that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion more than 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 60 of the households have three cell phones; this data is then used to test the claim that more than 30% of households have three cell phones. At the 5% significance level, is there sufficient evidence to support the company's claim?

Solution 9.17
The phone company's claim is that $p > 0.30$, which will be represented by the null hypothesis H_a. The alternative hypothesis will be contradictory statement, giving us the hypotheses:

$$H_0\text{: } p \leq 0.30 \qquad H_a\text{: } p > 0.30$$

Since the claim involves a population proportion, the appropriate sample statistic will be a sample proportion, $\hat{p}$. Note that np =150(.30) = 45 and $n(1 - p)$ = 150(.70) = 105, so the sampling distribution for $\hat{p}$ will be approximately normal with mean μ = 0.30 and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.30(.70)}{150}} \approx 0.03742$. Note that the sample proportion is $\hat{p}$ = 57/150 = 0.38. And since this is a right-tailed test, the p-value is:

$$p\text{–value} = \text{P}(\hat{p} > 0.38) = \text{normalcdf}(0.38, 10\text{^}99, .30, .03742) = 0.0163.$$

This p-value is less than the significance level, so our decision is to **reject Ho**. Moreover, this p-value tells us that if the null hypothesis is true, there is only about a 1.6% chance that the sample proportion $\hat{p}$ for a randomly selected sample will either be at least as far off from the expected proportion of .30.

Thus, at the α = .05 level of significance, the sample data provides sufficient evidence that the percentage of households with three cell phones is more than 30%. In other words, we have significant statistical evidence that the company's claim is correct.

As was the case with the Z-test and T-Test, the *p*-value and test statistic for a hypothesis test involving a proportion can be easily calculated using the TI-83 or TI-84 calculator:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the test statistic and *p*-value for a test involving a proportion, we go to the STAT menu, and then to the TESTS menu. Select **1-PropZTest** (item 5).
Enter the proportion hypothesized in H_0 as p_0; this must be a decimal.
Enter the number of successes in the sample as x; if a sample proportion $\hat{p}$ is given, then use $x = n\hat{p}$. If necessary, round this value to the nearest whole number.
Enter the sample size for *n*, and select the appropriate alternative on the last line.
Put the cursor on Calculate and press Enter – the test statistic and *p*-value will be displayed.

Caution: The output screen will show two values labeled as *p*. The first one is the *p*-value for the test. The second is the sample proportion $\hat{p}$ (but the ^ is very small and easy to miss)

Try It Σ

9.16 Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that percentage is actually lower. A random sample of 200 American adults are surveyed, of which 174 report owning cell phones. Test the manufacturer's claim using a 5% level of significance. State the null and alternative hypothesis, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

Example 9.18

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Solution 9.18

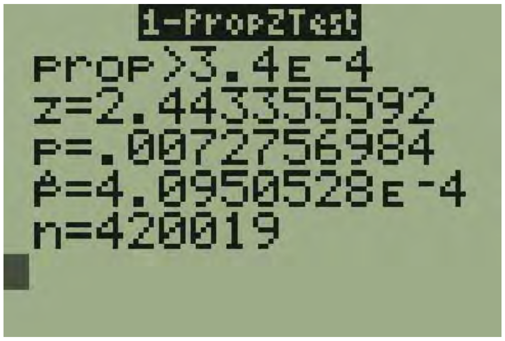
We will follow the Five Step Method.

1. Let *p* be the proportion of cell phone users that develop brain cancer; then the claim is that $p > 0.00034$. This claim will be represented by the alternative hypothesis, so we get:

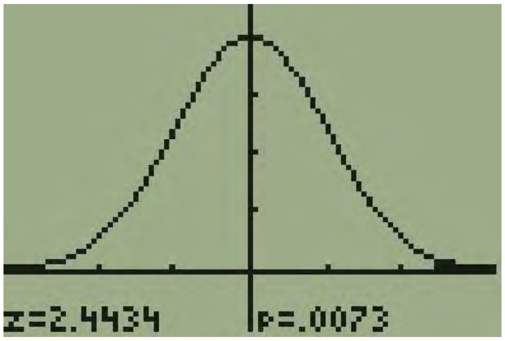
   $$H_0\text{: } p \le 0.00034, \qquad H_a\text{: } p > 0.00034$$

2. We are testing a claim about a proportion, so we use a one-proportion *z*-test.
   Note that the sample is sufficiently large because $np = 420{,}019(0.00034) = 142.8$, and $n(1-p) = 420{,}019(0.99966) = 419{,}876.2$.

3. In the TI-84, we go to STAT >> TESTS and select 1-PropZTest.
   Enter $p_0 = 0.00034$, $x = 172$ and $n = 420{,}019$. Select the ">" option for the alternative, place the cursor either on Calculate or Draw and press Enter to get the following output:



```
1-PropZTest
prop>3.4E-4
z=2.443355592
p=.0072756984
p̂=4.0950528E-4
n=420019
```



4. Since the p-value = 0.0073 is greater than our alpha value = 0.005, we do not reject H_0.

5. We conclude that there is not enough evidence to support the claim of higher brain cancer rates for cell phone users.

Finally, if we commit a Type I error, then we think that the rate of brain cancer is no worse for cell-phone users than for non-cell phone users, when in fact the rate is *higher* for cell-phone users. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.

Example 9.19

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that 207,754 rapes occur each year (male and female), on average, for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Solution 9.19

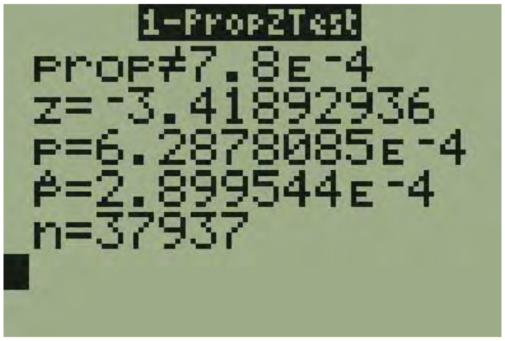
We again follow the Five Step Method.

1. The claim is that the proportion of sexual assaults in Daviess County, KY is significantly different from the national average. The proportion for the nation is 207,754/268,608,618, which is about 0.00078. This means that the claim is $p \neq 0.00078$, and so the hypotheses are:

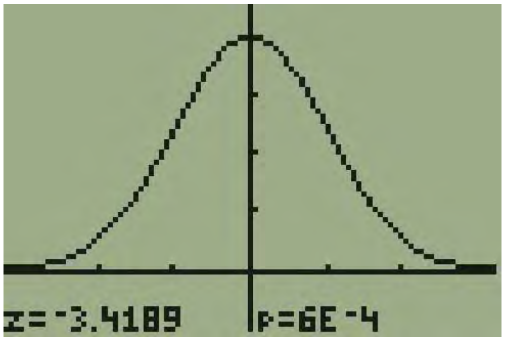
$$H_0\text{: } p = 0.00078, \quad H_a\text{: } p \neq 0.00078$$

2. Since we are working with proportions, we will use a one-proportion z-test.

3. Go to STAT >> TESTS and select 1-PropZTest. Enter 0.00078 as the hypothesized proportion p_0, enter 11 for x and 37,937 for n. Select the ≠ option for the alternative. Scroll down to Calculate and press Enter to get the following output screen:



```
1-PropZTest
prop≠7.8E-4
z=-3.41892936
p=6.2878085E-4
p̂=2.899544E-4
n=37937
```



The figure on the right shows the output obtained by using the DRAW option.

4. Since the *p*-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should **reject the null hypothesis**.

5. The sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

Summary of One-Proportion Z-test

Used to test a claim about a population proportion *p*.

Assumptions: Data comes from a random sample
The sample size an hypothesized proportion satisfy: $np_0 > 10$ and $n(1 - p_0) > 10$.

Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$ Calculator function: **1-PropZTest**

KEY TERMS and FORMULA REVIEW

Hypothesis Test: A procedure for determining whether the hypothesis stated is a reasonable statement and should not 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 alternative hypothesis (notation Ha).

Level of Significance: In hypothesis testing, the Level of Significance is a threshold value – it is the maximum value allowed for a Type I error. And when the p-value is below the level of significance, the null hypothesis must be rejected.

***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.

Student's *t*-Distribution: Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of this random variable *t* are:

- It is continuous, bell shaped curve.
- The pdf is symmetrical about its mean of zero. The variance is more than that of the normal distribution.
- 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 sample size.
- The smaller the sample size, the greater the variation
- As the sample size *n* gets larger, the *t*-distribution approaches the standard normal distribution.

Type 1 Error: The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error: The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

T-Test: A test for a population mean, μ that is used when the population SD, σ, is unknown.

Z-Test: A test for a population mean μ that is used when the population SD, σ, is known.

CHAPTER REVIEW

All hypothesis tests use the same basic steps, which we have called the **Five Step Method:**

1. Set up the hypotheses H_0 and H_a.
2. Select the correct test and identify the significance level α.
3. Calculate the test statistic and p-value.
4. Use these results to make the decision about H_0.
5. Interpret the decision about H_0 in the context of the given problem to state a conclusion.

The first step is to set up the hypotheses. The null hypothesis H_0 and the alternative hypothesis H_a are contradictory claims about one or more unknown population parameters. There are a few basic rules for setting these up:

- The null hypothesis must include an equal sign.
- One of the hypotheses must represent the claim being tested.
- The hypotheses should be contradictory statements.

The second step is to select the correct test:

- When testing a population mean μ, where the population SD σ is known, we use a Z-Test
- When testing a population mean μ, with unknown population SD σ, we use a T-Test
- When testing a population proportion, p, we use a one-proportion z-test.

The third step is to calculate the test statistic and p-value. The test statistic is a measure of relative position, and the formula depends on the type of test being used (see below). The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample. We can get both of these very easily using the TI-84 family of calculators.

The fourth step is to make the decision about H_0. In a test, there are only two possible decisions: Either we reject H_0 or we do not reject H_0. This decision is made by comparing the p-value found in Step 3 to a predetermined significance level α according to the following rule:

- We reject H_0 at significance level α if $p < \alpha$.
- If $p \geq \alpha$, then we do not reject H_0.

The fifth step is to write a conclusion. After we make our decision about H_0, we interpret the decision in the context of the given problem to write a meaningful conclusion for the test in terms of the given problem. That is, we should write a statement, in plain English, that explains the result of the test. A few things to keep in mind:

- If we reject H_0, then there is sufficient evidence to conclude that H_0 is incorrect. Thus, rejecting H_0 provides evidence that the alternative hypothesis H_a is true.
- If we fail to reject H_0, then there is not sufficient evidence to conclude that the alternative hypothesis H_a is true.
- Failing to reject H_0 does **not** mean that we have evidence that H_0 is true. It simply means that the sample data have failed to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Summary of Tests:

Z-Test:

- Used to test a claim about a population mean μ when σ (the population SD) is known.
- Data comes from a random sample
- Sampling from a normal distribution, or sample size $n \geq 30$
- Test statistic: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
- Calculator function: **Z-Test**

	Rejection Regions for Z-Test		
Significance Level	**Left-Tailed**	**Right Tailed**	**Two-Tailed**
$\alpha = .10$	$z < -1.28$	$z > 1.28$	$z < -1.645$ or $z > 1.645$
$\alpha = .05$	$z < -1.645$	$z > 1.645$	$z < -1.96$ or $z > 1.96$
$\alpha = .01$	$z < -2.33$	$z > 2.33$	$z < -2.576$ or $z > 2.576$

T-Test:

- Used to test a claim about a population mean μ when σ (the population SD) is not known.
- Data comes from a random sample
- Sampling from a normal distribution
- Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$
- Calculator function: **T-test**

One-Proportion Z-test:

- Used to test a claim about a population proportion *p*.
- Data comes from a random sample
- The sample size and hypothesized proportion satisfy: $np_0 > 10$ and $n(1 - p_0) > 10$.
- Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$
- Calculator function: **1-PropZTest**

Chapter 9 Exercises

1. A test is being conducted to determine whether the mean speed of a cable internet connection is more than three Megabits per second.

a. What is the random variable? Describe in words.

b. State the null and alternative hypotheses for the test.

2. The mean entry level annual salary of an employee at a large company is $58,000. An economist wishes to test the claim that the mean entry level salary is higher for IT professionals in the company. State the null and alternative hypotheses.

3. A test will be conducted to test the claim that the mean number of children for American families is 2.

a. What is the random variable? Describe in words.

b. State the null and alternative hypotheses for the test.

4. A sociologist claims that the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. A tourism expert thinks that the proportion is actually less. Assume that the tourism expert's claim is tested using an appropriate hypothesis test.

a. What is the random variable? Describe in words.

b. State the null and alternative hypotheses for the test.

5. In a population of fish, it is widely believed that approximately 42% are female. A biologist will conduct a hypothesis test to determine if the proportion of female fish is less than 42%. State the null and alternative hypotheses for the test.

6. An article in the 1990's stated that the mean time spent in jail by a first–time convicted burglar was 2.5 years. A study was then done to see if the mean time in jail has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years. Assume that the distribution of the population is normal and that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be?

7. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the mean time spent on death is equal to 15 years, what would the null and alternative hypotheses be?

a. H_0: __________ b. H_a: __________

8. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that we are conducting a hypothesis test to determine if the true proportion of people in a given town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: _________ b. H_a: __________

9. A market analyst claims that the mean price of mid-sized cars in the midwest region is $32,000. A suitable hypothesis test will be conducted to determine if the claim is true. State the Type I and Type II errors in complete sentences.

10. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis H_0, in words, is: The surgical procedure will go well.

a. State the Type I and Type II errors in complete sentences.
b. Which is the error with the greater consequence?

11. The power of a test is 0.981. What is the probability of a Type II error?

12. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors..

13. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002.

a. What is the power of this test?
b. Which is the error with the greater consequence?

14. A population mean is hypothesized to be $\mu = 13$. A random sample of 20 measurements is selected and the sample mean is 12.8, and the sample standard deviation is two. What distribution should you use to perform the hypothesis test? Assume the underlying population is normal.

15. A population has a standard deviation of $\sigma = 5$. A test is to be conducted to test the claim that the mean is $\mu = 25$. A sample of 108 individuals yields a sample mean of 24. What distribution should you use to perform a hypothesis test?

16. It is thought that 42% of respondents in a taste test would prefer Brand *A*. Suppose we want to test the claim that the true population proportion is less than 42%. We select a random sample of 100 people, and find that 39% preferred Brand *A*. What distribution should we use to perform a hypothesis test?

17. Suppose we are conducting a hypothesis test of a single population mean using a Student's *t*-distribution. What must we assume about the distribution of the data?

18. Suppose we are conducting a hypothesis test for a single population proportion. What must be true about the quantities of np and $nq= n(1-p)$?

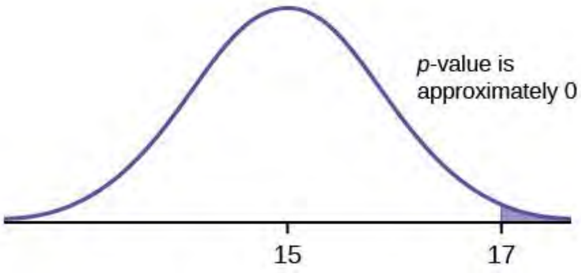
19. It is believed that the mean height of high school students who play basketball in a large urban district is 73 inches with and standard deviation of $\sigma = 1.8$ inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches?

a. State the null and alternative hypotheses for the test.
b. Suppose that the *p*-value is almost zero; state an appropriate conclusion.

20. It is conjectured that the mean age of graduate students at a large state university is at most 31 years. To test this claim, a random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is two years. Assume that the ages are normally distributed, and conduct an appropriate test.

a. State the null and alternative hypotheses for the test.
b. Assume that the *p*-value is 0.0366. Is the data significant at the .05 level? Explain.
c. Is the data significant at the .01 level? Explain.

21. Consider the following graph:



a. Does the shaded region represent a low or a high *p*-value compared to a level of significance of 1%?
b. Is this a right-tailed, left-tailed or two-tailed test? Explain.

22. Consider the statement, "If you do not reject the null hypothesis, then H_0 must be true." Is this statement correct? Explain why or why not.

23. Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year is selected. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

a. Is this a test for a mean or a proportion?
b. What symbol represents the random variable for this test?
c. In words, define the random variable for this test.
d. Is the population standard deviation known and, if so, what is it?
e. Find the following:
    a. $\bar{x}$ = _______ b. σ = _______ c. s = _______ d. n = _______
f. Since both σ and *s* are given, which should be used? Explain.
g. Which test should be used?
h. Find the *p*-value.
i. Using a significance level of $\alpha = 0.05$, what would be the decision about H_0?
j. Interpret this decision to state a conclusion for the test.

24. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the mean time on death row for all inmates is at most 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses: H_0: _____ H_a : _______
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Find the following:
(i) $\bar{x}$ = __ (ii) s = ___ (iii) n = __
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. Using a significance level of $\alpha = 0.05$, what would be the decision about H_0?
l. Interpret this decision to state a conclusion for the test.

25. Suppose that we conduct at test with hypotheses H_0: $\mu = 9$ and H_a: $\mu < 9$.
Is this a left-tailed, right-tailed, or two-tailed test?

26. Suppose that we conduct at test with hypotheses H_0: $\mu \leq 6$ and H_a: $\mu > 6$.
Is this a left-tailed, right-tailed, or two-tailed test?

27. Suppose that we conduct at test with hypotheses H_0: $p = 0.25$ and H_a: $p \neq 0.25$.
Is this a left-tailed, right-tailed, or two-tailed test?

28. A certain brand bottle of water is labeled as containing 16 fluid ounces of water. You believe that the bottles contain less than 16 oz. Suppose that you will conduct a hypothesis test for your claim. Would this be a left-tailed, right-tailed, or two-tailed test?

29. A golf pro claims that his mean golf score is 63. A sports writer wants to test the claim that his mean score is higher 63. Would this be a left-tailed, right-tailed, or two-tailed test?

30. A bathroom scale claims to be able to identify correctly any weight within a pound. You suspect that it cannot be that accurate and will conduct an appropriate test to test your claim. Would this be a left-tailed, right-tailed, or two-tailed test?

31. You flip a coin and record whether it shows heads or tails. The coin is supposed to be balanced, but you suspect that the probability of getting for this particular coin is less than 50%. If you were to test your claim, would this be a left-tailed, right-tailed, or two-tailed test?

32. If the alternative hypothesis has a not equals ($\neq$) symbol, what type of test does this signify?

33. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

34. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

35. Each of the following statements refers to a claim for a hypothesis test; so each could be either the null hypothesis or the alternative hypothesis. In each case, state the null hypothesis H_0 and the alternative hypothesis H_a, in terms of the appropriate parameter (either μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults in Los Angeles ride the bus to work.
f. The mean number of cars a person owns in her lifetime is at most ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation of six weeks each year.
i. Fewer than 11% of all American women will develop breast cancer.
j. The mean tuition cost for private American universities is more than $20,000 per year.

36. Refer to problem 35; classify each test as either left-tailed, right-tailed or two-tailed.

37. Refer to problem 35; for each test, state the Type I and Type II errors in complete sentences.

38. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. This data is used to test the claim that more than thirty percent of the teen girls smoke to stay thin. The alternative hypothesis is:

a. $p < 0.30$ b. $p \leq 0.30$ c. $p \geq 0.30$ d. $p > 0.30$

39. A statistics instructor wishes to test the claim that fewer than 20% of students at her college attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis would be:

a. $p = 0.20$ b. $p > 0.20$ c. $p < 0.20$ d. $p \leq 0.20$

40. Some years ago, an organization reported that on average, teenagers spent 4.5 hours per week on the phone. The organization thinks that the current mean is higher. A test will be conducted to test whether the current mean is higher than the previously reported value of 4.5 hours per week. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. The null and alternative hypotheses are:

a. Ho: $\bar{x} = 4.5$, Ha: $\bar{x} > 4.5$ b. Ho: $\mu \geq 4.5$, Ha: $\mu < 4.5$

c. Ho: $\mu = 4.75$, Ha: $\mu > 4.75$ d. Ho: $\mu = 4.5$, Ha: $\mu > 4.5$

41. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is "the drug is unsafe." What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

42. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. To test this claim she surveys 84 of her students and finds that 11 of them attended the midnight showing. What is the Type I error for the test?

a. To conclude that the percent of EVC students who attended is at least 20%, when in fact, it is less than 20%.
b. To conclude that the percent of EVC students who attended is 20%, when in fact, it is 20%.
c. To conclude that the percent of EVC students who attended is less than 20%, when in fact, it is 20% or more.
d. To conclude that the percent of EVC students who attended is less than 20%, when in fact, it is less than 20%.

43. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. A test is conducted to test the claim that that LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average. Which of the following statements is true?

a. The Type II error is to not reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours is more than seven hours.
b. The Type II error is to not reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours is at most seven hours.
c. The Type II error is to not reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours is at least seven hours.
d. The Type II error is to not reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours is less than seven hours.

44. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Suppose that a hypothesis test is conducted to test the claim that the mean hours is more than 4.5 hours per week. What is the Type I error?

a. To conclude that the current mean is higher than 4.5 hrs/week, when in fact, it is higher
b. To conclude that the current mean is more than 4.5 hrs/week, when in fact, it still 4.5 hrs.
c. To conclude that the current mean hours per week is 4.5 hrs/week, when in fact, it is higher
d. To conclude that the current mean is no higher than 4.5 hrs/week, when in fact, it is not higher

45. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. A test is conducted to test the claim that that LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average. The distribution to be used for this test is: $X \sim$ _______

a. $N\left(7.24, \dfrac{1.93}{\sqrt{22}}\right)$ b. $N(7.24, 1.93)$ c. t_{22} d. t_{21}

46. The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses: $H0$: ______ Ha: ________
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Find the following:
(i) $x =$ ______ (ii) $n =$ ______ (iii) $\hat{p}$ = _____
g. What distribution is used for this test?
h. Find the p-value.
j. If the significance level is a = 0.05, what is the decision about H_0?
k. Interpret the decision to write a conclusion for the test.

For the remaining exercises, use the FIVE STEP METHOD and a suitable hypothesis test to answer each question. For problems involving a Student's-t distribution, you may assume that the underlying population is normally distributed.

47. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using $\alpha = 0.05$, is the data highly inconsistent with the claim?

48. From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant of around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

49. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level?

50. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

51. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows:

12; 4; 15; 3; 11; 8; 6; 8.

Let x = the number of sick days they took for the past year. Does this data provide evidence to support the personnel department's claim the mean number of sick days differs from ten? Conduct an appropriate test using a significance level of $\alpha = .05$.

52. In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was 10. Using a 5% significance level, does this data provide evidence that the mean number of hours worked by women each week has increased?

53. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

54. A Nissan Motor Corporation advertisement read, "The average man's IQ is 107. The average brown trout's IQ is 4. So why can't man catch brown trout?" Suppose you believe that the brown trout's mean IQ is greater than four. You catch 12 brown trout. A fish psychologist determines the IQ's for the fish as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Use this data to conduct a hypothesis test for the claim that the mean IQ for brown trout is greater than 4.

55. Refer to the previous exercise. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the1 brown trout's mean IQ is *different* from 4.

56. According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

57. A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

58. The mean work time each week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. The data is as follows:

70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Use this data to test the claim that the mean number of hours worked each week is less than 60.

59. Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

60. A recent report states that 68% of online courses taught at community colleges nationwide were taught by full-time faculty. To test if this proportion is the same for California, Long Beach City College (LBCC) in California, was randomly selected for comparison. In that same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Does this data provide evidence that the percentage of online courses at community colleges that are taught by full-time faculty is different from 68% in California?

61. According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Use this data to test the claim that the proportion of NYC adults who has decreased.

62. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test to test the instructor's claim.

63. Nationwide, registered nurses earn an average annual salary of $69,110. A survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct an appropriate test to test the claim that the mean salary for nurses in California is higher than the national average.

64. La Leche League International reports that the mean age of weaning a child from breast-feeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

65. Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Does this data provide sufficient evidence to conclude that more than thirty percent of teen girls smoke to stay thin? After conducting the test, your decision and conclusion are

a. Reject *H0*: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject *H0*: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject *H0*: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject *H0*: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

66. A statistics instructor believes that fewer than 20% of students at her community college attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

67. Previously, an organization reported that teenagers spent an average of 4.5 hours per week on the phone. The organization thinks that the mean is now higher. To test the claim, 15 randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Use this data to test the claim that the current mean is greater than 4.5 hours. At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75
b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

68. According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use this data with a significance level of 0.05 to test whether the local high school's percentage is less than 18%.

69. A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate? In other words, at the .05 level, does this data provide evidence that the true proportion of families who own stock differs from 48.8%?

70. Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

71. The US Department of Energy reported that 51.7% of homes nationwide were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does this data provide evidence to support the claim that the percentage of homes heated by natural gas differs from the national average? Use a significance level of $\alpha = 0.05$ to test.

72. For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use an $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

73. The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

74. A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

75. A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

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

At the $\alpha = 0.05$ level, can it be concluded that the sample mean is higher than 5.8 visits per year?

76. According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class at a large university resulted in the following family sizes:

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

At $\alpha = 0.05$ level, is there evidence that the class' mean family size greater than the national average? Does this mean that the Almanac result is incorrect? Why or why not?

77. The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical; they claim that freshmen study less than 2.5 hours per day on average. To test the claim the class selected a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is there evidence to support the Stats class' claim?

78. A billing company that collects bills for medical offices in the surrounding area claims that there is a change in the percent of bills being paid by Medicare. In the past, the percentage that is paid by Medicare is 30%. A study of 7500 recent bills shows that 33% of these bills are being paid by Medicare. Test the claim using a 5% level of significance. Show all 5 steps.

REFERENCES

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking- rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013). Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013). Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013). Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June27, 2013).

Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at http://toastmasters.org/artisan/ detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/ 3868.htm (accessed June 27, 2013).

“Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault (accessed June 27, 2013).

10 | HYPOTHESIS TESTING WITH TWO SAMPLES

Figure 10.1 If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Chapter 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 when the population standard deviations are known.
- Conduct and interpret hypothesis tests for two population means when the population standard deviations are unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.
- Construct and interpret two-sample confidence intervals.

Introduction

Studies often require the comparison of two unknown population means or two unknown population proportions. For example, medical researchers have done several studies to measure the effect aspirin has in preventing heart attacks. Typically, the treatment group is given aspirin and the control group is given a placebo. Then the heart attack rate for each group is monitored and compared 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.

Now that we have learned how to conduct hypothesis tests on single means and single proportions, we will expand our methods to test claims involving two unknown means or two unknown proportions. The basic framework – the Five Step Method – will be the same, but the sampling distributions in which we do the calculations will be slightly different.

Here we will be comparing parameters from two different populations, using random samples chosen from each. In the first three sections we will assume that the samples selected are **independent** of one another. That is, the sample values selected from one population are not related in any way to sample values selected from the other population. In Section 10.4 we will compare two population means using "**matched pairs**"; for this test the data consist of two samples that are dependent on one another.

NOTE

This chapter relies heavily on either a calculator or a compute to calculate the degrees of freedom, the test statistics, and *p*-values. The TI-83+ and TI-84 calculators have utilities for doing tests involving two independent samples. Instructions for the TI-83+ and TI-84 calculators are included, as well as formulas for the test statistic of each test. We will also be able to use these calculators to do the calculations related to matched pairs tests.

This chapter will develop the following hypothesis tests:

- Test of two population means using independent samples (either 2SampTTest or 2SampZTest)
- Test of two population proportions using independent samples (2PropZTest)
- Test of two population means using a Matched Pairs design (samples are dependent).
- Test of the two population proportions by testing one population mean of differences.

10.1 | Two Population Means with Known Standard Deviations

Suppose we want to test a claim about two unknown population means, μ_1 and μ_2, where μ_1 is the mean for population 1 and μ_2 is the mean for population 2. Note that any inequality relating these two parameters can be rewritten in terms of $\mu_1 - \mu_2$. For example, the statement $\mu_1 < \mu_2$ can be rewritten as $\mu_1 - \mu_2 < 0$. Similarly, $\mu_1 > \mu_2$ can be rewritten as $\mu_1 - \mu_2 > 0$. Thus, we can really think of claims comparing two different population means, μ_1 and μ_2, as claims involving a single new parameter $\mu_1 - \mu_2$. An obvious point estimate for this parameter is the difference of sample means, $\bar{x}_1 - \bar{x}_2$, where $\bar{x}_1$ and $\bar{x}_2$ are calculated from random samples selected from populations 1 and 2, respectively.

Assumptions:

1. We are using independent, random samples selected from the two populations.

2. For the two populations from which we are sampling:

   - If the sample sizes are small, the distributions from which we are sampling should be approximately normal.
   - If the sample sizes are both large ($n \geq 30$), then the distributions are not important; they need not be normal.

Since the parameter $\mu_1 - \mu_2$ is estimated by the difference of sample means, $\bar{x}_1 - \bar{x}_2$, we will be working with the sampling distribution for the random variable $\bar{X}_1 - \bar{X}_2$. This random variable represents the differences of all possible sample means when random samples of size n_1 and n_2 are selected from the respective populations. Assuming that the population SD's for the two populations are σ_1 and σ_2, respectively the basic properties of this random variable are as follows:

Sampling Distribution for Difference of Sample Means

Random Variable: $\bar{X}_1 - \bar{X}_2$

Mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ Standard Deviation: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$

When we are sampling from normal distributions, or the sample size is large, then the sampling distribution is approximately normal: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}})$

Some important notes:

- μ_1 and μ_2 are the unknown population means.
- σ_1 and σ_2 are the population standard deviations.
- n_1 and n_2 are the sample sizes.
- The variance of the distribution for $\overline{X}_1 - \overline{X}_2$ is the sum of the variances of the sampling distributions for $\overline{X}_1$ and $\overline{X}_2$.

When testing claims about a single population mean, we used one of two sampling distributions: either a normal distribution (*z*-test) or a *t*-distribution, depending on whether or not the population standard deviation is known. This will be the case for two-sample tests as well. When both of the population standard deviations are known, we will use a **Two-Sample Z-Test**; the test statistic for this test is: $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$. Note that for most tests, we will start with a null hypothesis that includes an equal sign; that is, our null hypothesis will generally include the case $\mu_1 - \mu_2 = 0$. So the numerator will be simplified, and we have the test statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{{\sigma_1}^2}{n_1} + \dfrac{{\sigma_2}^2}{n_2}}}$$

Once we calculate our test statistic, we could use the fact that the sampling distribution for $\overline{X}_1 - \overline{X}_2$ is approximately normal and either compute the p-value or use an appropriate critical value for our test. However, the TI-83 and TI-84 calculators have a built-in utility that will give us both the test statistic and p-value for our test.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the test statistic *z* and *p*-value for a Two-Sample Z-test:

Go to the STAT, menu and then to TESTS. Select option 3, which is **2-SampZTest**.

Enter the population SD's, the sample means and the sample sizes; select the appropriate alternative hypothesis. Arrow down to Calculate and press ENTER. The test statistic and *p*-value will appear on the output screen. If we select the Draw option instead of Calculate, we can view a graph showing the region corresponding to the *p*-value.

NOTE: The first inputs we are prompted for are the population standard deviations, σ_1 and σ_2. If these are not known, we should not be using 2SampZTest.

Example 10.1

The mean lasting time of two competing floor waxes is to be compared. The amount of time that each brand lasts is normally distributed. Each brand of wax is applied to 20 randomly chosen floors for testing tested for how the finish lasts under normal wear; the data are recorded in the table below:

Wax	Sample Mean	Population SD
Brand A	3 months	0.33 months
Brand B	2.7 months	0.36 months

Using a 5% significance level, does this data suggest that Brand A wax lasts longer, on average, than Brand B?

Solution 10.1:

This is a test of two means, using independent samples. The random variable we will be working with is $\overline{X}_1 - \overline{X}_2$ = difference in the mean number of months the competing floor waxes last; and since the population standard deviations known, we can use **2SampZTest**.

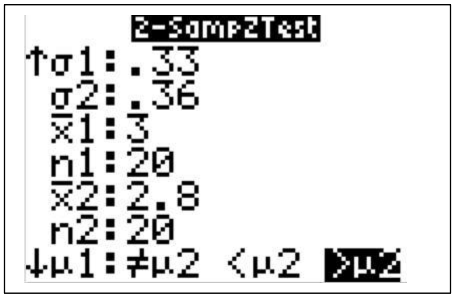
The claim is that Brand A wax **lasts longer** Brand B, on average. I.e. the claim is $\mu_A > \mu_B$. So this is a right tail test, and the hypotheses can be written as either:

H_0: $\mu_A \leq \mu_B$ or H_0: $\mu_A - \mu_B \leq 0$

Ha: $\mu_A > \mu_B$ $\quad$ Ha: $\mu_A - \mu_B > 0$

However, the first is preferred since it more naturally shows the claim and also matches the inputs for the calculator. Go to 2SampZTest, and enter the following:

σ1: 0.33
σ1: 0.36
$\bar{x}_1$: 3
n1: 20
$\bar{x}_2$: 2.8
n2: 20
μ1: > **μ2**



```
2-SampZTest
↑σ1:.33
 σ2:.36
 x̄1:3
 n1:20
 x̄2:2.8
 n2:20
↓μ1:≠μ2 <μ2 >μ2
```

Put the cursor on Calculate and press enter to get $z = 1.83$ and *p-value* = 0.0335

Since *p-value* < .05, we **reject *H*₀**. At the 5% significance level, there is sufficient evidence to conclude that Brand A wax lasts longer, on average, than Brand B wax.

Try It Σ

10.1 The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. The table below shows the result. Using a 5% level of significance, do the data indicate that Engine 2 has higher RPM than Engine 1?

Engine	Sample Mean	Population SD
1	1500	50
2	1600	60

Example 10.2

A field researcher is gathering data on the trunk diameters of mature pine and spruce trees in a certain area. The following are the results of his random sampling.

	Pine Trees	**Spruce Trees**
Sample mean (cm)	35	30
Sample size (n)	40	80
Population Variance	160	160

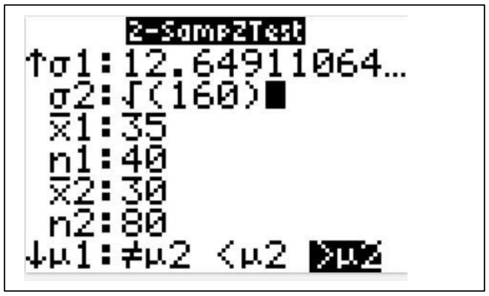
Using a .10 significance level, can he conclude that the average trunk diameter of a pine tree is greater than the average diameter of a spruce tree?

Solution 10.2: This is a test of two means, using independent samples. The random variable we will be working with is $\overline{X}_1 - \overline{X}_2$, where $\overline{X}_1$ represents a random sample mean diameter of pines, and $\overline{X}_2$ is the random sample mean diameter of spruces. Since the population standard deviations known, we can use 2SampZTest.

Let μ_1 be the population mean of diameters of mature pine trees, and μ_2 the population mean of diameters of mature spruce trees. The claim is that pines have larger mean diameter, or $\mu_1 > \mu_2$. So this is a right tail test, and the hypotheses are:

H_0: $\mu_1 \leq \mu_2$
Ha: $\mu_1 > \mu_2$

CALCULATOR: 2SampZTest, and enter the following:



```
     2-SampZTest
↑σ1:12.64911064…
 σ2:√(160)■
 x̄1:35
 n1:40
 x̄2:30
 n2:80
↓μ1:≠μ2 <μ2 >μ2
```

NOTE: The calculator asks for the *standard deviation*, which is the square root of the variance.

Put the cursor on Calculate and press enter to get $z = 2.94$ and *p-value* = 0.0206.

Since *p-value* < .10, we will **reject H_0**. At the .10 significance level, there is sufficient evidence to conclude that the mean diameter for pine trees is larger than the mean diameter for spruce trees.

Finally, we note that a two-sample Z-test can also be done using critical values. That is, we can make the decision whether or not to reject Ho by comparing the test statistic to an appropriate critical value. For example, in Example 10.2 above, we have a right tailed Z-test with significance level α = .10. The critical value for the test will be *z*-value that has an area of .10 to the right of it; so the area to the left of this value is .90. This is easily found using the calculator:

$$z_{.10} = \text{invNorm}(.90) = 1.282.$$

Thus, we would reject Ho if and only if the test statistic $z > 1.282$. Since our test statistic is $z = 2.94$, our decision again is to reject Ho.

For convenience we again show the rejection regions for the three most commonly used significance levels:

	Rejection Regions for Z-Test		
Significance Level	**Left-Tailed**	**Right Tailed**	**Two-Tailed**
α = .10	$z < -1.28$	$z > 1.28$	$z < -1.645$ or $z > 1.645$
α = .05	$z < -1.645$	$z > 1.645$	$z < -1.96$ or $z > 1.96$
α = .01	$z < -2.33$	$z > 2.33$	$z < -2.576$ or $z > 2.576$

Summary of Two-Sample Z-Test

- Used for testing a claim about two unknown means μ_1, μ_2 with data from independent samples.
- Use only when the population standard deviations σ_1, σ_2 are known.
- Random variable and distribution: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
- Test statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- Calculator function: **2-SampZTest**

10.2 | Two Population Means when Standard Deviations are Unknown

When we want to compare two unknown population means, it is often the case that the population standard deviations are also unknown. The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Two Sample T-Test, or Aspin-Welch *t*-test. (The formula for the degrees of freedom developed by Aspin and Welch)

Again we are using independent samples, and working with the random variable $\bar{X}_1 - \bar{X}_2$. And again the mean of this sampling distribution is $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$. But when we do not know the population standard deviations, we must estimate them using the two sample standard deviations (s_1 and s_2) from our independent samples. This in turn provides an estimate for the standard deviation of the sampling distribution:

$$\sigma_{\bar{x}_1-\bar{x}_2} \approx s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

The expression is called the **standard error of the difference of sample means**. The test statistic for the test will be $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{standard error}}$, where $\mu_1 - \mu_2$ is the difference in means hypothesized in H_0. As with the Two Sample Z-Test, we almost start with a null hypothesis that includes an equal sign, so for our calculations we will be assuming that $\mu_1 - \mu_2 = 0$, giving the simplified test statistic:

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

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

$$df = \frac{\left(\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}\right)^2}{\dfrac{{s_1}^4}{{n_1}^2(n_1 - 1)} + \dfrac{{s_2}^4}{{n_2}^2(n_2 - 1)}}$$

☞ It is not necessary to compute this by hand. The TI-83 and TI-84 calculator will compute it for us automatically (thank goodness!)

The calculator instructions for a Two-Sample T-Test are essentially the same as for a Two-Sample Z-test, except that there is one addition input. When using **2SampTTest**, we will be asked whether we want to "**Pool the Data**". If we know that the population variances are equal, then we can combine the two samples to get a better estimate of the common population variance. This is called "pooling the data"; **we *only* pool the data if we know the population variances are equal**.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the test statistic and *p*-value for a Two Sample T-Test:

Go to the STAT, menu and then to TESTS. Select option 4, which is **2-SampTTest**.

Enter the sample means, sample SD's and sample sizes; select the appropriate alternative hypothesis. For 2-SampTTest, we are also asked if we want to Pool the data, and use the rule:

- If we know that the population variances are equal, we say YES (pool the data)
- If we do not know, or if the population variances are not equal, we say NO.

Arrow down to Calculate and press ENTER. The test statistic and *p*-value will appear on the output screen. If we select the Draw option instead of Calculate, we can view a graph showing the region corresponding to the *p*-value.

NOTE: Any time we are unsure of whether to pool the data, we should select **NO**.

Example 10.3

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

	Sample Size	Sample Mean	Sample SD
Girls	9	2	0.873
Boys	16	3.2	1.0

Using a there a 5% level of significance, is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day?

Solution 10.3

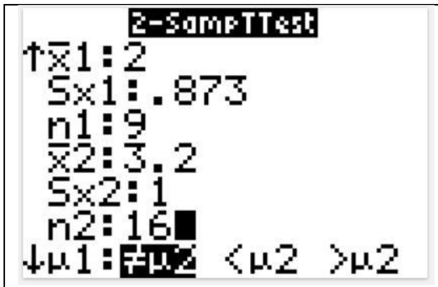
The population standard deviations are not known**.** Let *g* be the subscript for girls and *b* be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test for two population means using independent samples, and the population standard deviations are not known – so we use **2-SampTTest**.

This will use the random variable $\overline{X}_g - \overline{X}_b$, which is the difference in the sample means between boys and girls. We are testing whether there is a *difference*, so the claim is that $\mu_g \neq \mu_b$ and the hypotheses are:

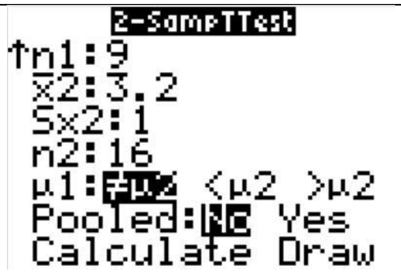
$$H_0\text{: } \mu_g = \mu_b$$
$$Ha\text{: } \mu_g \neq \mu_b$$

Note that this is a two-tailed test.

Go to 2SampTTest, and enter the following:



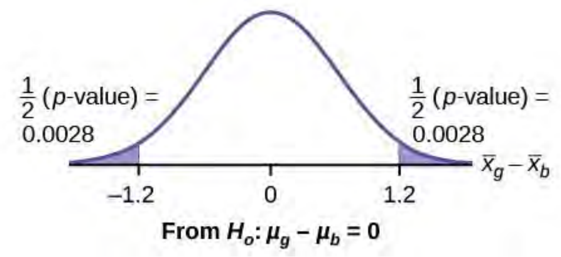
```
2-SampTTest
↑x̄1:2
 Sx1:.873
 n1:9
 x̄2:3.2
 Sx2:1
 n2:16
↓μ1:≠μ2 <μ2 >μ2
```



```
2-SampTTest
↑n1:9
 x̄2:3.2
 Sx2:1
 n2:16
 μ1:≠μ2 <μ2 >μ2
 Pooled:No Yes
 Calculate Draw
```

Put the cursor on Calculate and press enter to get $t = -3.14$ and *p-value* = 0.0056. Note also that the degrees of freedom is shown on the output screen: $df = 18.7264$.

Since *p-value* < .05, we will **reject H_0**. At the 5% significance level, there is sufficient evidence to conclude that there is a difference in the mean time spent playing sports between boys and girls. Finally, if we put the cursor on Draw instead of Calculate, we would see a graph similar to the picture below (the test statistic and *p*-value would still be displayed)



Try It Σ

10.3 Data from two independent samples are shown in the table below. Both have normal distributions. The means for the two populations are thought to be the same. Does this sample data provide evidence of a difference in the population means? Test at the 5% level of significance.

	Sample Size	Sample Mean	Sample SD
Population A	25	5	1
Population B	16	4.7	1.2

Example 10.4

A study is done by a community group in two neighboring colleges to determine which school graduates students with more math classes. College A samples 11 graduates; the average for this sample is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that, on average, students who graduate from College A have taken more math classes than students from College B. Assume that both populations have a normal distribution, and we are testing a 1% significance level. Answer the following questions.

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the hypotheses for the test?
f. Is this test right, left, or two-tailed?
g. What is the *p*-value?
h. Should H_0 be rejected or not?
i. What conclusion can we draw from the test?

Solution 10.4

a. Test for two means b. These are unknown c. Student's *t*-distribution d. $\overline{X}_A - \overline{X}_B$

e. $H_0 : \mu_A \leq \mu_B$ vs. $H_a : \mu_A > \mu_B$ f. Right-tailed g. From TI-84, $p = 0.1928$

h. Do not reject H_0 since *p-value* > .01

i. There is not sufficient evidence at the 1% level of significance 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.

Try It Σ

10.4 A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test of the claim using a 5% significance level. What is your conclusion?

Example 10.5

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

Online class:

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

Face-to-face Class:

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

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? We will test this claim with a 5% significance level.

Solution 10.5

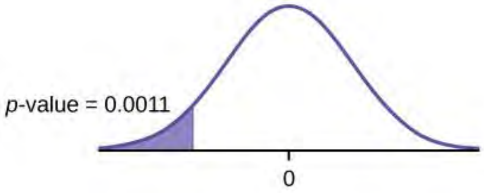
This is a test for two unknown population means, using independent samples and unknown population SD's, so we use a **2-SampTTest**. We let μ_1 be the population mean for online classes, and μ_2 be the population mean for face-to-face classes. The claim being tested is that scores for online classes are *lower*, on average, than scores for face-to-face classes. I.e. the claim is that $\mu_1 < \mu_2$, and so the hypotheses are:

H_0: $\mu_1 \geq \mu2$ and Ha: $\mu_1 < \mu_2$

First we go the STAT menu and then to EDIT; enter the sample data into two lists (e.g. L1, L2). Arrow over to TESTS and select option 4: 2SampTTest. Choose the Data option, and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Select the "<" alternative; do not pool the data (i.e. select Pooled: No) Arrow down to Calculate and press ENTER to get:

$t = -3.23$ and *p-value* = 0.0011

NOTE: If we use the Draw option instead of Calculate, we get the graph below; the test statistic and p-value are also displayed.



Since *p-value* < .05, we **reject H_0**. At the 5% level of significance, there is sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face- to-face class. So it appears that the professor was correct.

Cohen's Standards for Small, Medium, and Large Effect Sizes (Optional)

Cohen's *d* statistic is a measure of effect size based on the differences between two means. Named for American statistician Jacob Cohen, *d* measures the relative strength of the differences between the means of two populations based on sample data. This statistic is the difference of the observed sample means, divided by $s_{\text{pooled}} = \sqrt{\dfrac{(n_1-1)s_1{}^2 + (n_2-1)s_2{}^2}{n_1+n_2-2}}$. This is the sample standard deviation obtained by "pooling" the data – that is, by combining the two samples into a single, large sample. Note that we can get this value from the calculator; if we use 2-SampTTest and select YES for Pooling, the last entry in the output screen is Sxp, which is the pooled standard error.

Now, the Cohen's *d*-statistic is defined as $d = \dfrac{(\bar{x}_1 - \bar{x}_2)}{s_{\text{pooled}}}$.

The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes according to the following table:

Size of Effect	***D***
Small	0.2
Medium	0.5
Large	0.8

Cohen's Standard Effect Sizes

Example 10.6

Calculate Cohen's *d* for Example 10.4. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution 10.6

$\bar{x}_1 = 4,\ s_1 = 1.5,\ n_1 = 11;\quad \bar{x}_2 = 3.5,\ s_2 = 1,\ n_2 = 9$

Go to 2-SampTTest, and enter the data; select YES for pooling, then scroll to Calculate and press Enter to get s_{pooled} = Sxp = 1.302. Using the formula, we get

$$d = (\bar{x}_1 - \bar{x}_2)/s_{\text{pooled}} = (4 - 3.5)/1.302 = 0.384.$$

We would classify the effect as small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. Since the effect size for the data is small, this indicates that there is not a significant difference between the means for the two colleges.

Example 10.7

Calculate Cohen's *d* for Example 10.3. Is the size of the effect small, medium or large? Explain what the size of the effect means for this problem.

Solution 10.7

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

Try It Σ

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

94.2	75.2	69.6	52	48	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.6	26

Northeast

126	70.6	65.2	51.4	45.5	37	33	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39	34.1	31	25.3	23.4	21.5

West

Is there a difference in the mean weighted alpha for banks in the northeast and in the west? Test at a 5% significance level. Explain what the size of the effect means for this problem.

Summary of Two-Sample T-Test

- Used for testing a claim about two unknown means μ_1, μ_2 with data from independent samples.

- Use when sampling from two normal distributions and the population standard deviations σ_1, σ_2 are *not* known.

- Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ $\qquad df = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{s_1^4}{n_1^2(n_1-1)} + \dfrac{s_2^4}{n_2^2(n_2-1)}}$

- Calculator function: **2-SampTTest**

- Select the "Pooled" option ***only*** when the population variances are equal.
  If in doubt, do not pool the data.

10.3 | Comparing Two Population Proportions

In this section we develop a test for testing claims involving two unknown population proportions, p_1 and p_2. As we did for comparing unknown means, we will look at the parameter $p_1 - p_2$, the difference of the two population proportions. The obvious point estimate for this parameter is $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are sample proportions chosen from the respective populations. When conducting tests comparing two population proportions ($p_1 - p_2 =$ hypothesized value), we will assume the following:

1. We are using two independent, random samples.

2. The number of successes is at least 10, and the number of failures is at least 10, for each of the samples. That is, we should have $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$ as well as $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$.

These assumptions imply that the sampling distribution for $\hat{p}_1 - \hat{p}_2$ will be approximately normal. The mean of the sampling distribution will be $p_1 - p_2$, and the variance will be the sum of the variances of the sampling distributions for $\hat{p}_1$ and $\hat{p}_2$. That is, we have:

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \quad \text{and} \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\,.$$

Generally, the null hypothesis will state that the two population proportions are the same. That is, H_0: $p_1 = p_2$; and since the population proportions are equal, the variances of the individual sampling variations will also be equal. So we can get a better approximation to this common variance by using a pooled proportion, p_c. This is the total number of successes, divided by the total sample size: $p_c = \dfrac{x_1 + x_2}{n_1 + n_2}$. Inserting p_c in place of p_1 and p_2 in the formula above and simplifying, we get:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Now the test statistic for testing a claim about two proportions will be: $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}}$.

Because the null hypothesis assumes that p_1 and p_2 are equal, we have $p_1 - p_2 = 0$; substituting the expression for $\sigma_{\hat{p}_1 - \hat{p}_2}$ above, we get the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\,.$$

This may look complicated, but it is programmed into the TI-83 and TI-84 calculators. As with other tests, the calculator will give us both the test statistic and the p-value.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the test statistic and p-value for a Two Proportion Z-Test:

Go to the STAT, menu and then to TESTS. Select option 6, which is **2-PropZTest**.

Enter the values for x_1, n_1, x_2, and n_2. **Note** that both x_1 and x_2 must be whole numbers; so if you are given sample proportions, use $x_1 = n_1\hat{p}_1$ and $x_2 = n_2\hat{p}_2$, and round to the nearest whole number. Select the appropriate alternative hypothesis.

Arrow down to Calculate and press ENTER. The test statistic and p-value will appear on the output screen. Note that the pooled proportion p_c also shows on the output screen as $\hat{p}$ (without a subscript). As with other tests, we can select the Draw option instead of Calculate to view a graph showing the region corresponding to the p-value.

NOTE: the following summary is focusing on $H_0 : p_1 - p_2 = 0$ which also includes ≥ and ≤.

Test value: $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Assumptions: 1. The samples are independently chosen random samples

2. Both sample sizes are large:

$n_1\hat{p}_1 \geq 10;\ n_1(1-\hat{p}_1) \geq 10; n_2\hat{p}_2 \geq 10; n_2(1-\hat{p}_2) \geq 10.$

Example 10.8

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

Solution 10.8

This is a test of two proportions. Let p_A and p_B be population proportions of individual with hives after 30 minutes for medications A and B respectively. The problem asks if there is a *difference* in proportions, so the claim being tested is that $p_A \neq p_B$ and so the hypotheses are:

H_0: $p_A = p_B$ and Ha: $p_A \neq p_B$

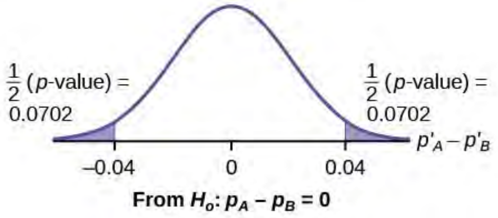
Note that this is a two-tailed test, and we have data from independent random samples. From the given information, we can see that the number of successes and failures in each sample is at least 10, so the sampling distribution for $\hat{p}_A - \hat{p}_B$ will be approximately normal. Thus, we will use 2-PropZTest. In the calculator, go to the STAT menu, then to TESTS and select **2-PropZTest**.

Enter the following:

x1: 20
n1: 200
x2: 12
n2: 200
p : $\neq$ $\boldsymbol{p_0}$

Note: both x_1 and x_2 must be whole numbers; so if you are given sample proportions, use $x_1 = n_1 \hat{p}_1$ and $x_2 = n_2 \hat{p}_2$, and round to the nearest whole number.

Scroll down to Calculate and press Enter to get: $z = 1.47$ and p-value = 0.1404.
Using the Draw feature, we get a graph that looks like:



Since *p-value* = 0.1404 > 0.01, we **do not reject H₀**.
Thus, at a 1% level of significance, this sample data does not provide sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication *A* and medication *B*.

Try It Σ

10.8 Two types of valves are being tested to determine if there is a difference in pressure tolerances. In a random sample of 100 Brand *A* valves, 14 cracked under 4,500 psi. In a random sample of 100 of Brand *B* valves, 19 cracked under 4,500 psi. Using a 5% level of significance, test the claim that there is a difference in the proportion of valves that crack at 4,500 psi.

Example 10.9

A research study was conducted about gender differences in "sexting." The researcher believed that the proportion of girls involved in "sexting" is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in the table:

	Males	**Females**
Sent "sexts"	183	156
Total surveyed	2231	2169

Is the proportion of boys sending sexts more than the proportion of girls sending sexts? Test at a 1% level of significance.

Solution 10.9

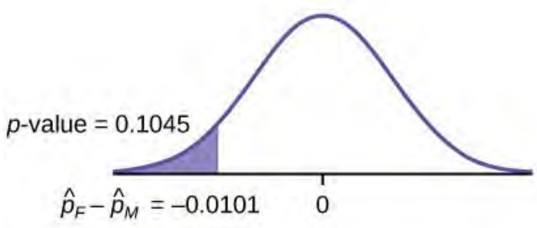
This is a test of two population proportions. Let M and F be the subscripts for males and females. Let p_M and p_F be the proportions for males and females respectively. The claim is that the proportion of boys sending "sexts" is *more* than the proportion of girls, so the claim is $p_F < p_M$. Thus we have the hypotheses:

$$H_0: p_F \geq p_M \quad vs. \quad H_a: p_F < p_M$$

Note that the test is right-tailed.
In the TI-84, go to STAT, then TESTS, and select 2-PropZTest. Enter the following:

x1: 156 , n1: 2169 , x2: 183, n2: 2231, p : < $\boldsymbol{p_0}$

Scroll down to Calculate and press Enter to get: $z = -1.26$ and p-value = 0.1045.
Using the Draw feature, we get the graph:



Since $p = .1045 > .10$, we **do not reject H_0.**
At the 1% level of significance, there is not sufficient evidence to conclude that the proportion of girls sending "sexts" is less than the proportion of boys sending "sexts."

Example 10.10

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. At a 5% level of significance, is there evidence that the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

Solution 10.10

Let p_W and p_A are be the proportions of people who own iPhones in the white and African-American populations, respectively. The claim we are investigating is that the proportion of whites who own iPhones is greater than the proportion for African-Americans: that is, the claim is that $p_W > p_A$. So we have the hypotheses:

$$H_0: p_W \leq p_A, \qquad H_a: p_W > p_A$$

Note that the test is right-tailed. The conditions for using a Two-Proportion Z-test are easily met, so we go to **1-PropZTest** and enter the following:

x1: 134 (this is 10% of 1343, rounded to a whole number)
n1: 1343
x2: 12 (this is 5% of 232, rounded to a whole number)
n2: 232
p : $> p_0$

Scroll down to Calculate and press Enter to get: $z = 2.33$ and p-value = 0.0099.

Since *p-value* < .05, we **reject H₀**. At the 5% level of significance, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

Try It Σ

10.10 A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Using a 5% significance level, is there a difference in the proportions?

Summary of Two-Proportion Z-Test

- Used for testing a claim about two unknown proportions p_1, p_2 with data from independent samples.

- Assume that $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$ and that $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$. That is, there are at least 10 successes and 10 failures in each sample.

- Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$, where $p_c = \dfrac{x_1 + x_2}{n_1 + n_2}$

- Calculator function: **2-PropZTest**

10.4 | Matched Pairs Test

Two samples are *dependent* when the subjects are paired or matched in some way. For example, to study whether or not a medication helps lower cholesterol, researchers might take a random sample of patients and measure their cholesterol. Then after a specified time taking the medication, the patients would have their cholesterol measured again – this would produce pairs of "before" and "after" data values, with one pair for each patient. We could then analyze the actual *differences*, D = before – after, of the pairs using an ordinary t-test. Thus the data is the gain or loss in cholesterol readings. Such a test is called a *matched-pairs* test. One of the key advantages of this experiment design is that it controls for a lot of individual attributes of the patients, such as diet, exercise, overall health, etc.

When using a hypothesis test for matched or paired samples, the following conditions must be met:

1. Simple random sampling is used.
2. Two measurements are drawn from the same pair of individuals.
3. Differences are calculated for each matched pair.
4. The differences comprise the sample that is used for the hypothesis test.
5. Either the population of differences are normally distributed, or else the number of differences in the sample is sufficiently large to insure that the sampling distribution is approximately normal.

Notation: We will use d to represent individual differences, so $\bar{d}$ is the sample mean of the differences, and μ_d is the population mean of *all* differences. From the Central Limit Theorem, we get the mean and standard deviation of the sampling distribution for $\bar{d}$: $\mu_{\bar{d}} = \mu_d$ and $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$.

From this, we get a formula for the test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$.

Of course, we will rarely (if ever) use this formula, since we can get the test statistic, as well as the corresponding p-value from the calculator.

Example 10.11

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table:

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

A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. Assume that the population of differences has a normal distribution. Using a 5% significance level, is there evidence that the sensory measurements are lower, on average, after hypnotism?

Solution 10.11

The corresponding "before" and "after" values form matched pairs; we will let

$$d = \text{"before"} - \text{"after"}$$

If the sensory readings are lower after hypnotism, then “before” readings should, on average, be larger than “after” readings and so we would expect the mean of all *d* values to be positive. Thus the claim we want to test is $\mu_d > 0$. So the hypotheses are:

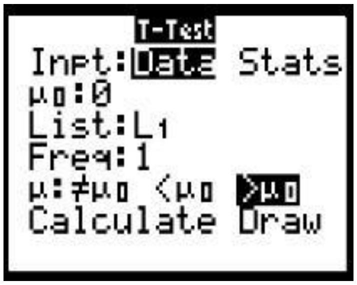
$$H_0\text{: } \mu_d \leq 0, \quad H_a\text{: } \mu_d > 0$$

Now we calculate the differences; we can easily do this by adding another row to our table:

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
Difference:	-0.2	4.1	1.6	1.8	3.2	2.0	2.9	9.6

Go to the STAT menu and then to EDIT. Enter the differences into L1.
Next, we go to STAT again, then to the TESTS menu and select T-test. Select the DATA option. Set μ_0: 0, enter the list name (type 2nd 1 to get L1), and select the “greater than” alternative. Scroll down to Calculate and hit Enter to get:

$t =$ 3.036 and *p-value* = 0.0095



T-Test
Inpt:Data Stats
μ_0:0
List:L1
Freq:1
μ:$\neq\mu_0$ $<\mu_0$ $>\mu_0$
Calculate Draw

Reject Ho because 0.0095 < 0.05

At the 5% level of significance, there is sufficient evidence to conclude that the sensory measurements are lower, on average, after hypnotism. Hypnotism appears to be effective in reducing pain.

NOTE

For the TI-84 calculator, you can either calculate the differences (before – after) ahead of time and put the differences directly into a list, or you can put the **before** data into one list L1 and the **after** data into a second list, L2. Then go to L3 and arrow up to the name; that is, put the cursor on L3 at the top of the screen.

Type L1 – L2 and press Enter; the calculator will do the subtractions and store the differences in List 3.

Try It Σ

10.11 A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Subject:	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Possible claim scenarios for Matched Pair Tests:

1.) There is a decrease from before to after, which means after is "lower" than before: this scenario will result in the average difference being positive, $\mu_d > 0$.

2.) There is an increase from before to after, which means after is "higher" than before: this scenario will result in the average difference being negative, $\mu_d < 0$.

3.) There is a difference from before to after, which means after is not the same as before: this scenario will result in the average difference being not equal to 0, $\mu_d \neq 0$.

4.) There is no difference from before to after, which means after is the same as before: this scenario will result in the average difference being equal to 0, $\mu_d = 0$.

Example 10.12

A pharmaceutical company wishes to test a new drug with the expectation of lowering cholesterol levels. Ten subjects are randomly selected and pretested. The subjects were placed on the drug for a period of 6 months, after which their cholesterol levels were tested again.

The test results, before and after, are listed below. (All units are milligrams per deciliter.)

Subject	1	2	3	4	5	6	7	8	9	10
Before	195	225	202	195	175	250	235	268	190	240
After	180	220	210	175	170	243	205	250	183	225

Use this data, along with a 1% significance level, to test the claim that the drug is effective in lowering cholesterol.

Solution 10.12

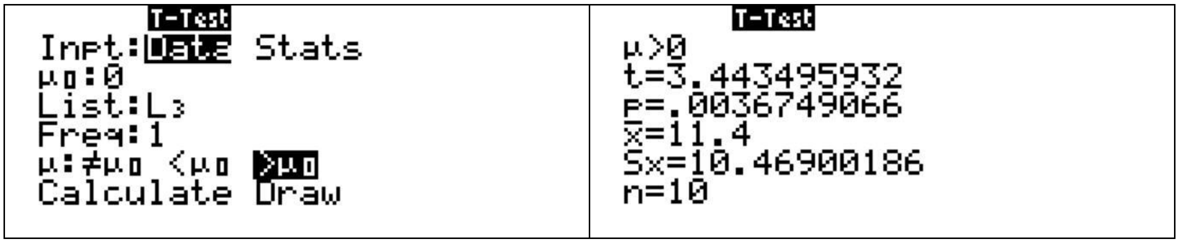
Here we are comparing two population means, using a matched pairs design. Thus, we will use a t-test on the differences: 15, 5, -8, 20, 5, 7, 30, 18, 7, 15

Let *d* = cholesterol level before taking drug – cholesterol level after taking drug.

If the drug lowers cholesterol, then the cholesterol level should be higher *before* the drug so if the claim is true, we expect mean of the differences to be positive, giving us the hypotheses:

Ho: $\mu_d \leq 0$
Ha: $\mu_d > 0$ (claim)

Enter the differences into a list, and go to STAT >> TESTS >> TTest. Enter 0 for μ_0, specify the list name, and select the $> \mu_0$ alternative.



```
T-Test
Inpt:Data Stats
μ0:0
List:L3
Freq:1
μ:≠μ0 <μ0 >μ0
Calculate Draw
```

```
T-Test
μ>0
t=3.443495932
p=.0036749066
x̄=11.4
Sx=10.46900186
n=10
```

This gives us a test statistic of $t = 3.44$, and a p-value = .0037. Since *p-value* < 0.01, we **reject Ho**. There is sufficient evidence to conclude that the mean of the differences is negative. Thus, there is sufficient evidence to conclude that the drug helps lower cholesterol.

Try It Σ

10.12 A new prep class was designed to improve SAT test scores. Nine students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. Are the scores, on average, higher after the class? Test at a 5% level. The data recorded in the table below:

Student	A	B	C	D	E	F	G	H	I
First Score	480	510	530	540	550	560	600	620	660
Second Score	490	520	550	530	580	580	610	640	690

Example 10.13

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded in the table below; the distances shown are in feet.

	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 there is a significant difference in mean distances between the children's dominant and weaker hands. Use a significance level of $\alpha = .05$.

Solution 10.13

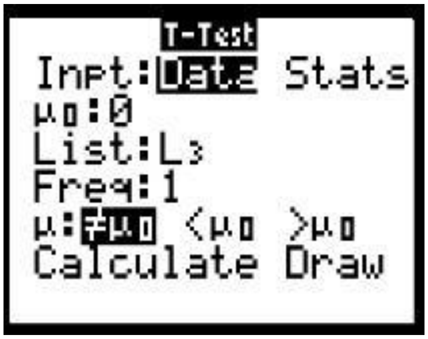
First, we find the differences for each pair. Here we let

d = distance for dominant hand – distance for weaker hand.

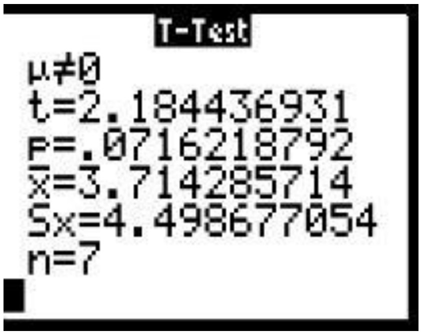
Since we asked whether there is a *difference* in the means, this will be a two-tailed test with hypotheses:

$$H_0: \mu_d = 0, \quad Ha: \mu_d \neq 0$$

We will assume that the differences have a normal distribution, and will use a T-test. Enter the differences into a list, and go to STAT >> TESTS >> TTest. Enter 0 for μ_0, specify the list name, and select the $\neq \mu_0$ alternative.



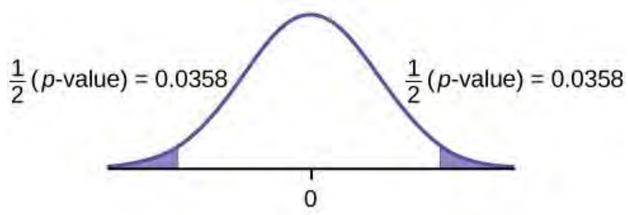
```
T-Test
Inpt:Data Stats
μ0:0
List:L3
Freq:1
μ:≠μ0 <μ0 >μ0
Calculate Draw
```



```
T-Test
μ≠0
t=2.184436931
p=.0716218792
x̄=3.714285714
Sx=4.498677054
n=7
```

This gives us a test statistic of $t = 2.18$, and a p-value of $p = .0716$. Since $p > 0.05$, we do not reject Ho. Thus, there is not enough evidence to conclude that there is a significant difference in the means.

Note that if we had selected the DRAW option, we would see the following graph:



The graph shows that had this been a right-tailed test, our decision would have been to reject Ho. However, our hypotheses are based on the question that has been asked, not on the sample data. And it would not be appropriate to change the hypotheses once we saw the data in order to get a different result.

Notice the other statistics that are given on the calculator screen ($\bar{x}$ and S_x). These values are the sample average difference and sample standard deviation of the differences. The symbols that should be used for the values are as follows:

$\bar{d} = 3.71$ and $S_d = 4.5$.

10.5 | Confidence Intervals from Two Samples

In this chapter, we have learned how to investigate claims involving two parameters; but sometimes we actually need an estimate of the difference of two parameters. For example, if we conducted a hypothesis test and conclude that μ₁ > μ₂, a natural follow-up question might be: *by how much*? Using a confidence interval, we can actually give bounds for how large the difference between these parameters really is. In this section we will extend the ideas of Chapter 8 to develop confidence interval estimates for the difference of two means and the difference of two proportions. Like the intervals developed in Chapter 8, these will have the basic form:

(point estimate – margin of error, point estimate + margin of error)

And we now have all of the ingredients needed. For example, if we want to find a confidence interval for $\mu_1 - \mu_2$ using data from two independent samples, we know that the appropriate point estimate is the difference of sample means, $\bar{x}_1 - \bar{x}_2$; and this random variable will either follow a normal distribution or a *t*-distribution. When the population standard deviations are known, we can use a z-distribution, and the margin of error will be $E = z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}$. Thus, the confidence interval will have the form:

$$(\bar{x}_1-\bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}} < \mu_1-\mu_2 < (\bar{x}_1-\bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}} \; .$$

When the population standard deviations are not known, then we would estimate σ_1 and σ_2 by s_1 and s_2 and use a critical *t* values in place of the *z* critical values:

$$(\bar{x}_1-\bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}} < \mu_1-\mu_2 < (\bar{x}_1-\bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}} \; .$$

While these formulas look complicated, we should remember that they are really the same basic format as the intervals for a single mean μ that we developed in Chapter 8 (but the point estimates and standard errors are different). Moreover, we never need to actually *use* these formulas, because they are built into the TI-83 and TI-84 calculators.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate a confidence interval for the difference of two means, μ₁ – μ₂:

Go to the STAT, menu and then to TESTS.
If the population SD's are known, then Select option 9, which is **2-SampZInt**.
If the population SD's are not known, then Select option 10, which is **2-SampTInt**.

Enter the SD's, the sample means and the sample sizes just as we did for two-sample tests. Specify the confidence level. Arrow down to Calculate and press ENTER. The confidence interval will appear on the screen.

To illustrate this, we revisit the question from Example 10.2; in that example, we used a hypothesis test to compare the mean trunk diameters for pine spruce trees. In that test, we concluded that the average trunk diameter for pine tree is in fact greater than the average diameter of spruce trees.

Example 10.14

A field researcher is gathering data on the trunk diameters of mature pine and spruce trees in a certain area. The following are the results of his random sampling.

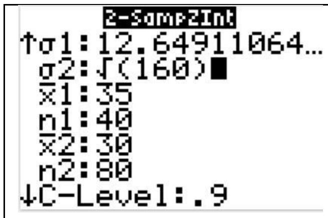
	Pine Trees	**Spruce Trees**
Sample mean (cm)	35	30
Sample size	40	80
Population Variance	160	160

Use this sample data to find a 90% confidence interval for the difference in mean trunk diameter for pine and spruce trees. Interpret this interval.

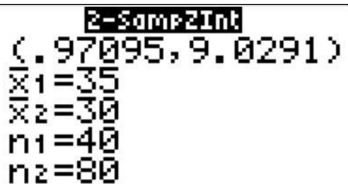
Solution 10.14

We let μ_1 be the mean diameter for pine trees, and μ_2 be the mean diameter for spruce trees. We will calculate a confidence interval for $\mu_1 - \mu_2$; the population variances are known, so we will use **2-SampZInt**.

Go to 2SampZInt, and enter the following:



```
      2-SampZInt
↑σ1:12.64911064…
 σ2:√(160)■
 x̄1:35
 n1:40
 x̄2:30
 n2:80
↓C-Level:.9
```



```
    2-SampZInt
(.97095,9.0291)
x̄1=35
x̄2=30
n1=40
n2=80
```

Hit CALCULATE to get the interval. So we are 90% confident that $.971 < \mu_1 - \mu_2 < 9.03$.

That is, the mean for pine trees is *at least* .971 inches more than the mean for spruce trees. And, the mean for pine trees is *at most* 9.03 inches more than the mean for spruce trees.

Recall also Example 10.5; in that example, a professor at a large college wanted to compare the mean final exam scores for two populations of students – online students and students in traditional, face to face courses. We conducted a test and found evidence that the mean score for online classes was lower than the mean for traditional courses. Here we will calculate a confidence interval for the difference of means.

Example 10.15

A professor at a large community college wanted to compare the mean final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. The randomly selected 30 final exam scores from each group are listed below:

Online class:

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

Face-to-face Class:

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

Calculate a 95% confidence interval for the difference in mean scores; interpret this interval.

Solution 10.15

We let μ_1 be the population mean for online classes, and μ_2 be the population mean for face-to-face classes. We want a confidence interval for $\mu_1 - \mu_2$; the population SD's are not known, so will use **2-SampT-Int**. Moreover, instead of sample statistics, we are given raw data, so we will use the DATA option.

- First go the STAT menu and then to EDIT; enter the sample data into two lists (e.g. L1, L2).
- Go to STAT again, then to TESTS; select option 10: 2SampTInt. Choose the DATA option.
- Arrow down and enter L1 for the first list and L2 for the second list.
- Set **C-Level: .95**
- Set **Pooled: No** (since we have no reason to think the population variances are equal)
- Arrow down to Calculate and press Enter to get the interval: **(-19.67, -4.59).**

Thus, we are 95% confident that $-19.67 < \mu_1 - \mu_2 < -4.59$.

The mean for online courses is *at least* 4.59 points lower than the mean for traditional courses. And, the mean for online courses is *at most* 19.67 points lower for traditional courses.

We can also find a confidence interval for $p_1 - p_2$, the difference of the two population proportions. The point estimate for this parameter is $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are sample proportions chosen from the respective populations. And we know that if the sample sizes are sufficiently large (see Section 10.3), then the sampling distribution for $\hat{p}_1 - \hat{p}_2$ will be approximately normal, so the margin of error will be $E = z_{\alpha/2}\sigma_{\hat{p}_1 - \hat{p}_2} = z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$. Thus we get the following formula for the confidence interval:

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

Again, we would rarely (if ever) use this formula, as it is built into the TI-84 calculator:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate a confidence interval for $p_1 - p_2$, the difference of the two population proportions:

Go to the STAT, menu and then to TESTS. Select option B, which is **2-PropZInt**.

Enter the values for x_1, n_1, x_2, and n_2. Note that both x_1 and x_2 must be whole numbers; so if you are given sample proportions, use $x_1 = n_1\hat{p}_1$ and $x_2 = n_2\hat{p}_2$, and round to the nearest whole number. Specify the Confidence level, as a decimal. Arrow down to Calculate and press ENTER.

The confidence interval will appear on the output screen.

To illustrate this, we revisit the situation from Example 10.10:

Example 10.16

Researchers for a cell phone company conducted a study of smartphone use among adults. They wish to compare the proportion of smartphone users among two different populations: adult whites (non-Hispanic) and African American adults. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Use this data to construct a 95% confidence interval for the difference in population proportions.

Solution 10.16

Let p_W and p_A are be the proportions of people who own iPhones in the white and African-American populations, respectively. Then we want a confidence interval for $p_W - p_A$.

The conditions for using a Two-Proportion Z-Interval are easily met, so we go STAT, then to TESTS, and scroll down to to **2-PropZInt.** Enter the following:

```
    x1:  134       (this is 10% of 1343, rounded to a whole number)
    n1:  1343
    x2:  12          (this is 5% of 232, rounded to a whole number)
    n2:  232
 C-Level: .95
```

Arrow down to Calculate and press Enter to get the interval: **(0.0156, 0.0875).**

Thus, we are 95% confident that $.0156 < p_W - p_A < .0875$.

KEY TERMS and FORMULA REVIEW

Cohen's effect size: This is a measure of the relative strength of the difference between two means based on sample data. Cohen's *d*-statistic is defined as $d = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\text{pooled}}}$, where s_{pooled} is the pooled standard error, $s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^{\ 2} + (n_2 - 1)s_2^{\ 2}}{n_1 + n_2 - 2}}$.

Degrees of freedom: The number of objects in a sample that are free to vary. For a one-sample *t*-test, the degrees of freedom was simple: df = *n* – 1. However, for a two-sample t-test, the calculation is more complicated, given by Welch's formula:

$$df = \frac{\left(\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}\right)^2}{\frac{s_1^{\ 4}}{n_1^{\ 2}(n_1 - 1)} + \frac{s_2^{\ 4}}{n_2^{\ 2}(n_2 - 1)}}$$

Pooled proportion: Used in a hypothesis test for two proportions to estimate the common value of p_1 and p_2: $p_c = \frac{x_1 + x_2}{n_1 + n_2}$.

Sampling Distribution for $\hat{p}_1 - \hat{p}_2$: The mean and standard deviation are:

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \quad \text{and} \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} .$$

Provided np_1 and np_2 are both greater than 10, the distribution will be approximately normal.

Sampling Distribution for the random variable $\bar{X}_1 - \bar{X}_2$:

The mean and standard deviation are: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}}$

If we are sampling from normal distributions, or the sample sizes are both large, then the sampling distribution is approximately normal: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}})$

Standard Error: This is an estimate of $\sigma_{\bar{x}_1-\bar{x}_2}$ obtained by estimating the unknown population variances by the respective sample variances; the standard error is: $s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}$.

Two Sample Confidence Intervals:

Z-Interval for estimating difference of two means:

$$(\bar{x}_1-\bar{x}_2)-z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}$$

T-Interval for estimating difference of two means:

$$(\bar{x}_1-\bar{x}_2)-t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}$$

Z-Interval for estimating difference of two proportions:

$$(\hat{p}_1-\hat{p}_2)-z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}<p_1-p_2<(\hat{p}_1-\hat{p}_2)+z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

CHAPTER REVIEW

10.1 Two Population Means with known Standard Deviations

If we are testing a claim comparing two unknown population means,μ_1 and μ_2, where the population standard deviations σ_1 and σ_2 are known, then we use a **Two-Sample Z-Test**. The key features of the test are:

- The data must come from independent samples
- Use only when the population standard deviations σ_1, σ_2 are known.
- Random variable and distribution: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
- Test statistic: $z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- Calculator function: 2-SampZTest

10.2 Two Population Means with unknown Standard Deviations

If we are testing a claim comparing two unknown population means,μ_1 and μ_2, where the population standard deviations σ_1 and σ_2 are *not* known, then we use a **Two-Sample T-Test**; the key features of the test are:

- Assume that the data comes from independent samples.
- Assume that the populations from which we are sampling are approximately normal, and the population standard deviations σ_1, σ_2 are *not* known.
- Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- $df = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$
- Calculator function: 2-SampTTest
- Select the "Pooled" option ***only*** when the population variances are equal. If in doubt, *do not* pool the data.

10.3 Comparing Two Independent Population Proportions

If we are testing a claim comparing two unknown population proportions, p_1 and p_2, we use a **Two-Proportion Z-Test**; the key features of the test are

- Assume the data comes from from independent samples.
- Assume that $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$ and that $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10$.
  That is, there are at least 10 successes and 10 failures in each sample.
- Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, where $p_c = \dfrac{x_1 + x_2}{n_1 + n_2}$
- Calculator function: 2-PropZTest

10.4 Matched or Paired Samples

If we are testing a claim about two unknown population means, using dependent samples consisting of matched pairs of data values, then we use an ordinary t-test on the differences. The key features of the test are:

- Test the differences d by subtracting one measurement from the other measurement in each pair.
- Random Variable: $\bar{d}$ = sample mean of the differences
- Distribution: Student's t-distribution with $n - 1$ degrees of freedom, where n is the number of matched pairs.
- If n (the number of differences) is less than 30, then we must assume that the differences are normally distributed.
- Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$.

Exercises for Chapter 10

The next seven exercises all describe a claim or scenario that will be investigated using a hypothesis test. In each case:

a. State the hypotheses in terms of the appropriate parameter(s)
b. Determine if the samples are independent or dependent
c. Identify the random variable, and its distribution (either normal or t-distribution), along and explain this choice.
d. Identify the test to be used.

1. It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. A test will be conducted to check whether the proportions are in fact equal.

2. The manufacturer of a new windshield treatment claims that their product will repel water more effectively. To test the claim, ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. This data is then used to test the manufacturer's claim.

3. The known standard deviation in salary for all mid-level professionals in the financial industry is \$11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is \$80,000. The sample mean salary for mid-level professionals in Company B is \$96,000. This data is used to test the claim that mid-level professionals at Company A and Company B are paid differently, on average.

4. It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

5. It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

6. A sample of 12 in-state graduate school programs at school A has a mean tuition of \$64,000 with a standard deviation of \$8,000. At school B, a sample of 16 in-state graduate programs has a mean of \$80,000 with a standard deviation of \$6,000. On average, are the mean tuitions different?

7. The manufacturer of a new medicine claims that the drug helps improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after. This data is then used to test the manufacturer's claim.

8. A study is done to determine which of two soft drinks has more sugar. The researchers believe that Beverage B has more sugar than Beverage A, on average. A random sample of 13 cans of Beverage A is selected; the mean amount of sugar in the sample is 36 grams with a standard deviation of 0.6 grams. A random sample of 10 cans of Beverage B is selected; the mean amount of sugar in this sample is 38 grams with a standard deviation of 0.8 grams. Both populations have normal distributions.

a. What are the hypotheses for this test?
b. What test should be used? Why?
c. Is this a one-tailed or two-tailed test?

9. The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

a. State the null and alternative hypotheses.
b. Which distribution (normal or Student's *t*) would we use for this hypothesis test? Explain.
c. Calculate the test statistic and *p*-value.
d. Using a significance level of $\alpha = 0.05$, what is the decision about Ho?
e. What can we conclude from this test?

10. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. The populations (of fastball speeds) for each pitcher have normal distributions. A sample of 14 fastball pitches is measured from each pitcher; the results are shown in the table:

Pitcher	Sample mean (mph)	Population SD (mph)
Wesley	86	3
Rodriguez	91	7

Scouters believe that Rodriguez pitches a speedier fastball than Wesley; use the data above to test this claim.

a. State the null and alternative hypotheses in terms of the appropriate parameters.
b. What test should be used, and why?
c. Calculate the test statistic and the *p*-value.
d. At the 1% significance level, what is the decision about Ho?
e. What can we conclude from this test?

11. A researcher is testing the effects of plant food on plant growth; the researcher thinks the food makes the plants grow taller. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants (in inches) are recorded after eight weeks. The following table shows the results.

Plant Group	Sample Mean	Population SD
Food	16	2.5
No Food	14	1.5

Assume the populations have normal distributions and answer the following:

a. State the null and alternative hypotheses.
b. What type of test should be used, and why?
c. Find the test statistic and the *p*-value.
d. At the 1% significance level, what is the decision about Ho?
e. What can we conclude from the test?

12. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared, and the melting points of each alloy have normal distributions. 15 pieces of each metal are being tested; the results are shown in the following table. This data will be used to test the claim that Alloy Zeta has a different melting point than Alloy Gamma.

	Sample Mean (°F)	Population SD (°F)
Alloy Gamma	800	95
Alloy Zeta	900	105

a. State the null and alternative hypotheses.
b. Which type of test should be used, and why?
c. Calculate the test statistic and the *p*-value.
d. Using a 1% significance level, what is our decision about Ho?
e. State a conclusion for the test.

13. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS2 had system failures within the first eight hours of operation. OS2 is believed to be more stable (have fewer crashes) than OS1.

a. State the null and alternative hypotheses.
b. What type of test should be used, and why?
c. Find the test statistic and the *p*-value.
d. At the 5% significance level, what is the decision about Ho?
e. What can you conclude about the two operating systems?

14. In the recent Census, three percent of the U.S. population reported being of two or more races. However, this proportion varie from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percentages are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

a. State the null and alternative hypotheses.
b. Is this a right-tailed, left-tailed, or two-tailed test?
c. Which test would we use for this hypothesis test, and why?
d. Calculate the test statistic and *p*-value.
e. Using a significance level of $\alpha = 0.05$, what is the decision?
f. Write a conclusion for the test.

15. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in the table below. Each installation was tested before and after the patch, and the number of system failures recorded in the table below. Using a 1% significance level, test the claim that the average number of system failures is reduced after installing the patch. Assume the differences have a normal distribution.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

a. What is the random variable?
b. State the null and alternative hypotheses.
c. What kind of test should be used and why?
d. What are the test statistic and *p*-value?
e. What is the decision for Ho?
f. What conclusion can we draw about the software patch?

16. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. Assume that the differences have a normal distribution, and test the claim that the mean number of balls increased after the class; use a 5% significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

a. State the null and alternative hypotheses.
b. What kind of test should be used and why?
c. What are the test statistic and *p*-value?
d. What is the decision for Ho?
e. What conclusion can we draw about the juggling class?

17. A doctor wants to know if a medication is effective in lowering blood pressure. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test the doctor’s claim at the 1% significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

a. State the null and alternative hypotheses.
b. What is the test statistic?
c. What is the *p*-value?
d. What is the decision about Ho?
e. What is the conclusion?

NOTE

In the following problems, if you are using a Student's *t*-distribution, including for paired data, you may assume that the underlying population is normally distributed. If no significance level is given, assume that α = .05.

18. The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Is there a significant difference in the means?

19. A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

20. At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time	Jumping Time
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

21. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if there is evidence that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary. (Use α = .05)

22. Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the following information to answer the next two exercises: The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals:

Western	**Eastern**
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

23. Suppose that a hypothesis test is conducted to test the claim that Western Division teams score more goals, on average, than Eastern Division teams. The **exact** distribution for this hypothesis test would be:

a. the normal distribution
b. the Student's *t*-distribution
c. the uniform distribution
d. the exponential distribution

24. If the level of significance for the test in **#23** is $\alpha = 0.05$, then the conclusion would be:

a. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
b. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
c. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
d. Unable to determine

Use the following information to answer the next two exercises: A statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the day students and the "night" subscript refers to the night students.

25. If this data is used to test the instructor's claim, which of the following would be the appropriate conclusion?

a. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
b. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
c. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

26. An appropriate alternative hypothesis for the hypothesis test is:

a. $\mu_{day} > \mu_{night}$ b. $\mu_{day} < \mu_{night}$ c. $\mu_{day} = \mu_{night}$ d. $\mu_{day} \neq \mu_{night}$

27. Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

28. A test is conducted to compare two types of diet. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds. Use this data and a 5% significance level to test the claim that the liquid diet yields a higher mean weight loss than the powder diet.

29. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Does this data provide evidence that the mean cost for auto insurance for teenage boys is greater than that for teenage girls?

30. A group of transfer bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1,011, respectively. From past studies, the population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

31. Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of 7 mpg. And 31 non-hybrid sedans get a mean of 22 mpg with a standard deviation of 4 mpg. Suppose that the population standard deviations are known to be 6 and 3, respectively. Conduct a hypothesis test to evaluate the manufacturers' claim.

32. A study is done to determine if students in the California State University system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California State University system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California State University system students took on average 4.5 years with a standard deviation of 0.8. The private university students took an average of 4.1 years with a standard deviation of 0.3. Use this data to test the claim.

33. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). The table below contains ten of the paired responses for husbands and wives. Use this data to conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's score	2	2	3	3	4	2	1	1	2	4
Husband's score	2	2	1	3	2	1	1	1	2	4

34. A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to test whether the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them. Does this data provide evidence to support the claim?

35. A test is conducted to determine whether there is a difference in the proportions of female suicide victims that are ages 15 to 24 are the same for whites and for African-Americans in the United States. We randomly pick one year, 1992, to compare the populations. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24. We will let female suicide victims be our population.

Use the following information to answer the next three exercises. Neuro-invasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuro-invasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuro-invasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuro-invasive West Nile virus cases more than the 2010 proportion of neuro-invasive West Nile virus cases? Using a 1% level of significance, we conduct an appropriate hypothesis test.

36. This is:

a. a test of two proportions
b. a test of two independent means
c. a test of a single mean
d. a matched pairs test.

37. An appropriate null hypothesis is:

a. $p_{2011} \leq p_{2010}$
b. $p_{2011} \geq p_{2010}$
c. $\mu_{2011} \leq \mu_{2010}$
d. $p_{2011} > p_{2010}$

38. The *p*-value is 0.0022. At a 1% level of significance, the appropriate conclusion is:

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuro-invasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuro-invasive West Nile disease.

b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuro-invasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuro-invasive West Nile disease.

c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuro-invasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuro-invasive West Nile disease.

d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuro-invasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuro-invasive West Nile disease.

39. A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. Does this data provide evidence that the percent of Hispanic students at the two colleges is significantly different? Explain.

40. Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in the table below. At the1% level of significance, is there evidence to support the claim?

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

41. Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders. Does this data indicate that there is a significant difference in the proportions between the age groups?

42. A group of friends debated whether a higher percentage of men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Use this data to test the friends' conjecture at the 5% level of significance.

43. Two computer users were discussing tablet computers. They conjectured that a higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. The table below shows the number of tablet owners for each age group. Test the claim using a 5% level of significance.

	16-29 yr old	30 yrs and older
Own a Tablet	69	231
Sample size	628	2309

44. A teacher is interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was \$31.14 with a standard deviation of \$4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was \$33.86 with a standard deviation of \$10.87. Use this data to test the claim that children's educational software costs less, on average, than children's entertainment software.

45. A social scientist recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey at a college to test her claim. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Does this data provide evidence to reject her claim?

46. Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded in the table blow. Does this data provide evidence that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Use the following information to answer the next two exercises. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same. Let the subscript *t* = treated patient and *ut* = untreated patient.

47. The appropriate hypotheses are:

a. H_0: $p_t < p_{ut}$ and Ha: $p_t \geq p_{ut}$
b. H_0: $p_t \leq p_{ut}$ and Ha: $p_t > p_{ut}$
c. H_0: $p_t = p_{ut}$ and Ha: $p_t \neq p_{ut}$
d. H_0: $p_t = p_{ut}$ and Ha: $p_t < p_{ut}$

48. If the *p*-value is 0.0062 and we use $\alpha = .05$, what is the conclusion?

a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.
c. There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients who develop AIDS after four years.
d. There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.

49. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows:

	Player 1	Player 2	Player 3	Player 4
Score before class	83	78	93	87
Score after class	80	80	86	86

Does this data provide evidence that the technique is effective?

50. A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are shown in the table:

Southern States	**2012**	**2013**
Alabama	3450	3720
Arkansas	2150	2280
Florida	15,540	15,710
Georgia	6970	7310
Kentucky	3160	3300
Lousiana	3320	3630
Mississippi	1990	2080
North Carolina	7090	7430
Oklahoma	2630	2690
South Carolina	3570	3580
Tennessee	4680	5070
Texas	15,050	14,980
Virginia	6190	6280

Using a 5% significance level, does this data provide evidence that there were more cases in 2013 than in 2012?

51. A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices (in dollars) for his two favorite hotel chains is in the table below. At a 5% level of significance, is there evidence that the prices are different, on average, between the two chains?

City	Hyatt Regency Price	Hilton Price
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, D.C.	245	239

52. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting officer wants to estimate the difference in the means. He randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were \$46,100 and \$46,700, respectively, and their standard deviations were \$3,450 and \$4,210, respectively. Construct a 95% confidence interval for the difference in the means. Does this interval suggest that the means could be the same? Explain.

53. Marketing companies have collected data implying that on average, teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Use this data to construct a 95% confidence interval for the difference in means between girls and boys. According to this interval, is there evidence that on average, girls use more ring tones that boys? How big is this difference?

54. Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of 7 mpg. And 31 non-hybrid sedans get a mean of 22 mpg with a standard deviation of 4 mpg. Suppose that the population standard deviations are known to be 6 and 3, respectively. Use this data to construct a 95% confidence interval for the difference in mean mpg between hybrid cars and non-hybrid cars.

55. A study is done to determine if students in the California State University system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California State University system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California State University system students took on average 4.5 years with a standard deviation of 0.8. The private university students took an average of 4.1 years with a standard deviation of 0.3. Use this data to find a 95% confidence interval for the difference in means between the Cal State system and private universities.

56. A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. Use this data to construct a 90% confidence interval for the difference in the proportions.

57. Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in the table below. Use this data to construct a 95% confidence interval for the difference in obesity rates between men and women.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

58. An accounting firm is trying to decide between IT training that is conducted in-house or conducted by consultants. The table below shows the average annual training cost per employee. Are the mean costs significantly different using 5% level of significance? Assume the population variances are equal. Show all hypothesis testing steps.

	n	mean	Sample stdev
In-house	200	$485	$32
Consultants	185	$502	$40

59. According to CNNMoney article, the millennials who are eligible for retirement plan from workplace are saving at the same participation rate as older generations. Test this claim at a 1% level of significance. Show all hypothesis testing steps. Suppose the following data is collected:

	n	proportion
Millennials	300	94%
Other Generations	330	92%

REFERENCES

10.1 Two Population Means with Unknown Standard Deviations

Data from Graduating Engineer + Computer Careers. Available online at http://www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

“List of current United States Senators by Age.” Wikipedia. Available online at http://en.wikipedia.org/wiki/ List_of_current_United_States_Senators_by_age (accessed June 17, 2013).

“Sectoring by Industry Groups.” Nasdaq.Available online at http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry (accessed June 17, 2013).

“Strip Clubs: Where Prostitution and Trafficking Happen.” Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

“World Series History.” Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

10.2 Two Population Means with Known Standard Deviations

Data from the United States Census Bureau. Available online at http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Hinduja, Sameer. “Sexting Research and Gender Differences.” Cyberbulling Research Center, 2013. Available online at http://cyberbullying.us/blog/sexting-research-and-gender-differences/ (accessed June 17, 2013).

“Smart Phone Users, By the Numbers.” Visually, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed June 17, 2013).

Smith, Aaron. “35% of American adults own a Smartphone.” Pew Internet, 2013. Available online at http://www.pewinternet.org/~/media/Files/Reports/2011/PIP_Smartphones.pdf (accessed June 17, 2013).

“State-Specific Prevalence of Obesity Among Adults—Unites States, 2007.” MMWR, CDC. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm (accessed June 17, 2013).

“Texas Crime Rates 1960–1012.” FBI, Uniform Crime Reports, 2013. Available online at: http://www.disastercenter.com/ crime/txcrime.htm (accessed June 17, 2013).

10.3 Comparing Two Independent Population Proportions

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013). Data from Hyatt Hotels. Available online at http://hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services. Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013). Data from the Chancellor’s Office, California Community Colleges, November 1994.

“State of the States.”Gallup,2013. Available online at http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive (accessed June 17, 2013).

“West Nile Virus.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/ westnile/index.htm (accessed June 17, 2013).

That's no surprise. After paying bills. “66% Of Millennials Have Nothing Saved for Retirement.” *CNNMoney*, Cable News Network, money.cnn.com/2018/03/07/retirement/millennial-retirement-savings/index.html?sr=twCNN030718millennial-retirement-savings0103PMStory (accessed July 18, 2018).

11 | THE CHI-SQUARE DISTRIBUTION

Figure: The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

Chapter Objectives

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

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square single variance hypothesis tests.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.

11.1 | Introduction to Chi-Square Distribution

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the **chi-square distribution**.

In this chapter, you will learn the three major applications of the chi-square distribution:

1. the test of a single variance, which tests variability, such as in the coffee example
2. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
3. the test of independence, which determines if categorical variables are independent, such as in the movie example

NOTE: Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a Chi-Square Distribution table available (see Appendix). TI-83+ and TI-84 calculator instructions are included in the text.

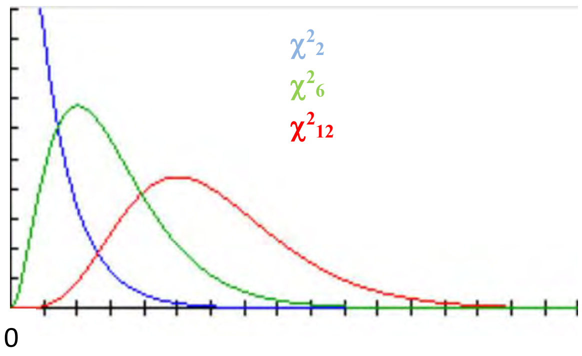
The Chi-Square Distribution

Let's conduct the following statistical experiment. We select samples of size n from a normal population, which has a standard deviation of σ. We find that the standard deviation in our sample is equal to s. Given these data, we can define a statistic, called chi-square, using the following formula:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

This new distribution has the following characteristics:

1.) All chi-square values are greater than or equal to 0
2.) There is a different chi-square curve for each degrees of freedom, n – 1 (figure below)
3.) The curve is nonsymmetrical and skewed to the right
4.) The total area under the curve is 1, or 100%
5.) The notation is $X \sim \chi^2_{df}$



NOTE: The degrees of freedom for the three major applications of Chi-square distribution are each calculated differently.

For the χ^2 distribution, the population mean is μ = df (degrees of freedom) and the population standard deviation is σ = $\sqrt{2(df)}$. The random variable is shown as χ^2, but may also be any upper case letter.

Critical values of the Chi-Square Distribution

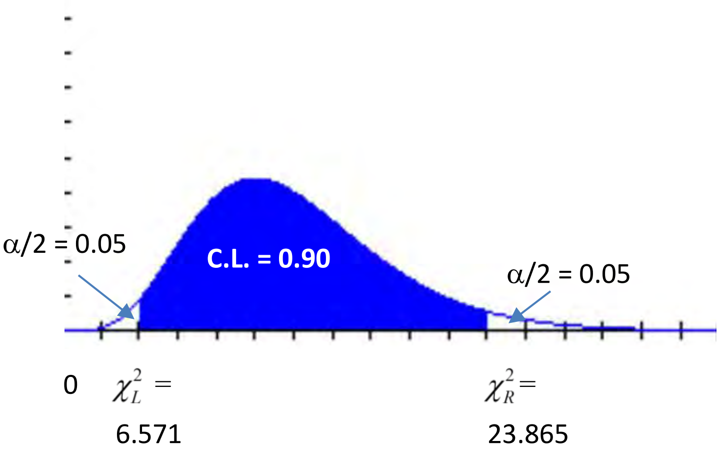
Recall that critical values are values from the distribution that separate the confidence area and the non-confidence area. We found $z_{\alpha/2}$ by using ±invnorm(α/2) and $t_{\alpha/2}$ by using t-distribution chart or by ±invT(α/2, df). Since the z-distribution and the t-distribution are symmetrical, the left and right critical values are opposite values. However, in the chi-square distribution, values are only positive.

χ_L^2 represents the left critical value.

χ_R^2 represents the right critical value.

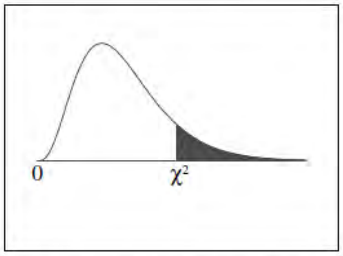
The critical values for the χ^2 distribution are recorded in a table; to find these values, you need area in the right tail and the degrees of freedom.

For example, let C.L. = 0.90, α = 0.10, α/2 = 0.05 and df = 14

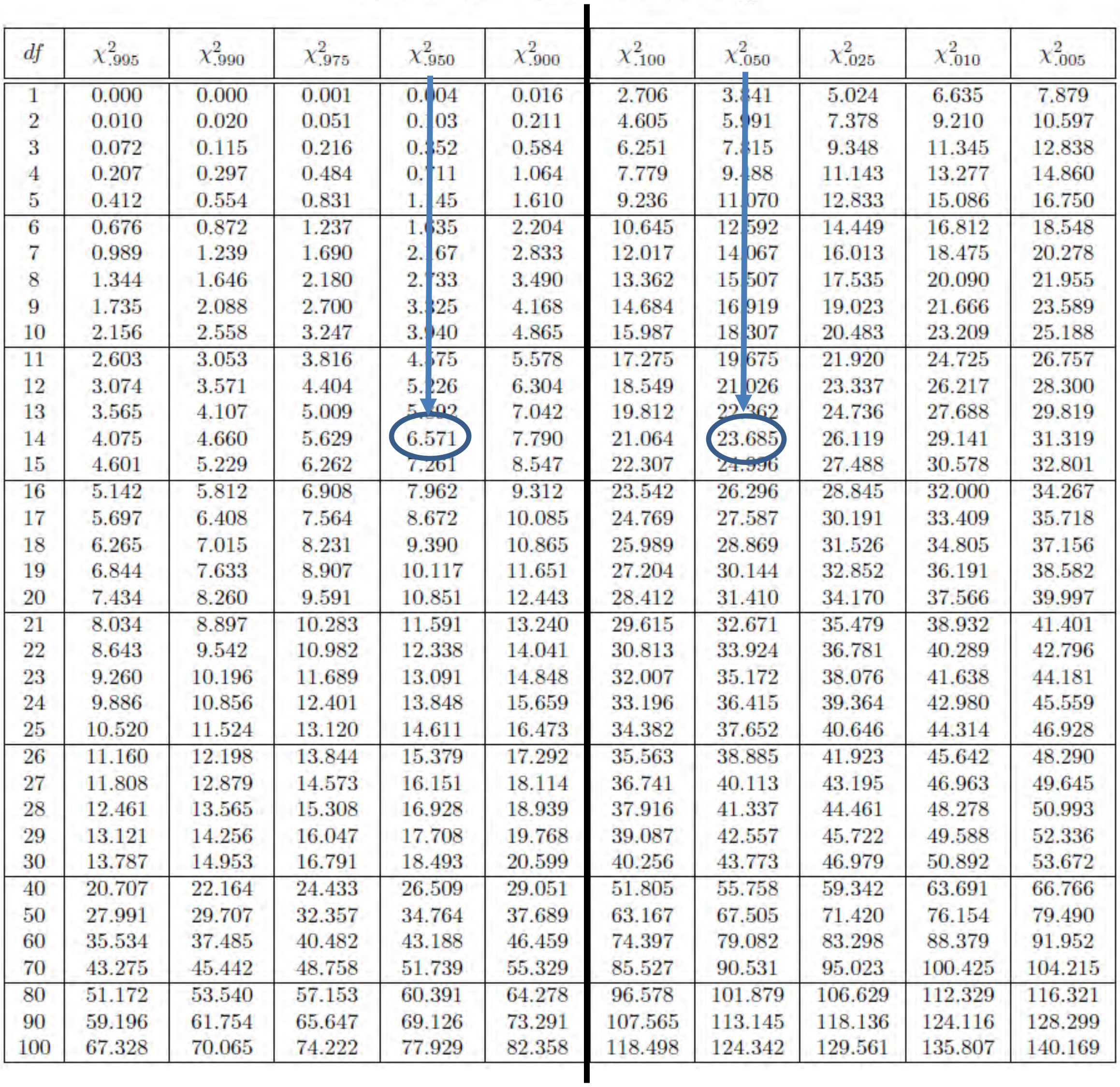


From the next page you can see, that the right side of the table is to find χ_R^2. The left side of the table is to find χ^2_L. The area to the right for χ_R^2 is .05 and df = 14 so χ_R^2 = **23.865**. The area to the right of χ_L^2 is .95 since the area in the left tail is .05. χ_L^2 = **6.571**.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.



df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Try It Σ

Find the left and right critical values for the chi-square distribution for c = .98 and df = 6

11.2 | Confidence Interval for Variance/Standard Deviation

Recall that an interval is a range of numbers. A **confidence interval** is a range of values, calculated from sample data, that is used to estimate an unknown population parameter. The interval gives us a lower bound value and upper bound for the parameter with a certain level of confidence.

Here we present intervals for estimating σ^2 and σ between a lower and upper bounds.

Unlike other confidence intervals we have seen, these bounds can't be found using the calculator. They depend on the chi-square critical values.

Calculating the Confidence Interval for variance

Lower bound < σ^2 < Upper bound

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

where χ_L^2, χ_R^2 are critical values, where d.f. = n – 1, and s^2 is the point estimate.

NOTE: χ_L^2, χ_R^2 are not to be squared. The right critical value is on the left side and the left on the right side. Also notice the numerators are the same. When you divide the same value by a larger number, you get a smaller value.

Calculating the Confidence Interval for standard deviation

Lower bound < σ < Upper bound

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$$

where s is the point estimate.

Example 11.1

The weights (in pounds) of 15 dogs selected randomly from those adopted out by an animal shelter last week are shown in the list below. Construct a 98% confidence interval for the population variance.

25, 34, 27, 27, 31, 28, 27, 28, 33, 31, 28, 29, 32, 29, 29

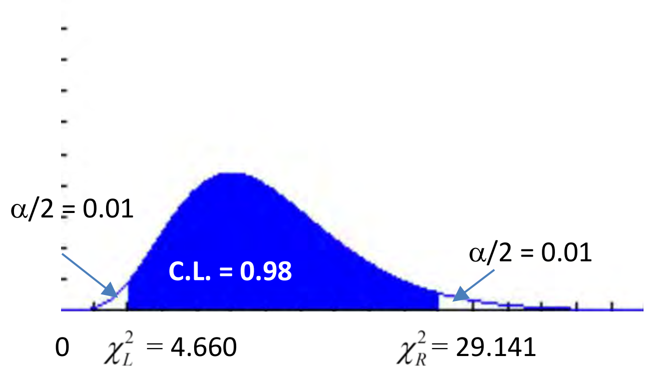
Solution 11. 1

First we need to find the sample variance (chapter 2) of the data set. Recall to find it by calculator you enter the data into L1 (Stat, Edit). 2nd Stat, Math, #8 Variance (L1):

$$s^2 = 6.314$$

Second, we find the critical values using the chi-square distribution chart.

Let C.L. = 0.98, $\alpha = 0.02$, $\alpha/2 = 0.01$ and df = n – 1 = 15 – 1 = 14



Third, we plug in the values into the formula.

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

$$\frac{14(6.314)}{29.141} < \sigma^2 < \frac{14(6.314)}{4.660}$$

$$3.033 < \sigma^2 < 18.969$$

Example 11.2

A random sample of 18 men have a mean height of 67.5 inches and a standard deviation of 1.5 inches. Construct a 99% confidence interval for the population standard deviation.

Solution 11.2

Since s = 1.5, $s^2 = 2.25$.

Let c = 0.99, $\alpha = 0.01$, $\alpha/2 = 0.005$ and df = n – 1 = 18 – 1 = 17

χ_L^2 **= 5.697** and χ_R^2 **= 35.718**

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}} \quad => \quad \sqrt{\frac{17(2.25)}{35.718}} < \sigma < \sqrt{\frac{17(2.25)}{5.697}}$$

$$1.035 < \sigma < 2.591$$

11.3 | Test of a Single Variance

A test of a single variance assumes that the underlying distribution is normal. The null and alternative hypotheses are stated in terms of the population variance, σ^2 (or population standard deviation, σ). The test statistic is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

where:

n = the total number of data

s^2 = sample variance

σ^2 = population variance

df = n - 1

A test of a single variance may be right-tailed, left-tailed, or two-tailed.

Example 11.3

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Solution 11.3

Even though we are given the population standard deviation, we can set up the test using the population variance as follows.

H_0: $\sigma^2 \leq 5^2$

H_a: $\sigma^2 > 5^2$ (claim)

right-tailed test

Try It Σ

A SCUBA instructor wants to record the collective depths each of his student's dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

Example 11.4

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.

Solution 11.4 using p-value

Since the claim is that a single line causes **less** variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

H_0: $\sigma \geq 7.2^2$

H_a: $\sigma < 7.2^2$ (claim)

The word "less" tells you this is a left-tailed test.

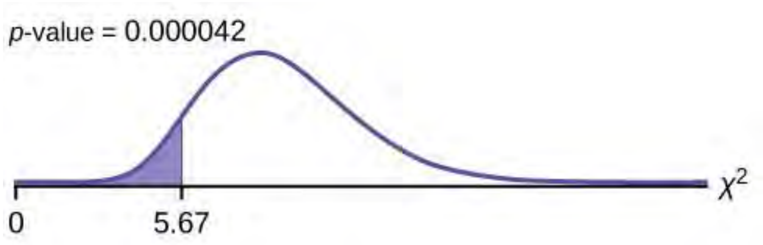
Distribution for the test: χ^2_{24}, where:

- n = the number of customers sampled
- df = n – 1 = 25 – 1 = 24

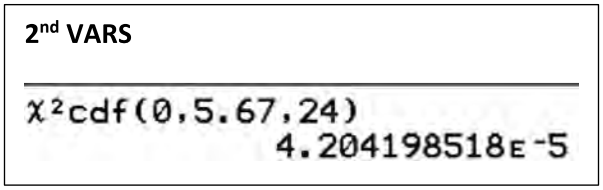
Calculate the test statistic: Substituting n = 25, s = 3.5, and σ = 7.2 (the value of σ in Ho), we get:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 5.67$$

Graph:



p-value = P ($\chi^2 < 5.67$) = χ^2cdf(0,5.67, 24) = 0.000042



2nd VARS

```
χ²cdf(0,5.67,24)
         4.204198518E-5
```

NOTE: remember that chi-square distribution starts at 0

Reject H_0 because p-value < α = 0.05

Conclusion: At a 5% level of significance, the data provides sufficient evidence to conclude that a single line causes a lower variation among the waiting times. That is, with a single line the standard deviation of customer waiting times is less than 7.2 minutes.

Solution 11.4 using critical value

H_0: σ ≥ 7.22

H_a: σ < 7.22 (claim)

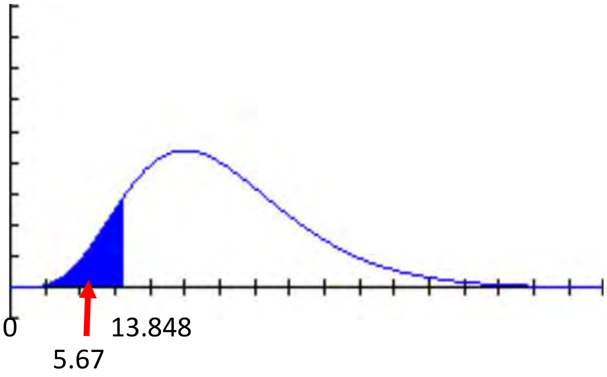
Calculate the test statistic: Substituting n = 25, s = 3.5, and σ = 7.2 (the hypothesized value of σ) we get: $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 5.67$

Find Critical Value:

We are only finding the left critical value since it is left-tailed. We do not split α = 0.05 since it is one-tailed. Therefore, using the chi-square (χ^2) chart, we need the area to the right which is 0.95 and df = n – 1 = 24.

$$\chi^2_L = \mathbf{13.848}$$

Since the test statistic, 5.67, is in the left critical region, we Reject H_0.



Conclusion is the same as above.

Try It Σ

The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August of 2012, the standard deviation of Internet speeds across Internet Service Providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, the test statistic, sketch the graph of the p-value, and draw a conclusion. Test at the 1% significance level.

11.4 | Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data "fits" a particular distribution or not. For example, you may suspect your unknown data fits a binomial or a uniform distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is: $\chi^2 = \sum \frac{(O-E)^2}{E}$,

where

- O = observed frequency
- E = expected values (from theory)
- Degrees of freedom = k – 1, where k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true.

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.

NOTE:

The expected value for each cell needs to be at least five in order for you to use this test.

Goodness of Fit Steps

1.) H_o: $p_1 = \%$, $p_2 = \%$, … $p_k = \%$
H_a: at least one percentage is different

or

H_o: the frequency distribution **fits** ____ distribution
H_a: the frequency distribution does **not fit** ___ distribution

2.) Test value is $\chi^2 = \sum \frac{(O-E)^2}{E}$

Calculator: χ^2 *GOF* Test
L1: Observed Values
L2: Expected Values

3.) P-value = χ^2cdf(Test Value, 10^99, df) since it is always right-tailed
4.) Reject H_o if p-value < α
5.) Interpret the decision to write a meaningful conclusion

Example 11.5

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to **Table 11.1**.

a. Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.
b. Can you use the information as it appears in the charts to conduct the goodness-of-fit test? If not, create a new frequency distribution to continue.
c. Find the degree of freedom.
d. Find the test value
e. Find the p-value
f. State the conclusion
g. Interpret the decision

Number of absences per term	Faculty perception
0 – 2	50
3 – 5	30
6 – 8	12
9 – 11	6
12+	2

Table 11.1

Number of absences per term	Observed number of students
0 – 2	35
3 - 5	40
6 – 8	20
9 – 11	1
12+	4

Table 11.2

A random survey across all mathematics courses was then done to determine the actual number (observed) of absences in a course. The chart in **Table 11.2** displays the results of that survey.

Solution 11.5

a.) H_0: Student absenteeism fits faculty perception.
Ha: Student absenteeism does not fit faculty perception.

b.) No, notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in **Table 11.3** and **Table 11.4.**

Number of absences per term	Faculty perception
0 – 2	50
3 – 5	30
6 – 8	12
9 +	8

Table 11.3

Number of absences per term	Observed number of students
0 – 2	35
3 - 5	40
6 – 8	20
9 +	5

Table 11.4

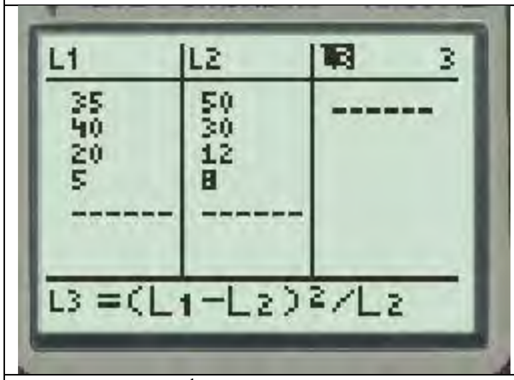
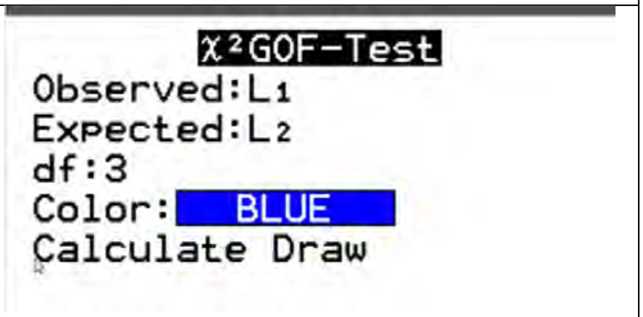
c.) Degrees of freedom (df) = k – 1 = number of categories – 1 = 4 – 1 = **3**

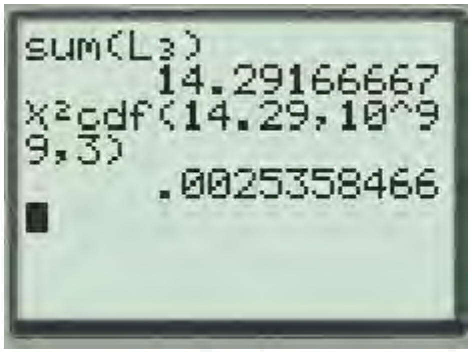
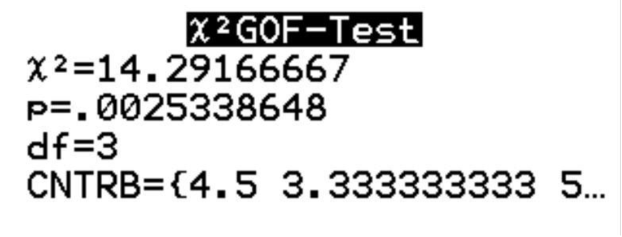
d.) Test Value using calculator
Older calculators like the TI-83 (plus) and some versions of TI-84 (plus) do not have χ^2-GOF Test, but there is still a way to find the χ^2 Test Value using your calculator.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate

Enter Observed frequency values into L1. Enter Expected frequency values in to L2.

TI-83 (plus)	TI-84 plus
In L3, enter the formula by highlighting L3 and then pressing Enter. Your cursor will now appear at the bottom of the screen. Type in the formula (L1 – L2)^2/L2. Then press enter to fill in L3. Take sum(L3) to find test value.	In the newer calculators, you can 2nd Quit after entering the Observed in L1 and Expected in L2. 2nd Quit, press Stat, Tests
 L1 L2 L3 3 35 50 ------ 40 30 20 12 5 8 ------ ------ L3 =(L1-L2)²/L2	 χ^2GOF-Test Observed:L1 Expected:L2 df:3 Color: BLUE Calculate Draw

Sum up L3 (2nd Stat, Math) to find the Test Value, (2nd Vars, χ^2CDF) to find the p-value

sum(L3)
14.29166667
χ^2cdf(14.29,10^99,3)
.0025358466
	
	χ^2GOF-Test
	χ^2=14.29166667
	p=.0025338648
	df=3
	CNTRB={4.5 3.333333333 5…

χ^2 Test Value = 14.29

e.) p-value = χ^2CDF(14.29, 10^99, 3) = .00254

f.) Reject Ho (Student absenteeism fits faculty perception.) because p-value < α

g.) Therefore, Student absenteeism does not fit faculty perception.

Example 11.6

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent uniformly during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the table below. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

n = 60

Solution 11.6

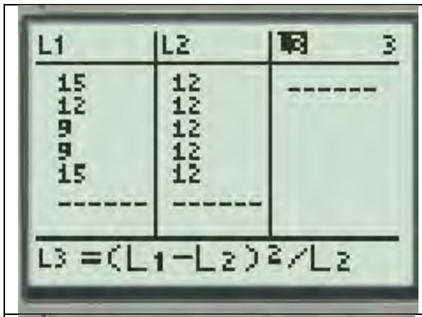
H_0: The absent days occur with equal frequencies, that is, they fit a uniform distribution.

H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

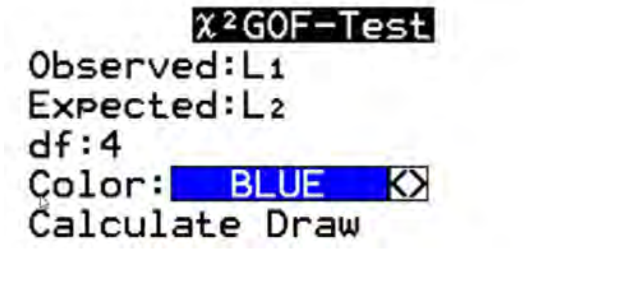
If the absent days occur with equal frequencies, then, out of 60 absent days there would be 12 absences expected on each day. These numbers are the expected (E) values. The values in the table are the observed (O) values or data.

L1: Observed (O) values (15, 12, 9, 9, 15)

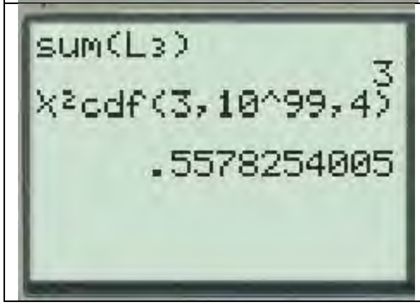
L2: Expected (E) values (12, 12, 12, 12, 12)



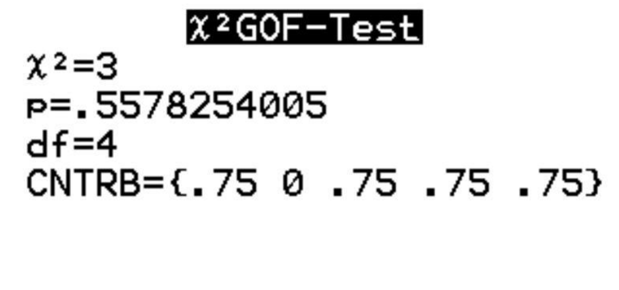
```
L1      L2      L3     3
15      12      ------
12      12
9       12
9       12
15      12
------  ------
L3 =(L1-L2)²/L2
```



```
χ²GOF-Test
Observed:L1
Expected:L2
df:4
Color: BLUE
Calculate Draw
```



```
sum(L3)
                 3
χ²cdf(3,10^99,4)
       .5578254005
```



```
χ²GOF-Test
χ²=3
p=.5578254005
df=4
CNTRB={.75 0 .75 .75 .75}
```

χ^2 Test Value = 3

p-value = χ^2CDF(3, 10^99, 4) = .5578

Do Not Reject Ho (The absent days occur with equal frequencies) because p-value ≥ α

Conclusion: At a 5% level of significance, the sample data does not provide sufficient evidence to conclude that the absences do not occur with equal frequencies.

Try It Σ

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 49 students were asked on which night of the week they did the most homework. The results were distributed as in the following table:

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Example 11.7

One study claims that the number of televisions that American families have is as follows:

10% of families have 0 televisions, 16% have 1 television, 55% have 2 televisions, 11% have 3 televisions, and 8% have 4+ televisions. A random sample of 600 families in the far western United States resulted in the following table.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15

Total = 600

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?

Solution 11.7

H_o: $p_1 = 10\%$, $p_2 = 16\%$, $p_3 = 55\%$, $p_4 = 11\%$, $p_5 = 8\%$

H_a: at least one percentage is different

To find the expected values, we need to take the percentages in H_0 (since H_0 is assumed true) and multiply each by n (n = 600):

Number of Televisions	Percent	Expected Frequency
0	10%	(0.10)(600) = 60
1	16%	(0.16)(600) = 96
2	55%	(0.55)(600) = 330
3	11%	(0.11)(600) = 66
4+	8%	(0.08)(600) = 48

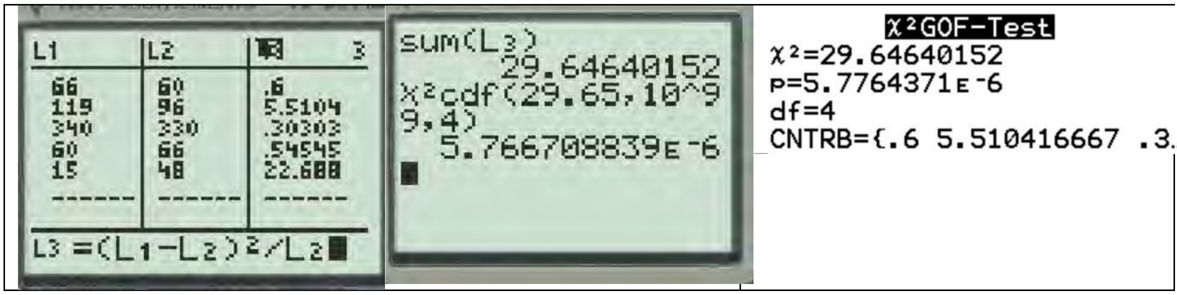
Remember that df for Goodness of Fit tests is k – 1 instead of n – 1.

df = 5 – 1 = 4; df ≠ 600 – 1

L1: 66, 119, 340, 60, 15

L2: 60, 96, 330, 66, 48

Calculate the test statistic: $\chi^2 = \chi^2$GOF-Test or Sum(L3) = 29.65



```
L1      L2      L3        3
 66      60     .6
 119     96     5.5104
 340     330    .30303
 60      66     .54545
 15      48     22.688
------  ------  ------
L3 =(L1-L2)²/L2
```

```
sum(L3)
      29.64640152
χ²cdf(29.65,10^9
9,4)
   5.766708839E-6
```

```
χ²GOF-Test
χ²=29.64640152
p=5.7764371E-6
df=4
CNTRB={.6 5.510416667 .3…
```

p-value = P(> 29.65) = χ^2CDF(29.65, 10^99, 4) = 0.000006

Reject H_0 because p-value < α

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.

Example 11.8

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.

Solution 11.8

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, 30TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the number of cells is three. Since X = the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

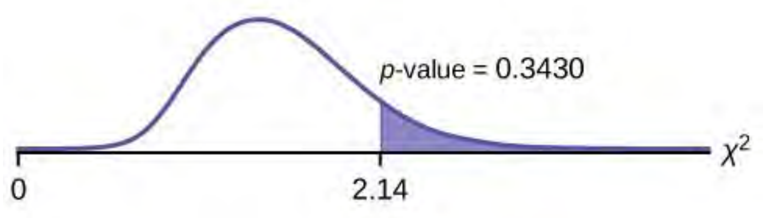
H_0: The coins are fair (uniformly distributed)

H_a: The coins are not fair.

Distribution for the test: χ^2 where df = 3 – 1 = 2.

Calculate the **test statistic**: $\chi^2 = 2.14$

Graph:



p-value = $P(\chi^2 > 2.14) = \chi^2$CDF(2.14, 10^99, 2) = 0.3430

Do Not Reject H_0 because p-value ≥ α

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

11.5 | Test of Independence and Test for Homogeneity

Tests of independence involve using a contingency table of observed (data) values. The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\chi^2 = \sum_{i \bullet j} \frac{(O-E)^2}{E}$$

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 · j terms of the form $\frac{(O-E)^2}{E}$

A test of independence determines whether two factors are independent or not.

NOTE:

The expected value for each cell needs to be at least five in order for you to use this test.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of two factors, the null hypothesis states that the factors are independent and the alternative hypothesis states that they are not independent (dependent). If we do a test of independence using the example, then the null hypothesis is:

H_0: *Factor 1* and *Factor 2* are independent

H_a: *Factor 1* and *Factor 2* are dependent

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of **degrees of freedom for the test of independence** is:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the expected number (E):

$$E = \frac{(row\ total)(column\ total)}{total\ number\ surveyed}$$

Try It

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety- seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Example 11.9

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In the following table is a sample of the adult volunteers and the number of hours they volunteer per week. Is the number of hours volunteered independent of the type of volunteer?

Type of Volunteer	1 – 3 hours	4 – 6 hours	7 – 9 hours
Community College Students	111	96	48
Four-Year College Students	96	133	61
Nonstudents	91	150	53

Table: Number of Hours Worked Per Week by Volunteer Type (Observed) The table contains observed (O) values (data).

Solution 11.9

The observed table and the question at the end of the problem, "Is the number of hour's volunteered independent of the type of volunteer?" tell you this is a **test of independence**. The two factors are number of hours volunteered and type of volunteer. This test is always **right-tailed.**

H_0: The number of hours volunteered is independent of the type of volunteer.

H_a: The number of hours volunteered is dependent on the type of volunteer.

Distribution for the test: χ^2_4 since df = (3 columns – 1)(3 rows – 1) = (2)(2) = 4

Calculate the test statistic: χ^2 = 12.99 (calculator or computer)

Using the TI-83, 83+, 84, 84+ Calculator

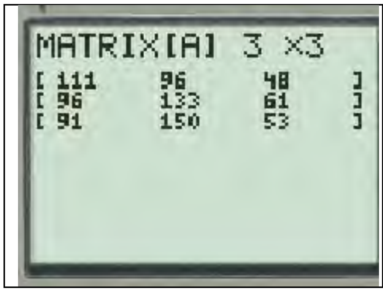
To calculate test value of independence test

Enter the 3 by 3 Contingency Table into Matrix [A] by going to 2nd Matrix, Edit

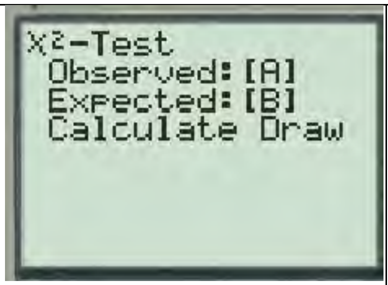
Press 2nd Quit to save the matrix and exit.

Go to STAT, TESTS, and select χ^2-Test; make sure Observed: [A]. Press Calculate.

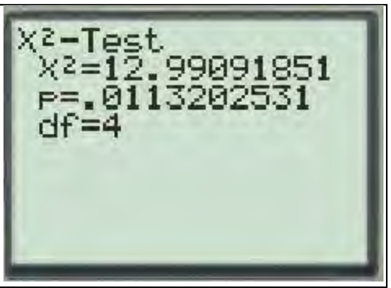
NOTE: You do not have to enter anything into Matrix [B]. Matrix [B] will automatically be filled with Expected Values.



```
MATRIX[A] 3 ×3
[ 111    96    48   ]
[ 96     133   61   ]
[ 91     150   53   ]
```

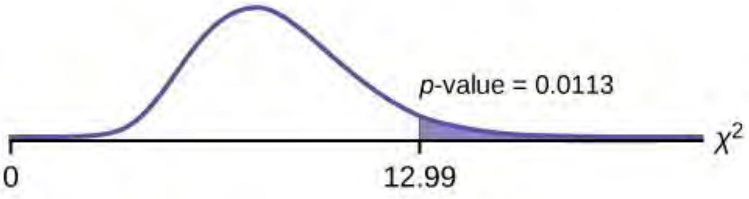


```
X²-Test
 Observed:[A]
 Expected:[B]
 Calculate Draw
```



```
X²-Test
 X²=12.99091851
 p=.0113202531
 df=4
```

Graph:



p-value = P($\chi^2 > 12.99$) = 0.0113

Since no α is given, assume α = 0.05. p-value = 0.0113.

Reject H₀ because p-value < α.

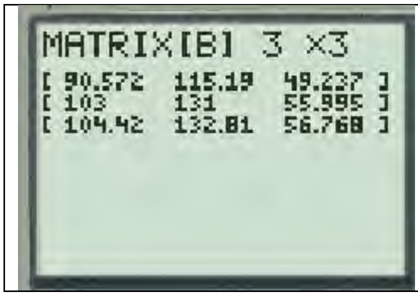
This means that the factors are not independent.

Conclusion: At a 5% level of significance, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

Using the TI-83, 83+, 84, 84+ Calculator

To see the Expected Values

Expected Values of the Independence test were placed into Matrix [B]. To view them, Press 2nd Matrix, Edit, [B]



```
MATRIX[B] 3 ×3
[ 90.572  115.19  49.237 ]
[ 103     131     55.995 ]
[ 104.42  132.81  56.768 ]
```

Type	1 – 3 hrs	4 – 6 hrs	7 – 9 hrs
CC students	90.572	115.19	49.237
4-YR students	103	131	55.995
Nonstudents	104.42	132.81	56.768

Try It Σ

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table below shows the results:

Industry Sector	**2000**	**2010**	**2020**
Nonagricultural wage and salary	13,243	13,044	15,018
Goods-producing, excluding agriculture	2457	1771	1950
Services-providing	10,786	11,273	13,068
Agriculture, forestry, fishing, and hunting	240	214	201
Nonagricultural self-employed and unpaid family worker	931	894	972
Secondary wage and salary jobs in agriculture and private household industries	14	11	11
Secondary jobs as a self-employed or unpaid family worker	196	144	152

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Example 11.10

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. Following table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Med Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

a. Test for independence using 1% level of significance using critical values

b. How many high anxiety level students are expected to have a high need to succeed in school?

Solution 11.10

a. Test for independence

H$_0$: The need to succeed in school is independent of anxiety level

H$_a$: The need to succeed in school is dependent of anxiety level

Distribution for the test: χ_4^2 since df = (5 columns – 1)(3 rows – 1) = (4)(2) = 8

Calculate the test statistic: χ^2 = 48.42 (calculator or computer)

Enter the 3 by 3 Contingency Table into Matrix [A] by going to 2nd Matrix, Edit

2nd Quit, STAT, TESTS, χ^2-Test, Calculate

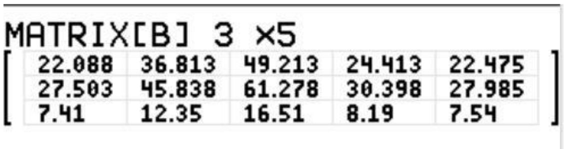
Find the Critical Value: α = 0.01 (do not split since it is right-tailed test), df = 8

$\chi_R^2 = 20.090$

Conclusion: Reject H$_0$ because 48.42 (test value) > 20.090 (critical value)

The need to succeed in school is related to the level of anxiety.

b. How many high anxiety level students are expected to have a high need to succeed in school? Here we are looking for an expected value. So we need to look at Matrix [B].



```
MATRIX[B] 3 ×5
[ 22.088  36.813  49.213  24.413  22.475 ]
[ 27.503  45.838  61.278  30.398  27.985 ]
[ 7.41    12.35   16.51   8.19    7.54   ]
```

Need to Succeed in School	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety
High Need	22.088	36.813	49.213	24.413	22.475
Med Need	27.503	45.838	61.278	30.398	27.985
Low Need	7.41	12.35	16.51	8.19	7.54

You can expect about 22 high anxiety level students to have high need to succeed in school.

A special type of independence test, called the **test for homogeneity**, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence. However, the hypotheses statements are stated differently.

Hypotheses

H$_0$: The distributions of the two populations are the same.

H$_a$: The distributions of the two populations are not the same.

Test Statistic (value)

Use a χ^2 test statistic. It is computed in the same way as the test for independence, using the matrix feature of the calculator.

Common Uses:

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Example 11.11

It is claimed that male and female college students have the same distribution of living arrangements. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other. The results are shown in the following table. Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05.

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Solution 11.11

H_0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.

H_a: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of Freedom, df = 3

Distribution for the test: χ^2

Calculate the test statistic: $\chi^2 = 10.1287$ (calculator or computer)

p-value = $P(\chi^2 > 10.1287) = 0.0175$

Since no α is given, assume $\alpha = 0.05$. p-value = 0.0175. p-value < α.

Reject H_0 because p-value < α. This means that the distributions are not the same.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Try It Σ

Do families and singles have the same distribution of cars? Use a level of significance of 0.05. Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, and van/SUV. The results are shown in the table below. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Example 11.12

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. The following table shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Solution 11.12

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

df = (number of columns – 1)(number of rows – 1) = 2(1) = 2

Distribution for the test: χ^2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator: χ^2 – Test)

p-value = $P(\chi^2 > 3.2603) = 0.1959$

$\alpha = 0.05$ and the p-value = 0.1959. p-value $\geq \alpha$

Do not reject Ho.

Conclusion: At a 5% level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

Key Terms

Critical Value is a value from the distribution that separates the confidence area from the non-confidence area.

Chi Square Distribution is a family of curves created by using variance and dependent on degrees of freedom. The shape of the curves is skewed right. The values of the distribution are greater than or equal to zero. μ = df, $\sigma = \sqrt{2(df)}$

Formula Review

Confidence Interval for Variance: $\dfrac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \dfrac{(n-1)s^2}{\chi_L^2}$

Confidence Interval for Standard Deviation: $\sqrt{\dfrac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\dfrac{(n-1)s^2}{\chi_L^2}}$

Review of Tests

You have seen the χ^2 test statistic used in four different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Test of a Single Variance/Standard Deviation**: Use the test to determine variation.

  Test statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$

  Degrees of freedom: df = n – 1

  Test may be left-, right-, or two-tailed

- **Goodness-of-Fit**: Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:

H_0: p_1 = %, p_2 = %, … p_k = %
H_a: at least 1 proportion is different than stated

or

H_0: The population fits the given distribution.
H_a: The population does not fit the given distribution.

Test statistic: $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ (calculator: χ^2GOF-Test)

Degrees of freedom: df = k – 1

Test is always right-tailed

- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:

  H_0: The two variables (factors) are independent.

  H_a: The two variables (factors) are dependent.

  Test statistic: $\chi^2 = \sum_{i \bullet j} \frac{(O-E)^2}{E}$ (calculator: χ^2-Test)

  $$E = \frac{(row\ total)(column\ total)}{total\ number\ surveyed}$$

  Degrees of Freedom: df = (number of rows – 1)(number of columns – 1)

  Test is always right-tailed

- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:

  H_0: The two populations follow the same distribution.

  H_a: The two populations have different distributions.

  Test statistic: $\chi^2 = \sum_{i \bullet j} \frac{(O-E)^2}{E}$ (calculator: χ^2-Test)

  $$E = \frac{(row\ total)(column\ total)}{total\ number\ surveyed}$$

  Degrees of Freedom: df = (number of rows – 1)(number of columns – 1)

  Test is always right-tailed

EXERCISES FOR CHAPTER 11

1. If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

2. If df > 90, the distribution is _______________. If df = 15, the distribution is ______________.

3. When does the chi-square curve approximate a normal distribution?

4. Find the critical values, χ^2_L and χ^2_R for c = 0.95 and n = 12.

5. Find the critical values, χ^2_L and χ^2_R for c = 0.98 and n = 20.

6. The grade point averages for 10 randomly selected students are listed below. Construct a 90% confidence interval for the population standard deviation, σ.
   2.0 3.2 1.8 2.9 0.9 4.0 3.3 2.9 3.6 0.8

7. An archer's standard deviation for his hits is six (data is measured in distance from the center of the target). An observer claims the standard deviation is less.
   a. What type of test should be used?
   b. State the null and alternative hypotheses.
   c. Is this a right-tailed, left-tailed, or two-tailed test?

8. The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.
   a. What type of test should be used?
   b. State the null and alternative hypotheses.
   c. df =

9. The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.
   a. What type of test should be used?
   b. What is the test statistic?
   c. What is the p-value?
   d. What can you conclude at the 5% significance level?

10. 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.
    a. Is the traveler disputing the claim about the average or about the variance?
    b. A sample standard deviation of 15 minutes is the same as a sample variance of ______ minutes.

c. Is this a right-tailed, left-tailed, or two-tailed test?
d. H_0:
e. df =
f. chi-square test statistic =
g. p-value =
h. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.
i. Let α = 0.05. What is the conclusion? Rewrite out the conclusion in complete sentences.
j. How did you know to test the variance instead of the mean?
k. If an additional test were done on the claim of the average delay, which distribution would you use?

11. A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15 oz. cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 oz. In order to determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated? Use α = 0.01.

12. Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

13. Isabella, an accomplished Bay to Breakers runner, claims that the standard deviation for her time to run the 7.5 mile race is at most three minutes. To test her claim, Rupinder looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes. Use α = 0.10.

14. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief. Use α = 0.05.

15. The number of births per woman in China is 1.6 down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether or not the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had had. The results are shown in the following table. Does the students' survey indicate that the standard deviation is greater than 0.75? Use α = 0.01.

# of Births	**Frequency**
0	5
1	30
2	10
3	5

16. According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two?
    Use $\alpha = 0.10$.
    Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; 11

17. The manager of "Frenchies" is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a ten-ounce order of fries is at most 1.5 oz., but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 oz. and a standard deviation of two oz.

    a. Find 90% confidence interval for population standard deviation.
    b. Test the claim at 5% level of significance.

18. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $1,249 with a very narrow standard deviation of $25. You find a website that has a price comparison for the same computer at a series of stores as follows: $1,299; $1,229.99; $1,193.08; $1,279; $1,224.95; $1,229.99; $1,269.95; $1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5% significance level. As a potential buyer, what would be the practical conclusion from your analysis?

19. A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 g, and there is a maximum allowed weight tolerance of 5% above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements? Conduct an appropriate hypothesis test.
    (a) at the 5% significance level
    (b) at the 1% significance level

    Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; 172;

20. *Determine the appropriate test to be used in the following exercises:*
    a. An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

    b. An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

c. A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

21. A teacher predicts that the distribution of grades on the final exam will be and they are recorded in the following table on the left and the actual distribution for a class of 20 is given on the right:

Predicted Distribution:

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

Actual Distribution:

Grade	Frequency
A	7
B	7
C	5
D	1

a. df =

b. State the null and alternative hypotheses.

c. χ^2 test statistic =

d. p-value =

e. At the 5% significance level, what can you conclude?

22. The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in the following table.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232

a. If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.
The percentage of each ethnic group in Santa Clara County is as in following table.

b. Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.
H0:
Ha:

c. Is this a right-tailed, left-tailed, or two-tailed test?

d. degrees of freedom =

e. χ^2 test statistic =

f. p-value =

g. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

h. Let $\alpha = 0.05$. Decision:

i. Conclusion (write out in complete sentences):

j. Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

23. A survey by Computerworld mobile data service, asked "how would you rate the overall trustworthiness of your mobile data provider?" The results from 2016 are shown in the following table:

Very trustworthy	31%
Somewhat trustworthy	41%
Somewhat untrustworthy	9%
Very untrustworthy	4%
No opinion	15%

Let's say there is a claim that the distribution has changed since then. You randomly select 500 students enrolled here this semester and survey them. The results are below:

Very trustworthy	38%
Somewhat trustworthy	35%
Somewhat untrustworthy	10%
Very untrustworthy	7%
No opinion	10%

Use a level of significance of 5% to test the claim. Show all steps.

24. *Determine the appropriate test to be used in the following exercises.*
  a. A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.
  b. The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.
  c. A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.

25. Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. The following table shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance they must travel.

Traveling Distance	Third class	Second class	First class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

  a. State the hypotheses, H_0 and H_a:
  b. df =

c. How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?
d. How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?
e. What is the test statistic?
f. What is the p-value?
g. What can you conclude at the 5% level of significance?

26. An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650 whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

a. Complete the table.

Smoking level per day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1 – 10						
11 – 20						
21 – 30						
31 +						
Totals						

b. State the hypotheses.
H_0:
H_a:
c. Find the Expected Values. Round to two decimal places. Place them into a table.
d. df =
e. χ^2 test statistic =
f. p-value =
g. Is this a right-tailed, left-tailed, or two-tailed test? Explain why.
h. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.
i. State the decision and conclusion (in a complete sentence) for $\alpha = 0.05$
j. State the decision and conclusion (in a complete sentence) for $\alpha = 0.01$

27. A math teacher wants to see if two of her classes have the same distribution of test scores.
a. What test should she use?
b. What are the null and alternative hypotheses?

28. A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

29. A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

30. Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in following table:

	20 – 29	30 – 39	40 – 49	50 - 59
Private Practice	16	40	38	6
Hospital	8	44	59	39

a. State the null and alternative hypotheses.
b. df =
c. What is the test statistic?
d. What is the p-value?
e. What can you conclude at the 5% significance level?

31. Which test would you use to decide whether two factors have a relationship?

32. Which test would you use to decide if two populations have the same distribution?

33. How are tests of independence similar to tests for homogeneity?

34. How are tests of independence different from tests for homogeneity?

35. Decide whether the following statements are true or false.
a. As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.
b. The standard deviation of the chi-square distribution is twice the mean.
c. The mean and the median of the chi-square distribution are the same if df = 24.

36. A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in the following table are the result of the 120 rolls.

Face Value	**Frequency**	**Expected frequency**
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

37. The marital status distribution of the U.S. male population, ages 15 and older, is as shown in the following table.

Marital Status	Percent	Expected frequency
Never married	31.3	
Married	56.1	
Widowed	2.5	
Divorced/separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in the Expected frequency in the above table, rounding to two decimal places.

Marital Status	Frequency
Never married	140
Married	238
Widowed	2
Divorced/separated	20

38. The columns in the following table contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Race/Ethnicity	AP Examinee	Overall Student	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

a. Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

b. Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

39. The City of South Lake Tahoe, CA, has an Asian population of 1,419 people, out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the Manhattan, NY, area yielded the data in the following table. Conduct a goodness-of-fit test to determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Race	Asian Indian	Chinese	Filipino	Japanese	Korean	Vietnamese	Other
Lake Tahoe Frequency	131	118	1045	80	12	9	24
Manhattan Frequency	174	557	518	54	29	21	66

40. Read the statement and decide whether it is true or false:
    a. In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.
    b. In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.
    c. Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week or not.
    d. The test to use to determine if a six-sided die is fair is a goodness-of-fit test.
    e. In a goodness-of fit test, if the p-value is 0.0113, in general, do not reject the null hypothesis.

41. UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in The Chronicle of Higher Education (2/2/2006). Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for part a and part b. The second column in each table does not add to 100% because of rounding.
    a. Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Women – Expected Major	Women – Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

    b. Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Men – Expected Major	Men – Actual Major
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	16.5%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

42. A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means any one type of recyclable material such as plastic or aluminum. The following table shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5% significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business Type	**Number in class**	**Observed # that recycle**	**Expected # that recycle**
Office	35	19	17.5
Retail/Wholesale	48	27	24
Food/Restaurants	53	35	26.5
Manufacturing/Medical	52	21	26
Hotel/Mixed	24	9	12

43. The following table contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (Years)	**Obese (Percentage)**	**Expected USA average (%)**
21 – 30	75	32.6
31 – 40	26.5	32.6
41 - 50	13.6	36.6
51 – 60	21.9	36.6
61 – 70	21	39.7

44. A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	**Beginner**	**Intermediate**	**Advanced**
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

45. Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in the following table. Conduct a test of independence.

Family Size	**Sub & Compact**	**Mid-size**	**Fulll-size**	**Van & Truck**
1	20	35	40	35
2	20	50	70	80
3-4	20	50	100	90
5 +	20	30	70	70

46. College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. The following table shows the data. Conduct a test of independence using $\alpha = .05$.

Major	**< \$50,000**	**\$50,000 - \$68, 999**	**\$69, 000 +**
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

47. Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in the following table. Conduct a test of independence using $\alpha = .10$.

Location	**20 – 29**	**30 – 39**	**40 – 49**	**50 and over**
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

48. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence using $\alpha = .05$.

Sport	**18 -25**	**26 – 30**	**31 – 40**	**41 and over**
Racquetball	42	58	30	46
Tennis	58	76	38	65
Swimming	72	60	65	33

49. A major food manufacturer is concerned that the sales for its skinny French fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in table below. Conduct a test of independence using $\alpha = .05$.

Type of Fries	**Northeast**	**South**	**Central**	**West**
Skinny Fries	70	50	20	25
Curly Fries	100	60	15	30
Steak Fries	20	40	10	10

50. According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence using $\alpha = .01$.

Age of Males	None	< $200,000	$200,000- $400,000	$400,001 - $1,000,000	$1,000,001+
20-29	40	15	40	0	5
30-39	35	5	20	20	10
40-49	20	0	30	0	30
50+	40	30	15	15	10

51. Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence using $\alpha = .05$

Annual Salary	Not a HS graduate	HS graduate	College graduate	Master's/Doctorate
< $30,000	15	25	10	5
$30,000 - $39, 999	20	40	70	30
$40,000 - $49,999	10	20	40	55
$50.000 - $59,999	5	10	20	60
$60,000 +	0	5	10	150

52. Read the statement and decide whether it is true or false
    a. The number of degrees of freedom for a test of independence is equal to the sample size minus one.
    b. The test for independence uses tables of observed and expected data values.
    c. The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.
    d. In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

53. An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on the following table, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

US Region	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Total	45	112	101	46	60	27	391

54. The following table provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Age Group/Net Worth Value (in millions)	**1 – 5**	**6 – 24**	**25 +**
17 – 25	8	7	5
26 – 30	6	5	9

55. A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/Ethnicity	**Asian-American**	**White/Non-Hispanic**	**African-American**	**Latino**
Against Tax	48	433	41	160
In Favor of tax	54	234	24	147
No Opinion	16	43	16	19
Total	118	710	71	272

56. A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in the following table. Conduct a test of homogeneity. Test at a 5% level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

57. Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place is shown in the table below. Conduct a test for homogeneity at a 5% level of significance.

	French Toast	Pancakes	Waffles	Omelets
Men	47	35	28	53
Women	65	59	55	60

58. A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5% level of significance.

59. In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In the following table you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5% significance level. Did you expect the result you obtained?

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)
Concern about the environment of other schools	1321	309
Dissatisfaction with academic instruction at other schools	1096	258
To provide religious or moral instruction	1257	540
Child has special needs, other than physical or mental	315	55
Nontraditional approach to child's education	984	99
Other reasons (e.g., finances, travel, family time, etc.)	485	216

60. When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in the following table shows the average energy use (in units of kg of oil equivalent per capita) in the USA and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5% significance level.

Year	European Union	United States
2010	3413	7164
2009	3302	7057
2008	3505	7488
2007	3537	7758
2006	3595	7697
2005	3613	7847

61. The Insurance Institute for Highway Safety collects safety information about all types of cars every year, and publishes a report of Top Safety Picks among all cars, makes, and models. In the following table presents the number of Top Safety Picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the Top Safety Picks safety award has remained the same between 2009 and 2013. Derive your results at the 5% significance level.

Year/Car Type	Small	Midsize	Large	Small SUV	Mid-Size SUV	Large SUV
2009	12	22	10	10	27	6
2013	31	30	19	11	29	4

REFERENCES

11.1 The Chi-Square Distribution

Data from *Parade Magazine.*

“HIV/AIDS Epidemiology Santa Clara County.” Santa Clara County Public Health Department, May 2011.

11.3 Test of a Single Variance

“AppleInsider Price Guides.” Apple Insider, 2013. Available online at http://appleinsider.com/mac_price_guide (accessed May 14, 2013).

Data from the World Bank, June 5, 2012.

11.4 Goodness-of-Fit Test

Data from the U.S. Census Bureau

Data from the College Board. Available online at http://www.collegeboard.com. Data from the U.S. Census Bureau, Current Population Reports.

Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I.S. Ockene, “Association between Eating Patterns and Obesity in a Free-living US Adult Population.” American Journal of Epidemiology volume 158, no. 1, pages 85-92.

Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, “Prevalence of Obesity in the United States, 2009–2010.” NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).

Stevens, Barbara J., “Multi-family and Commercial Solid Waste and Recycling Survey.” Arlington Count, VA. Available online at http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf (accessed May 24,2013).

11.5 Test of Independence/Homogeneity

DiCamilo, Mark, Mervin Field, “Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs.” The Field Poll, released Feb. 14, 2013. Available online at http://field.com/fieldpollonline/subscribers/ Rls2436.pdf (accessed May 24, 2013).

Harris Interactive, “Favorite Flavor of Ice Cream.” Available online at http://www.statisticbrain.com/favorite-flavor-of-ice- cream (accessed May 24, 2013)

“Youngest Online Entrepreneurs List.” Available online at http://www.statisticbrain.com/youngest-online-entrepreneur-list (accessed May 24, 2013).

Data from the Insurance Institute for Highway Safety, 2013. Available online at www.iihs.org/iihs/ratings (accessed May 24, 2013).

“Energy use (kg of oil equivalent per capita).” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE/countries (accessed May 24, 2013).

"Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubsearch/ pubsinfo.asp?pubid=2009030 (accessed May 24, 2013).

"Parent and Family Involvement Survey of 2007 National Household Education Survey Program (NHES)," U.S. Department of Education, National Center for Education Statistics. Available online at http://nces.ed.gov/pubs2009/2009030_sup.pdf (accessed May 24, 2013).

Gralla, Preston. "The Votes Are in: Which Mobile Data Provider Is Best?" *Computerworld*, Computerworld, 21 Dec. 2016, www.computerworld.com/article/3150992/wireless-carriers/the-votes-are-in-which-mobile-data-provider-is-best.html.

12 | LINEAR REGRESSION AND CORRELATION

Figure 12.1 Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

Chapter Objectives

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

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Determine if a correlation is significant
- Use the regression model to make predictions.
- Identify and interpret outliers.

INTRODUCTION

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be influenced 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. The type of data described in this chapter is **bivariate** data – the prefix "bi" indicating there are two variables. Although we will only study models involving two variables, in many real-world situations statisticians use **multivariate** data, meaning many variables.

In this chapter, we will be studying the simplest form of regression, "linear regression" with one independent variable. More specifically, given sample data measured on two variables, x and y, we wish to find a linear equation that best fits the observed sample data. We will also develop statistical measures and tests that measure how strong the linear relationship is between the variables.

12.1 | Linear Equations

We start by reviewing the basic facts about linear equations. A **linear equation** is an equation of the form:

$$y = a + bx$$

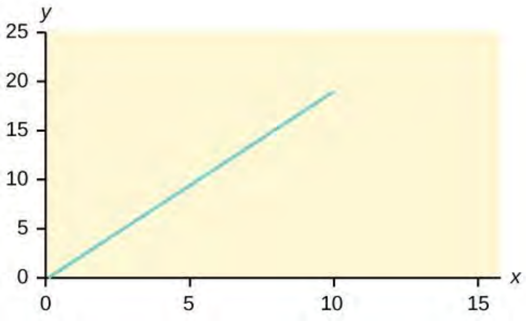
where a and b are constant numbers. This is called a linear equation, because the graph of the equation is a straight line.

The variable x is called the **independent variable**, and y is the **dependent variable.** That is, as written, y depends on x. Given any value, we can substitute it for the independent variable x to obtain the corresponding value for the dependent variable.

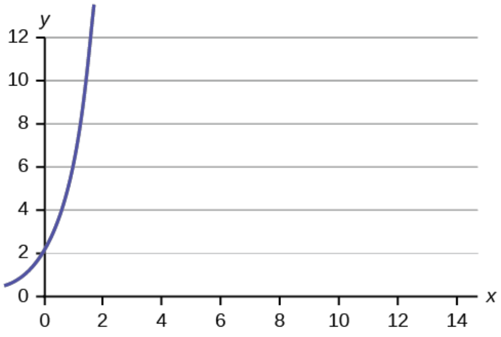
Example 12.1

Graph the following equations; which of these is linear?

a. $y = -1 + 2x$



b. $y = 2e^x$



Try It Σ

12.1 Is the following an example of a linear equation? If so, sketch the graph.

$$y = -0.125 - 3.5x$$

Example 12.2

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job. Find the equation that expresses the total cost in terms of the number of hours required to complete the job.

Solution 12.2

Let x = the number of hours it takes to get the job done. Let y = the total cost to the customer.

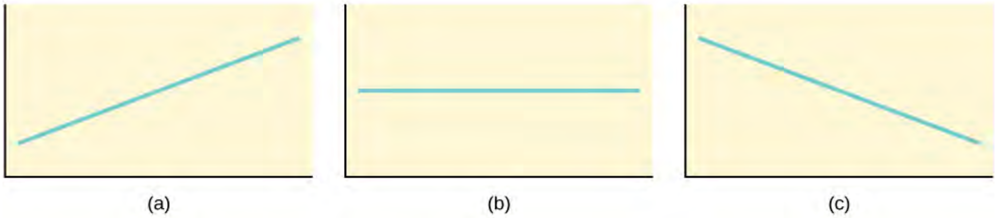
The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Try It Σ

12.2 Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Slope and Intercept of a Linear Equation

Given a linear equation $y = a + bx$, we call b the slope, and a is called the y-intercept. From basic algebra, recall that the slope is a number that describes the steepness of the line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis. The sign of the slope b determines whether the line slopes upward or downward, as shown in the figure below:



(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.

More specifically, when the slope b is positive, an increase in x results in an increase in y. And if the slope b is negative, then an increase in x results in a *decrease* in y.

Example 12.3

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 these using complete sentences.

Solution 12.3

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 additional hour she tutors, Svetlana earns an additional $15.

Try It Σ

12.3 Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

12.2 | Scatter Plots

Before we begin our discussion of linear regression and correlation, we need to a way to display the relation between two variables *x* and *y*. The most commonly used graph is a **scatter plot** (also called a scatter diagram). To make a scatterplot, we just plot the points, for each data pair using the *x*-value as the *x*-coordinate and the *y*-value as the *y*-coordinate. For small examples, this is easy to do by hand; but most statistical software packages will make very nice scatterplots, and they can be done on the TI-84 as well. We will illustrate this with an example.

Example 12.4

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table below show different depths with the maximum dive times in minutes.

X (depth in feet)	**Y** (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Construct a scatter plot. Let x = depth in feet and let y = maximum dive time in minutes.

Using the TI-83, 83+, 84, 84+ Calculator

To create a scatterplot:

1. Enter your X data into list L1 and your Y data into list L2.

2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Keep the other plots set to OFF.)

3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.

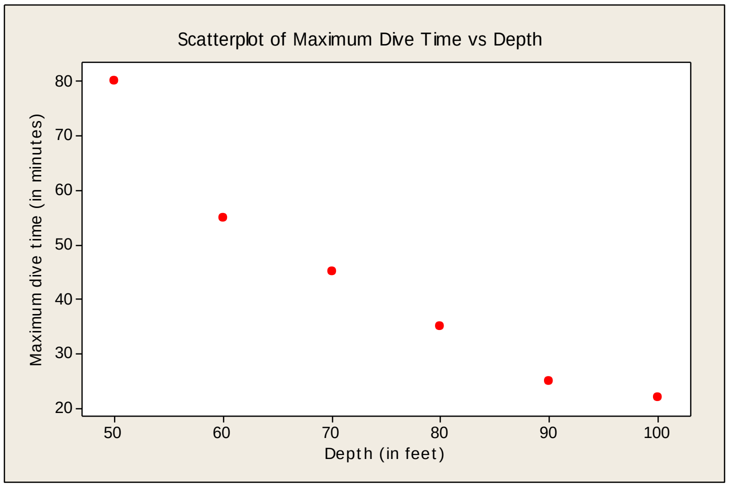
4. For Xlist: enter L1 and for Ylist: enter L2.

5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.

6. Make sure there are no other equations that could be plotted.
   Press Y = and clear any equations out.

7. Press the ZOOM key and then select option 9, "ZoomStat"; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Following these instructions, we get a graph like:



The points in the scatterplot exhibit a linear pattern; this suggests that a linear equation would be a good fit to this data. We also can see that as the depth increases, the maximum dive time de*creases*.

Try It Σ

12.4 Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	**Y** (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

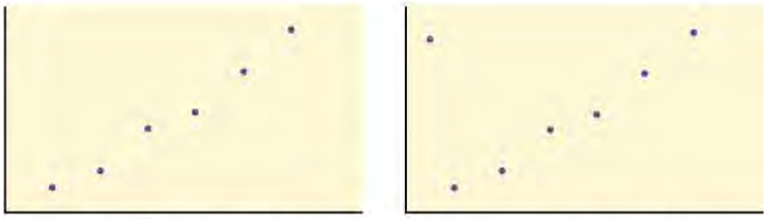
Construct a scatter plot and state whether Amelia's conjecture appears to be true.

A scatter plot shows the **direction** of the relationship between the variables.

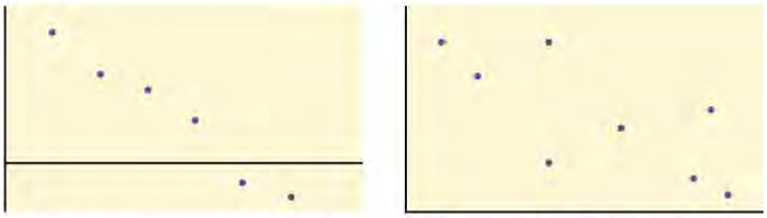
- If the points in the graph slope upward as we move left to right, then we say that there is a *positive* relationship between the variables *x* and *y*.
- If the points in the graph slope downward as we move left to right,, then we say that there is a *negative* relationship between the variables *x* and *y*.

We can also roughly determine the **strength** of the linear relationship by looking at the scatter plot; the more closely the points conform to a linear pattern, the better a linear equation will describe the relationship. If the points in the graph appear to follow a non-linear pattern (i.e. there is curvature) then perhaps a linear model is not appropriate.

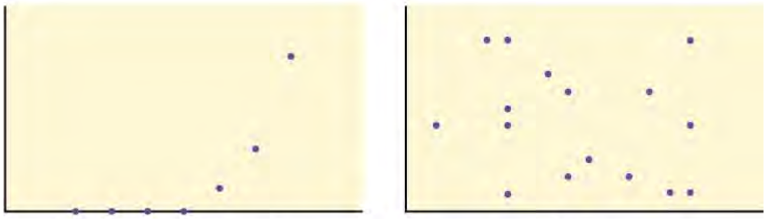
When we look at a scatterplot, we want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



(a) Positive linear pattern (strong) (b) Linear pattern w/ one deviation



(a) Negative linear pattern (strong) (b) Negative linear pattern (weak)



(a) Exponential growth pattern (b) No pattern

Again, in this chapter, we are interested in scatter plots that show a linear pattern. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to fit a line to the points in the scatter plot. This line can be calculated through a process called **linear regression**, which we will learn in the next section. We will also develop a numerical measure that provides a more objective measure of how well the line fits the data. And we will only use the fitted line when we have determined that there is a statistically significant relationship between the variables *x* and *y*.

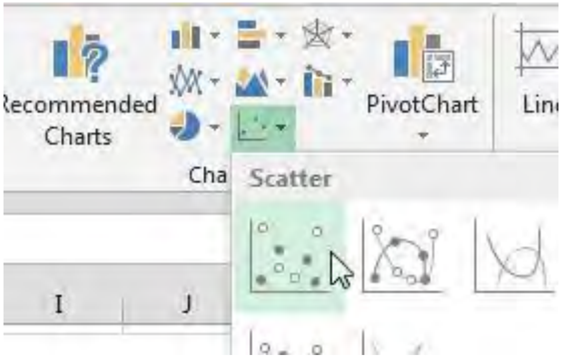
Example 12.5

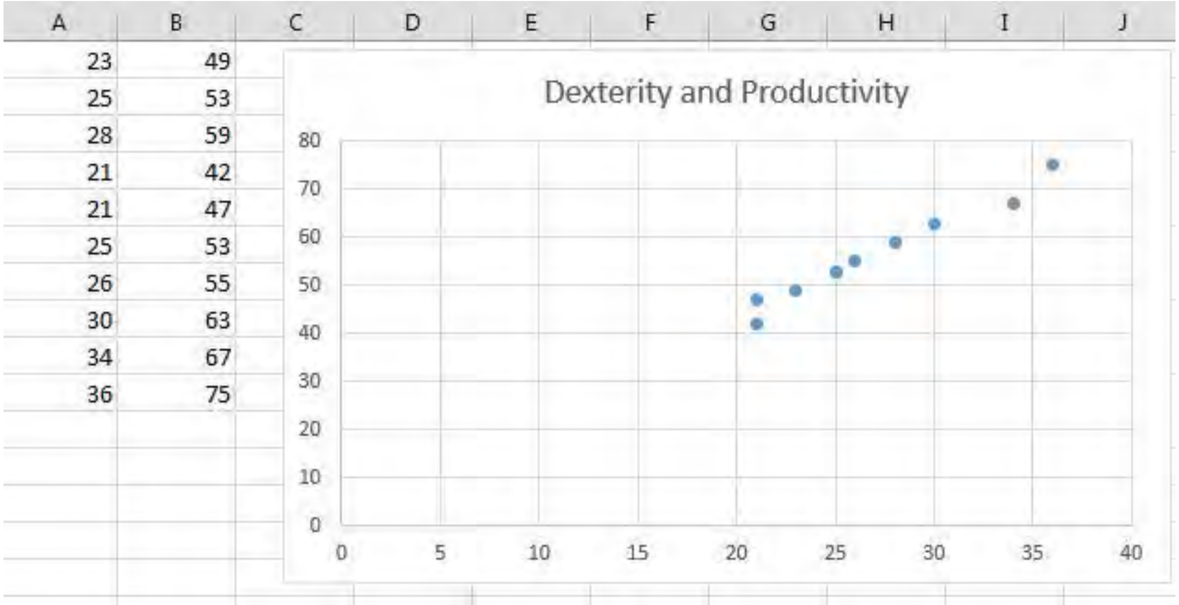
Two different tests are designed to measure employee productivity and dexterity. Several employees of a company are randomly selected and asked to complete the tests. The results are below. Draw a scatterplot using Excel, describe the scatterplot.

Dexterity	23	25	28	21	21	25	26	30	34	36
Productivity	49	53	59	42	47	53	55	63	67	75

Solution 12.5

First, place Dexterity values in column A and place Productivity values in column B. Second, highlight both columns, Insert Scatter. Choose the Scatter that is not connected.





A	B
23	49
25	53
28	59
21	42
21	47
25	53
26	55
30	63
34	67
36	75

Description: Strong positive linear pattern

12.3 | The Regression Equation

Data rarely fit a straight line exactly; usually we must be satisfied with estimates. In this section we discuss how to calculate the **Line of Best Fit**. This equation is also known as the **Least-Squares Line**.

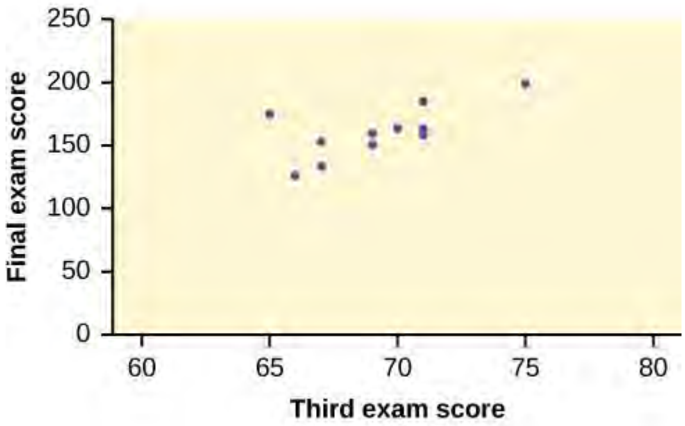
Example 12.6

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. Using the scatterplot, does it appear that we can predict the final exam score of a random student if we know the third exam score?

X (third exam score)	65	67	71	71	66	75	67	70	71	69	69
Y (final exam score)	175	133	185	163	126	198	153	163	159	151	159

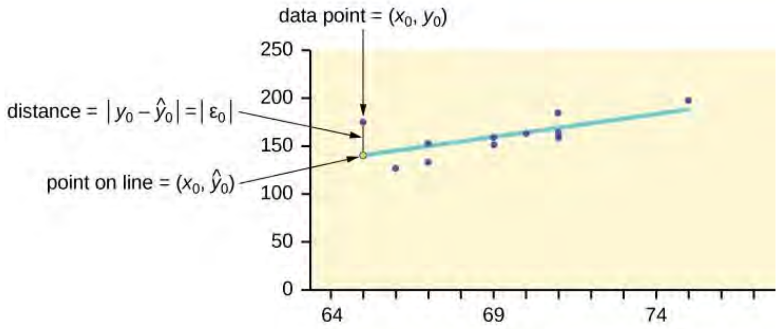
Solution 12.6

The scatterplot below exhibits a fairly tight, linear pattern. This indicates a linear model could be used to predict the final exam score from the score on the third exam.



Next we wish to find a line that best "fits" the data. To do this, we use what is called a **least-squares regression line** to obtain the line of best fit.

Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least- squares linear regression has the form $(x, \hat{y})$. The $\hat{y}$ is read "*y*-hat" and is the **estimated value of *y*;** that is, it is the value of *y* obtained by plugging *x* into the regression line. It is not generally equal to *y* from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the "error" or **residual** (ε is the Greek letter epsilon). It is not an error in the sense of a mistake; instead, it is the difference between the estimate and the true value of y. 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; since the point lies above the line, the residual is positive.

To get the line of best fit, our first impulse might be to minimize the sum of all the residuals – that is, minimize the sum of all vertical distances $y - \hat{y}$. However, this sum would end up very close to zero. The reason is that if the line really passes through the points, then the positive errors would cancel out the negative residuals, resulting in a sum of zero. To avoid this cancellation effect, we might instead minimize the sum of the actual vertical *distances* from the data points to the points on the line; these distances are given by the absolute value of the residuals. However, in order to minimize the sum, we need to use Calculus; and in Calculus the absolute value function is difficult to work with – so instead, we will take the square of each residual, and minimize the sum of these squares. This is where the name **least-squares line** comes from.

For each data point, we calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$; then we square these residuals and add them up. The regression line is the line for which the sum of the residuals,

$$\Sigma (y_i - \hat{y}_i)^2 = \Sigma \varepsilon_i^2$$

is minimized. This expression is called the **Sum of Squared Errors** and denoted as **SSE**.

Using Calculus, we can determine the values of a and b that make the **SSE** a minimum; the resulting line, $\hat{y} = a + bx$, is called the **least squares regression line**. The coefficients are:

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Here, $\bar{x}$ and $\bar{y}$ are the mean of the observed x- and y-values, respectively. Note that the regression line always passes through the point ($\bar{x}$, $\bar{y}$).

The process of fitting a line to a set of points in the plane is called **linear regression**. The least-squares approach described above is widely used, and so it is no surprise that these calculations are programmed into most statistical software packages, as well as into manycalculators. So there is no need to memorize the formulas above; instead, we will use the TI-83/TI-84 calculators:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the regression line $\hat{y} = a + bx$:

1. Go to STAT and then to the EDIT menu. Enter the *x*-values into list L1 and the *y*-values into list L2. Make sure that the order is preserved – i.e. these must remain paired so that the corresponding (*x*,*y*) values are next to each other in the lists.

2. Go to STAT and then to the TESTS menu; scroll down select **LinRegTTest**.

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.

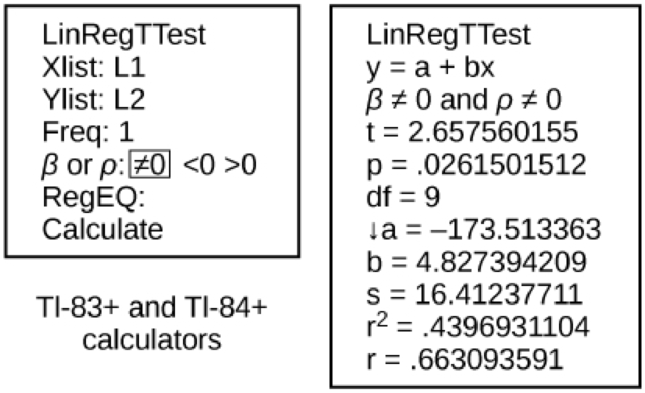
The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items. In particular, at the top of the screen you will see **y = a + bx** which tells us that the intercept is *a* and the slope is *b*.

To illustrate, recall the situation from **Example 12.6**, the "Third Exam/Final Exam" example. There we had the data:

X (third exam score)	65	67	71	71	66	75	67	70	71	69	69
Y (final exam score)	175	133	185	163	126	198	153	163	159	151	159

Following the instructions above, we have the following input and output screens:

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
```

Scroll down to find the values a = -173.513, and b = 4.8273; the equation of the best fit line is

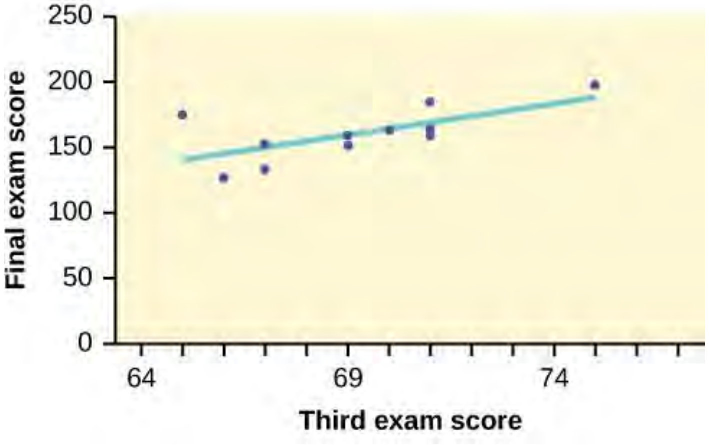
$$\hat{y} = \mathbf{-173.51 + 4.83}x$$

We have already seen how to make a scatterplot for the data; we can graph the regression line on the same screen. To graph the best-fit line, press the "Y=" key and type the equation –173.5 + 4.83X into equation Y1. Press ZOOM 9 again to graph it.

Another way to graph the line after you create a scatter plot is to use LinRegTTest:

1. Make sure you have already made the scatter plot. Check it on your screen.
2. Go to LinRegTTest and enter the lists.
3. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit.
4. Press Y = (you will see the regression equation).
5. Press GRAPH. The line will be drawn.

Following either of these procedures, the graph of the line of best fit for the third-exam/final-exam example is as follows:



Important Note

It is generally a good idea 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 the 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.** The reason for this is that we don't know how the variables are related outside this range; it may be that for other data values the points in the scatterplot no longer have a linear pattern at all, and sharply curve away from the line. So, for values outside the range of observed x-values, the equation will be unreliable at best.

Now that we have the regression line, we want to give a practical interpretation of the slope and intercept.

In general, the intercept *a* is the value of *y* that corresponds to *x* = 0. But in this case, this value does not have any practical meaning – of course, it would not make sense for a final exam score to be negative! Moreover, 0 is very far outside the range of observed *x*-values, so we would not want to use the model to make such a prediction in any case.

In general, the slope *b* is the expected change in *y* that corresponds to a one-unit increase in *x*. For this problem, we have *b* = 4.83, so for each one-point increase in the third exam score, we would expect the final exam to increase by about 4.83 points, on average.

Example 12.7

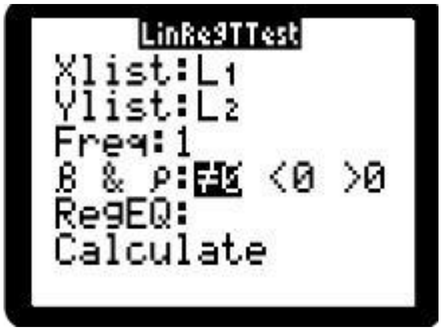
In an area of the Midwest, records were kept on the relationship between the annual rainfall (in inches) and the yield of wheat (bushels per acre).

Rain (inches)	10.5	8.8	13.4	12.5	18.8	10.3	7.0	15.6	16.0
Yield (bushels/acre)	50.5	46.2	58.8	59.0	82.4	49.2	31.9	76.0	78.8

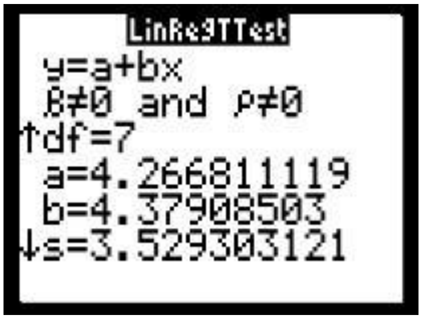
a. Calculate the regression equation.
b. Interpret the slope of the equation in the context of the given question.

Solution 12.7

Go to the STAT menu, and then the EDIT menu. Enter the Temp values into L1, and the Growth values into L2. Go to the STAT menu again, and then to the TESTS menu; scroll down to LinRegTTest; set the alternative to the "≠", scroll to Calculate and press Enter.



```
LinRegTTest
Xlist:L1
Ylist:L2
Freq:1
β & ρ:≠0 <0 >0
RegEQ:
Calculate
```



```
LinRegTTest
y=a+bx
β≠0 and ρ≠0
↑df=7
a=4.266811119
b=4.37908503
↓s=3.529303121
```

a. Scroll down in the output screen to find *a* = 4.26681 and *b* = 4.37908. Thus, the regression equation is: $\hat{y}$ = **4.267 + 4.379*x*.**

b. Slope is the coefficient of x. **Slope = 4.379** this mean that for each additional inch of rain each year, we expect the yield of wheat to increase by about 4.4 bushels/acre on average.

Try It Σ

12.6 The data below are ages and systolic blood pressures of 9 randomly selected adults.

Age	38	41	45	48	51	53	57	61	65
Pressure	116	120	123	131	142	145	148	150	152

a. Calculate the regression equation.
b. Interpret the slope of the equation in the context of the given question.

The Correlation Coefficient *r*

A scatterplot is a valuable visual tool; however, scatterplots can also be misleading. For example, by changing the horizontal or vertical scale, we can make the linear pattern look stronger or weaker than it really is. Other than examining the scatter plot and seeing that a linear model seems reasonable, how else can we tell if there is a linear relationship? We need a numerical measure of the strength of the relationship between *x* and *y*, and that numerical measure is the correlation coefficient.

The **correlation coefficient, *r*,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable *x* and the dependent variable *y*.

The correlation coefficient is calculated as

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

where n = the number of data pairs. We start by listing some important properties of r.

Properties of the Correlation Coefficient:

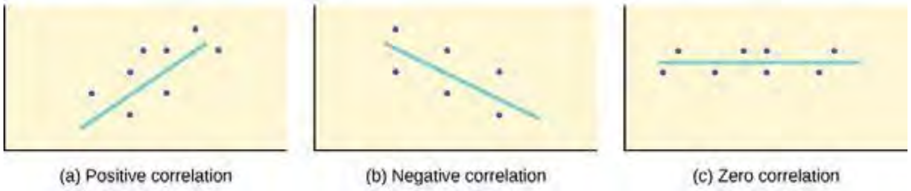
- $-1 \leq r \leq 1$. That is, the value of r is always between –1 and +1.
- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease. So there is a **positive correlation**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase. So there is a **negative correlation**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.

- If $r = 0$ there is absolutely no linear relationship between x and y **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = –1$, there is perfect negative correlation. In both cases, all of the original data points lie on a straight line (this is very rare in the real world)
- r is the same, no matter what units are used to measure x and y; we can easily see that whatever units are used, they will all cancel out of the numerator and denominator.
- r is the same, no matter which variable which of x and y is the independent and which is the dependent variable. I.e. if we switch x and y, we can see that the value will be the same.

NOTE

Even if we have a very strong correlation, this does not suggest that x causes y or y causes x. Remember, **"correlation does not imply causation."**

The figures below illustrate how r shows the direction of the relationship between x and y:



(a) Positive correlation

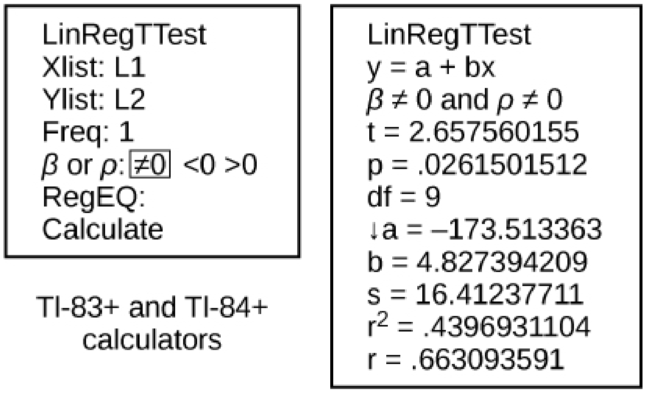
(b) Negative correlation

(c) Zero correlation

(a) A scatter plot showing data with a positive correlation. $0 < r < 1$
(b) A scatter plot showing data with a negative correlation. $–1 < r < 0$
(c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, it is easily calculated using the TI-83 and TI-84 c calculators. In fact, we already know how to do it, because it appears at the bottom of the output screens for LinRegTTest. For example, if we return to the output screen for the Third Exam/Final Exam example, we see that $r = 0.663$. Thus, we would say that there is a moderately strong, positive correlation between scores on the third exam and scores on the final exam.

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
```

The Coefficient of Determination

The statistic r^2 is called the **coefficient of determination**; as the notation suggests, this is just the square of the correlation coefficient. This statistic is usually stated as a percent, rather than in decimal form, and it measures something very specific in the in the context of the data.

When we calculate a regression, some of the variation in the *y*-values is due to variation in the *x*-values. This variation is called **explained variation.** But unless the linear relationship between *x* and *y* is a perfect correlation (very rare!) there will also be variation in the *y*-values that has nothing to do with *x*. These variations could be simple sampling variation, or could be from other factors that influence the variable *y*. This type of variation is called **unexplained variation.**

The coefficient of variation r^2 measures the percentage of variation in the dependent variable *y* that is explained by variation in the independent variable *x*.

This is why r^2 is usually written as a percentage. Note also that this interpretation of r^2 aligns with the properties of the correlation coefficient *r*. If the regression equation is a good fit to the observed data, then we would expect that a large percentage of the variation in *y* would be due to variation in *x*. That is, when there is a strong correlation, we would expect r^2 to be close to 100%. And r^2 is close to 1 exactly when *r* is close to either -1 or 1.

Once we have calculated *r,* finding r^2 is very easy; but the coefficient of determination also appears on the output for LinRegTTest. Again, referring to the Third Exam/Final Exam example, we can read the value of r^2 from the output screen: $r^2 = 0.6631^2 = 0.4397$. Since $.4397 \approx 44\%$, this means that approximately 44% of the variation in the final exam grades can be explained by the variation in the grades on the third exam.

Example 12.8

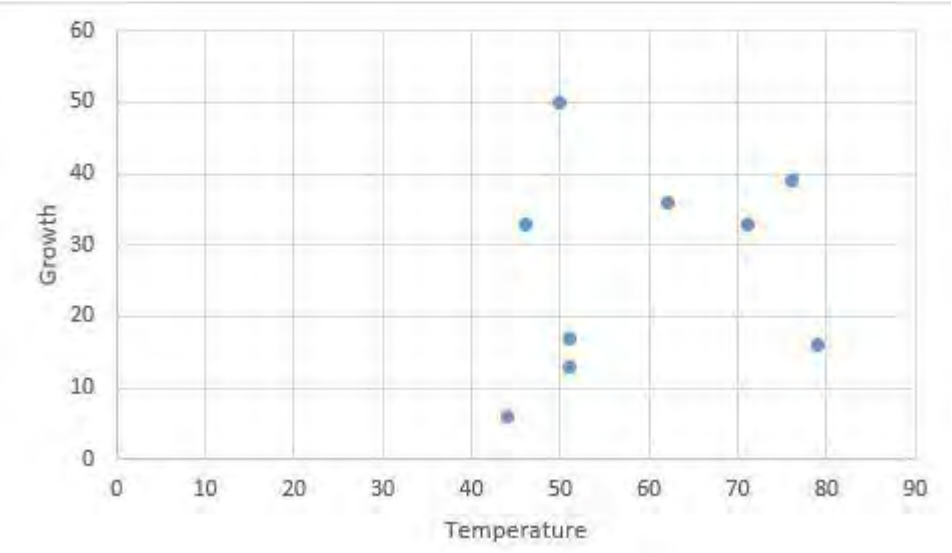
The paired data below consist of the temperatures on randomly chosen days and the amount a certain kind of plant grew (in millimeters):

Temp	62	76	50	51	71	46	51	44	79
Growth	36	39	50	13	33	33	17	6	16

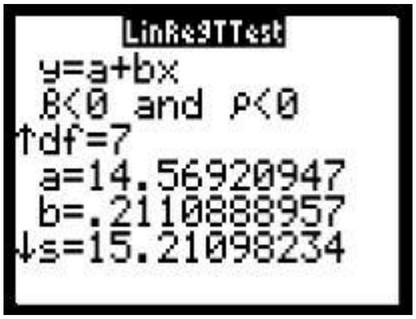
a. Draw a scatterplot
b. Find linear regression equation
c. Find the correlation coefficient *r,* and interpret.
d. Find the coefficient of determination r^2, and interpret.

Solution 12.8

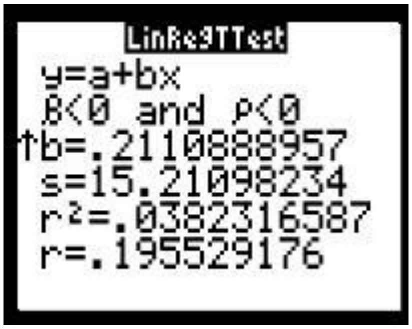
a. Scatterplot



Go to the STAT menu, and then the EDIT menu. Enter the Temp values into L1, and the Growth values into L2. Go to the STAT menu again, and then to the TESTS menu; scroll down to LinRegTTest; set the alternative to the "≠", scroll to Calculate and press Enter.



```
LinRegTTest
 y=a+bx
 β<0 and ρ<0
↑df=7
 a=14.56920947
 b=.2110888957
↓s=15.21098234
```



```
LinRegTTest
 y=a+bx
 β<0 and ρ<0
↑b=.2110888957
 s=15.21098234
 r²=.0382316587
 r=.195529176
```

b. Linear regression equation: $\hat{y} = 14.57 + .211x$

c. The correlation coefficient is at the bottom of the output screens: $r = 0.1955$.
There is a weak, positive correlation between temperature and growth.

d. The coefficient of determination is the second to last value in the output screens: $r^2 = 0.038$.
Only about 3.8% of the total variation in growth of the plant is explained by variation in temperature.

Try It Σ

12.7 The data below are ages and systolic blood pressures of 9 randomly selected adults.

Age	38	41	45	48	51	53	57	61	65
Pressure	116	120	123	131	142	145	148	150	152

a. Find the correlation coefficient *r,* and interpret.
b. Find the coefficient of determination r^2, and interpret.

Finally, we will point out that the correlation coefficient is part of an alternative formula for calculating the slope of the regression line. Let $\bar{x}$ and $\bar{y}$ be the mean of the observed *x*- and *y*-values, respectively, and let s_x and s_y be the standard deviations of the observed *x*- and *y*-values. Then we can calculate the slope using the formula:

$$b = r\frac{s_y}{s_x}.$$

Note that this formula clearly shows one of the properties listed earlier; namely, that the sign of *r* is the same as the sign of the slope, *b*, of the best-fit line.

12.4 | Testing the Significance of the Correlation Coefficient

The correlation coefficient, *r*, tells us about the strength and direction of the linear relationship between *x* and *y*. However, our interpretation of this statistic is inherently subjective. For example, we know that when *r* is "close to 1", this indicates a strong correlation. But this begs the question, *how close*? That is, just how close must *r* be to 1 in order for the correlation to be considered "strong"? Similarly, how close does *r* need to be to 0 to call the correlation "weak"? More importantly, when is the correlation strong enough to warrant using the regression equation for predictions? What we need is an objective criterion; in short, we need a hypothesis test for whether or not the correlation is *statistically significant*.

This procedure is called the test of the **"significance of the correlation coefficient",** and will allow us to quickly and efficiently decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute *r*, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient; this unknown parameter is denoted by ρ, the Greek letter "rho." The sample statistic *r* is a point estimate for ρ.

The hypotheses for the test are:

$$\mathbf{H_0: \rho = 0 \quad vs. \quad H_a: \rho \neq 0}$$

The null hypothesis states that the population correlation coefficient ρ is 0; this means that if we were to calculate the regression line using all data pairs in the population, then the points in the scatterplot would be randomly scattered about a horizontal line. That is, knowing the value of *x* would not be of any use in predicting *y*. So the null hypothesis essentially states that there really is no linear relationship between *x* and *y*.

On the other hand, the alternative hypothesis states that the population correlation coefficient is *significantly different from zero*. I.e., Ha states that there *is* a significant linear relationship between *x* and *y*.

Thus, if we reject H_0, we have evidence that there is a significant correlation between *x* and *y*. We sometimes shorten this by simply saying that *r* is significant.

As described above, the test of significance is two-tailed (because the alternative hypothesis is a "≠" statement). And it is a *t*-test with df = $n - 2$, where *n* is the number of data pairs. The test statistic is:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

To decide the test, we can use either critical values or *p*-values. Once the test statistic is found, we can get the p-values can be found with the TI-83 and TI-84 calculators using the tcdf function (in the DISTR menu). Even better, we can get *both* the test stat and *p*-value using LinRegTTest, and the following instructions:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the regression line $\hat{y} = a + bx$:

1. Go to STAT and then to the EDIT menu. Enter the *x*-values into list L1 and the *y*-values into list L2.
2. Go to STAT and then to the TESTS menu; scroll down select **LinRegTTest**.
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.

The test statistic *t* and *p*-value will appear on the output screen.

Example 12.9

Recall the Third Exam/Final Exam case. We were given data on exam scores:

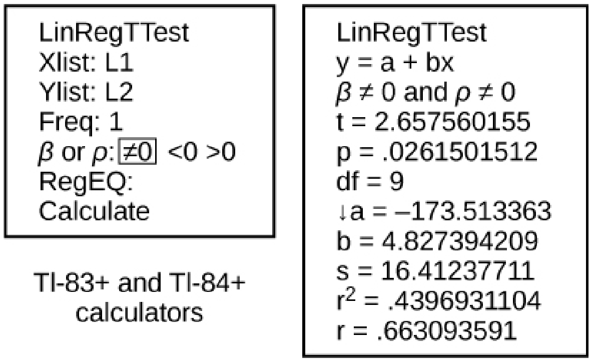
X (third exam score)	65	67	71	71	66	75	67	70	71	69	69
Y (final exam score)	175	133	185	163	126	198	153	163	159	151	159

We have already calculated the regression line for the data: $\boldsymbol{\hat{y} = -173.51 + 4.83x}$, as well as the correlation coefficient $r = 0.663$. Use this information to determine whether the correlation is significant at the 5% significance level.

Solution 12.9

The hypotheses are H_0: $\rho = 0$ and H_a: $\rho \neq 0$, and the significance level is $\alpha = 0.05$. Following the instructions above, we have the following input and output screens:

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
```

From the calculator output, we have $t = 2.66$ and $p = 0.0262$. Since $p < .05$, we **reject** $\mathbf{H_0}$. Thus, there is a significant correlation between the Exam 3 scores and the Final Exam scores. And because r is significant, **the regression line can be used to predict final exam scores.**

In Summary,

- If r is not significant, the regression equation should **not** be used for prediction.
- If r is significant then the regression equation 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 significant, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

The p-value is very easy to obtain using technology such as the TI-84 calculator. However, the calculator requires us to have the raw data available. But there are some cases when we just have the correlation coefficient and want to use it to determine whether the correlation is significant. In these instances we could calculate the test statistic t and then compare it to an appropriate critical value; but there are also tables of critical values for r. That is, we can compare the correlation coefficient r directly to the critical value to make the decision about H_0.

There is a table for critical values for the Pearson correlation coefficient at the end of this chapter. We reproduce an excerpt here for convenience:

Table of 95% Critical Values for Pearson Correlation Coefficient

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

Because this is a two-tailed *t*-test, we will reject H_0 at the $\alpha = .05$ level if either:

$r < -\text{CV}$ **or** $r > \text{CV}$

More concisely, we reject H_0 at the $\alpha = .05$ level if $|\, r \,| > \text{CV}$.

The table again shows how the sample size is also a factor in whether or not a correlation is significant or not.

Consider once more the third exam/final exam example. The line of best fit is $\hat{y} = -173.51 + 4.83x$, $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction?

We must determine if the correlation is significant; the hypotheses are: H_0: $\rho = 0$ vs. Ha: $\rho \neq 0$. But instead of using the *p*-value as before, we will instead use the critical value. We know that $\text{df} = n - 2 = 11 - 2 = 9$.

From the table, the critical values are $\pm$ 0.602; since $0.6631 > 0.602$, we reject H_0 and again conclude that *r* is significant.

There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score and the final exam score. Because *r* is significant, the regression line can be used to predict final exam scores.

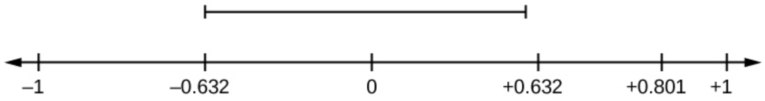
Example 12.10

Suppose you computed the following correlation coefficients using samples of the given size. For each of the following cases, use the table at the end of the chapter to locate the critical value and use it to determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

a. $r = 0.801$, calculated using $n = 10$ data points.
b. $r = -0.624$, calculated using $n = 14$ data points.
c. $r = 0.776$, calculated using $n = 6$ data points.

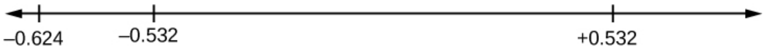
Solution 12.10

a. We computed $r = 0.801$ using $n = 10$ data points. Then df $= n - 2 = 10 - 2 = 8$, and the critical values associated with df $= 8$ are ± 0.632. So we would reject H_0 if either $r < -0.632$ or $r > 0.632$.
Since $r = 0.801$ and $0.801 > 0.632$, we conclude that r is significant and the regression line may be used for prediction. It is helpful to view this on a number line:



$r = 0.801 > +0.632$. Therefore, r is significant.

b. We computed $r = -0.624$ with 14 data points. So df $=12$ and the critical values are ± 0.532. Since $-0.624 < -0.532$, r is significant and the line can be used for prediction.



$r = -0.624 < -0.532$. Therefore, r is significant.

c. We have $r = 0.776$, calculated using $n = 6$ data points. So df $= 4$.
The critical values are ± 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.

Try It Σ

12.9 For each of the following, use critical values to determine whether the correlation is significant. Then state whether or not the regression line can be used for prediction.

a. $r = 0.5204$ using $n = 9$ data points
b. $r = -0.7204$ using $n = 8$ data points
c. $r = 0.708$ and $n = 9$
d. $r = 0.434$ and $n = 14$.

Example 12.11

In an area of the Midwest, records were kept on the relationship between the rainfall (in inches) and the yield of wheat (bushels per acre).

Rain (inches)	10.5	8.8	13.4	12.5	18.8	10.3	7.0	15.6	16.0
Yield (bushels/acre)	50.5	46.2	58.8	59.0	82.4	49.2	31.9	76.0	78.8

a. Test the significance of the correlation between Rain and Yield using the *p*-value.
b. Test the significance of the correlation using the critical value.

Solution 12.11

Go to the STAT menu, and then the EDIT menu. Enter the Temp values into L1, and the Growth values into L2. Go to the STAT menu again, and then to the TESTS menu; scroll down to LinRegTTest; set the alternative to the "≠", scroll to Calculate and press Enter.

a. The test statistic and *p*-value are: $t = 13.31$ and $p = 3.16$ x 10-6 = 0.00000316. Since $p < .05$, we reject H_0. There is a significant relationship between Rain and Yield.

b. The correlation coefficient is $r = .9808$. Using df = 9 – 2 = 7, the critical values are ± 0.666. Since $r > 0.666$, we again reject H_0 and conclude that there is a significant relationship between Rain and Yield.

Try It

12.10 The data below are ages and systolic blood pressures of 9 randomly selected adults.

Age	38	41	45	48	51	53	57	61	65
Pressure	116	120	123	131	142	145	148	150	152

Test the significance of the correlation between Age and Blood Pressure.

NOTE
The table provided in this textbook shows critical only for the significance level of 5%, α = 0.05. But there are other tables available that show critical values for other significance levels as well. Of course, if we are using the *p*-value method, we could test the significance of the correlation using *any* desired significance level; we would not limited to using α = 0.05.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires certain assumptions about the population data. Remember, the purpose of this test is to determine whether the linear relationship that we see between x and y in the sample data provides strong enough evidence to conclude that there is in fact a linear relationship between x and y in the population.

The regression 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.

Assumptions underlying the test of significance are:

(1) There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular x-value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)

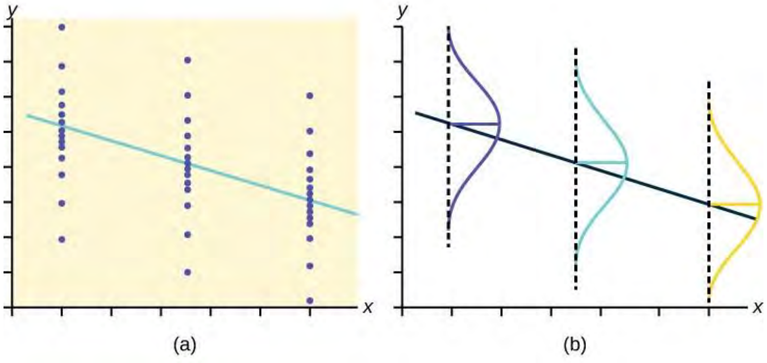
(2) The y values for any particular, *fixed* x-value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) states that these normal distributions are centered on the line; so for a fixed x-value, the mean of the normal distribution of corresponding y-values is the y-value predicted by the regression equation.

(3) The standard deviations of the normal distributions described in (2) are *all equal*. That is, the population of y-values corresponding to a fixed x will have the same standard deviation σ, regardless of the value of x.

(4) The residual errors are mutually independent (no pattern).

(5) The data are produced from a well-designed, random sample or randomized experiment.

Note that in the literature, Assumptions (2) and (3) are usually called the **Normality Assumption** and the **Constant Variance** assumption, respectively. Combining the first three assumptions, we see that no matter what the value of x, the distribution of corresponding y-values will have the same shape and spread about the regression line. This is illustrated in the following figure:



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 corresponding y values lies on the regression line.

Finally, the point estimate for the common standard deviation s in Assumption (3) is called the **standard error** of the estimate. Note that this is just the standard deviation of the residuals. This statistic provides a measure of the spread of y-values about the regression line, and can be calculated using the formula:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

We rarely need to use this formula, however. Like other key statistics related to regression, the standard error s appears on the output screens for LinRegTTest. For example, in the Third Exam/Final Exam example, the standard error is $s = 16.42$.

Test of Significance of Correlation

Hypotheses: **H_0: ρ = 0 vs. H_a: ρ ≠ 0**

The null hypothesis states that there is no correlation between x and y.
The alternative hypothesis states that there is a significant correlation between x and y.

Test statistic: $t = r\sqrt{\frac{n-2}{1-r^2}}$ Calculator Function: **LinRegTTest**

If we reject H_0, then we have evidence that there is a significant correlation between x and y.

12.5 | Prediction

Recall the **Third Exam/Final Exam** example. We examined the scatterplot and showed that there is a significant correlation between the Exam 3 scores and the Final Exam scores (i.e. the correlation coefficient is significant). We also found the equation of the best-fit line for the final exam grade as a function of the grade on the third exam: $\hat{y} = –173.51 + 4.83x$.

We can now use the least-squares regression line for prediction. For example, suppose we wanted to predict the mean final exam score of statistics students who received 73 on the third exam. Note that the sample exam scores (x-values) range from 65 to 75. Since $x = 73$ is in the range of observed x-values 65 and 75, and the correlation is significant, we can use the regression equation. So we substitute $x = 73$ into the equation to get:

$$\hat{y} = –173.51 + 4.83(73) = 179.08$$

Of course, this does not mean that any student who scores 73 on the third exam is guaranteed to score 179.08 on final exam. There will be other factors that influence a student's score on the final exam (e.g. the number of hours spent studying for the exam). What we know is that in the population, among all students who scored 73 on the third exam, the *average* score on the final exam for is the y-value on the regression line. So when we say that 179.08 is the predicted score, this means:

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 12.12

Recall the **Third Exam/Final Exam example** from Example 12.5. We have shown that the correlation is significant, and the regression equation is $\hat{y} = –173.51 + 4.83x$.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?
b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution 12.12

a. $\hat{y} = –173.51 + 4.83(66) =$ **145.27.**

b. The observed x-values in the sample data are between 65 and 75. Ninety is well outside of the domain of these observed x-values, so we cannot reliably predict the final exam score for this student.

Of course, we can easily enter 90 into the equation for x and calculate the corresponding y-value, but the predicted value will not be reliable. To illustrate how unreliable the prediction can be for x-values outside the range of observed data, make the substitution $x = 90$ into the equation:

$$y = –173.51 + 4.84(90) = 261.19$$

The final-exam score is predicted to be 261.19, which is not even possible - the largest the final-exam score can be is 200.

NOTE

The process of predicting inside of the observed x values observed in the data is called **interpolation**. The process of predicting outside of the observed x values observed in the data is called **extrapolation**. In general, extrapolation should be done with extreme caution, if at all.

Try It Σ

12.11 Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Important notes:

- If r is not significant, the regression equation should **not** be used for prediction. Use the mean of the observed y values for prediction.
- If r is significant then the regression equation 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 significant, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

12.6 | Outliers

In some data sets, there are values (observed data points) called **outliers**. Outliers are observed data points that are far from the least squares line. They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the 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 errors above or below the best-fit line as an outlier**.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

Example 12.13

In the **third exam/final exam example**, use the graphical method to determine if there is an outlier or not. If there is an outlier, we will delete it and fit the remaining data to a new line; the new line ought to fit the remaining data better. This means the SSE and standard error should both be smaller and the correlation coefficient ought to be closer to ± 1.

Solution 12.13

With the TI-83, 83+, 84+ graphing calculators, it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to 2*s* or more, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are considered outliers. We will call these lines Y2 and Y3:

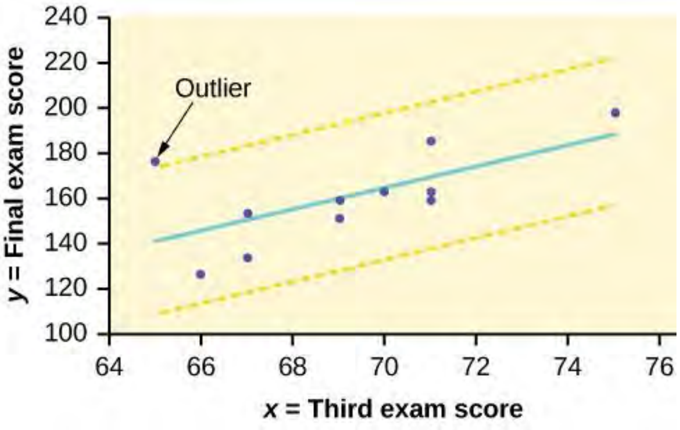
Recall that the regression equation for this example is $\hat{y} = -173.51 + 4.83x$, and the standard error is ***s* = 16.4**. (This statistic appears on the output screen for LinRegTTest)

We want to draw two lines that are parallel to the regression line, and such that the vertical distance to the regression line is 2*s* at each point; to do this, we add and subtract $2s = 2(16.4)$ to the equation for the regression line:

$$Y2 = -173.5 + 4.83x - 2(16.4)$$

and $$Y3 = -173.5 + 4.83x + 2(16.4)$$

Graph the scatterplot with the best fit line in equation Y1, as before.
Then enter the two extra lines as Y2 and Y3 in the "Y="equation editor and press ZOOM 9.



We see that the only data point that is not between lines Y2 and Y3 is the point $x = 65$, $y = 175$. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than two standard deviations away from the best-fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines; for this example, the point is just barely outside the parallel lines. On a computer, enlarging the graph may help; on the calculator screen it can help to zoom in. If we are in doubt, we just need to numerically verify that the point is an outlier.

In this case, we first calculate the predicted final exam score for a student who scores 65 on the third exam:

$$\hat{y} = -173.51 + 4.83(65) = 140.44.$$

Next we calculate the difference between the observed and predicted values (i.e. the residual):

$$y - \hat{y} = 175 - 140.44 = 34.56.$$

This residual really is more than $2s = 32.8$, which verifies that the observation is an outlier.

Next we use a numerical method to identify the outlier.

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:

Residual = observed y value − predicted y value = $y - \hat{y}$.

Recall that the standard error s is the standard deviation of the residuals, which is $s = 16.4$.

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	35
67	133	150	-17
71	185	169	16
71	163	169	-6
66	126	145	-19
75	198	189	9
67	153	150	3
70	163	164	-1
71	159	164	-10
69	151	160	-9
69	159	160	-1

We are looking for all data points for which the residual is greater than $2s = 2(16.4) = 32.8$ or less than -32.8. Compare these values to the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

NOTE: This method involves a lot of calculation, but it is easily automated. In fact, most statistical packages (such as Minitab or SPSS) will automatically calculate the residuals and flag data points that appear to be outliers.

Now that we have identified the point (65, 175) as an outlier, we would 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 if not delete the data. If the data is correct, we would leave it in the data set. For this example, let's suppose that after reviewing the data, we found that this data pair data really was an error and should be removed.

We remove the data point, and compute a new best-fit line and correlation coefficient using the ten remaining points; using the calculator as before, we obtain the new regression equation:

$$\hat{y} = -355.19 + 7.39x$$

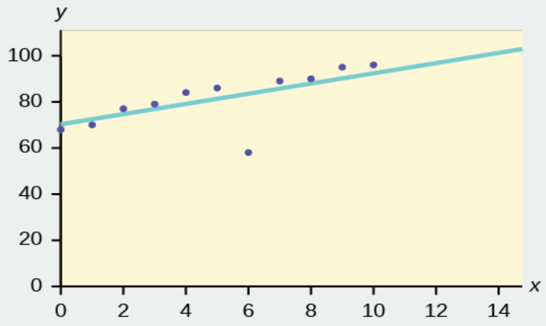
Moreover, the new model has correlation coefficient $r = 0.9121$. Recall that for the original data, we had $r = 0.6631$, so the new line exhibits a stronger correlation. This means that the new line is a better fit to the ten remaining data values. Moreover, the new standard error is $s = 9.275$, a significant reduction. Both of these tell us that the new regression line can better predict the final exam score given the third exam score.

NOTE

When outliers are deleted, the researcher should either record that data was deleted, and why, or should provide results both with and without the deleted data. If data is erroneous and the correct values are known (e.g., student one actually scored a 70 instead of a 65), then this correction can be made to the data.

Try It Σ

12.12 Identify the potential outlier in the scatter plot. The standard deviation of the residuals is approximately 8.6.



Example 12.14

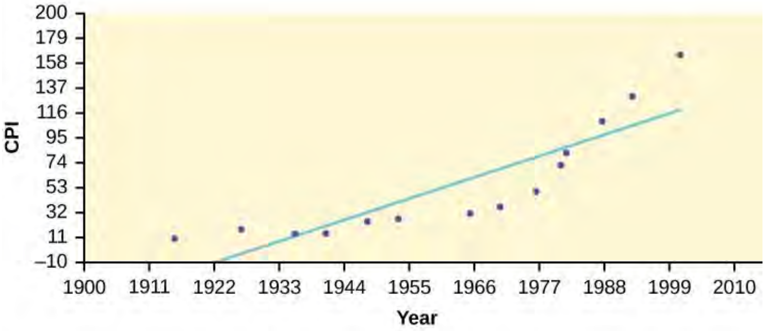
The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price 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.

x	y	x	y
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

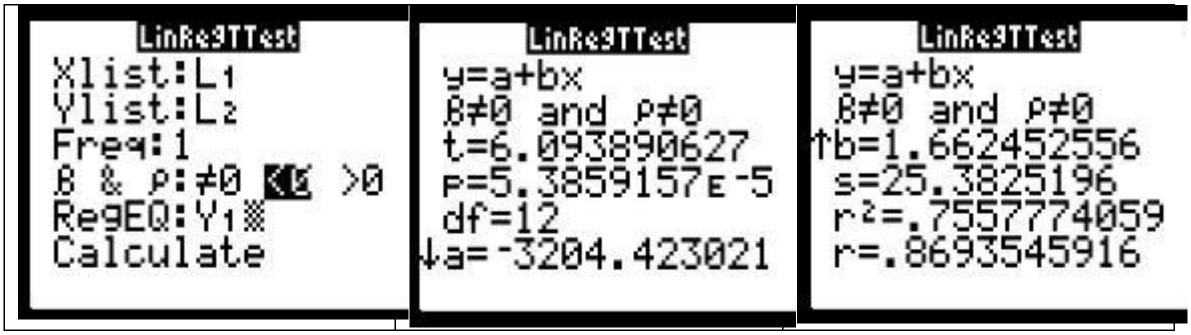
a. Draw a scatterplot of the data.
b. Calculate the least squares line; write the equation in the form $\hat{y} = a + bx$.
c. Graph the line on the same axes as the scatterplot.
d. Find the correlation coefficient. Is it significant?
e. What was the average CPI for the year 1990?
f. List the residuals and find any outliers.

Solution 12.14

Parts a and c:



Go to STAT, and then the EDIT menu. Enter the *x*-values into L1 and the *y*-values into L2. Go to STAT, and then the TESTS menu; scroll down to **LinRegTTest**. Specify the lists, select the "≠" alternative, place the cursor on Calculate and press Enter to get the output screens.



```
LinRegTTest
Xlist:L1
Ylist:L2
Freq:1
β & ρ:≠0 <0 >0
RegEQ:Y1
Calculate
```

```
LinRegTTest
y=a+bx
β≠0 and ρ≠0
t=6.093890627
p=5.3859157E-5
df=12
↓a=-3204.423021
```

```
LinRegTTest
y=a+bx
β≠0 and ρ≠0
↑b=1.662452556
s=25.3825196
r²=.7557774059
r=.8693545916
```

NOTE: To enter Y1 in RegEQ, press VARS, choose Y-Vars, choose Function Enter, Y1 enter

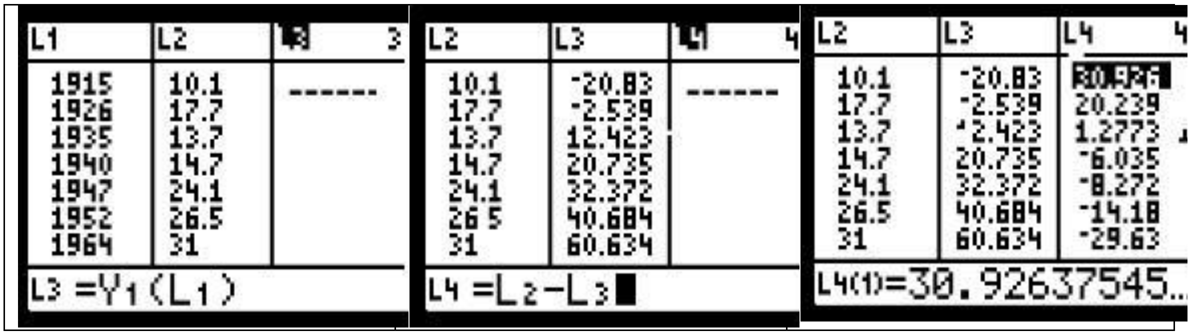
b. The least-squares regression line is: $\boldsymbol{\hat{y} = –3204 + 1.662x}$.

c. The correlation coefficient is $r = 0.8694$.

d. We can get the test statistic and p-value from the output screen: $t = 6.09$ and $p = 0.000054$. Since $p < .05$, the correlation is significant.
Alternatively, there are $n = 14$ data points, so df = 12; the corresponding critical value is 0.532. Since 0.8694 > 0.532, r is significant.

e. $\hat{y} = –3204 + 1.662(1990) = 103.4$ CPI

f. To list residuals using your calculator, you must first place Y1 in RegEQ as seen in the picture above AND make sure you Calculate. Then go back into your List, highlight L3, Press Enter, Press Vars, Function, Y-vars, Y1. Then Press (2nd L1), Press Enter. L3 will be filled with the predicted values. Highlight L4, Press Enter, L2 – L3, Press Enter. L4 will be filled with the residuals.



```
L1      L2      L3     3 | L2      L3      L4     4 | L2      L3      L4     4
1915    10.1    ------   | 10.1    ⁻20.83  ------   | 10.1    ⁻20.83  30.926
1926    17.7             | 17.7    ⁻2.539           | 17.7    ⁻2.539  20.239
1935    13.7             | 13.7    12.423           | 13.7    ⁻2.423  1.2773
1940    14.7             | 14.7    20.735           | 14.7    20.735  ⁻6.035
1947    24.1             | 24.1    32.372           | 24.1    32.372  ⁻8.272
1952    26.5             | 26 5    40.684           | 26.5    40.684  ⁻14.18
1964    31               | 31      60.634           | 31      60.634  ⁻29.63
L3 =Y1(L1)               | L4 =L2-L3                | L4(1)=30.92637545…
```

NOTE

In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve might be a more appropriate model to use than a line. If we were to try and fit a non-linear function to the data, we might obtain a more accurate model.

Try It Σ

12.14 The following table shows economic development measured in per capita income PCINC.

Year	PCINC	Year	PCINC
1870	340	1920	1050
1880	499	1930	1170
1890	592	1940	1364
1900	757	1950	1836
1910	927	1960	2132

a. What are the independent and dependent variables?

b. Draw a scatter plot.

c. Use regression to find the line of best fit and the correlation coefficient.

d. Is there a significant linear relationship between the variables? Explain.

e. Interpret the slope of the regression equation.

f. Use the line of best fit to estimate PCINC for the years 1925 and for 2000.

KEY TERMS and FORMULA REVIEW

Correlation Coefficient: A measure developed by Karl Pearson (early 1900s) that measures the strength and direction of linear relationship between two variables x and y. It is given by the formula:

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

Coefficient of determination: The square of the correlation coefficient, which is r^2. The coefficient of determination measures the percentage of variation in the dependent variable that is explained by variation in the dependent variable.

Constant Variance Assumption: This assumption states that for any fixed x-value, the population of corresponding y-values is normally has the same variance.

Least Squares Regression Line: This is the equation for the line of best fit to a set of n data points (x, y). The equation is $\hat{y} = a + bx$, where:

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

Here, $\bar{x}$ and $\bar{y}$ are the mean of the observed x- and y-values, respectively; the regression line always passes through the point ($\bar{x}$, $\bar{y}$).

Normality Assumption: This assumption states that for any fixed x-value, the population of corresponding y-values is normally distributed.

Outlier: An observation that does not fit the rest of the data. For regression, this is an observed data point that is located at a vertical distance more than $2s$ from the regression line, where s is the standard error of the estimate.

Residual. The difference $y - \hat{y}$ between an observed y-value and the y-value predicted by the regression line.

Scatterplot: A graph of the ordered pairs (x, y) of observed data, where x is the independent variable and y is the dependent variable.

Standard error of the estimate: This is the standard deviation of the residuals; it is also a point estimate for the common standard deviation from the Constant Variance assumption. This statistic provides a measure of the spread of y-values about the regression line, and is calculated using:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

CHAPTER REVIEW

12.2 Scatter Plots

If we have measurements on two quantitative variables, x and y for a sample of individuals in a population, then we can use a scatterplot to make an initial determination as to whether there is a linear relationship between the variables. In particular, a scatterplot can tell us the direction of the relationship between the x variables and the y variables, and provide a rough idea of the strength of the relationship.

12.3 The Regression Equation

The procedure of fitting a linear equation to a collection of observed data points is called *linear regression.*

The regression line is the "line of best fit" to the data. To understand what this means, we calculate the residuals, the differences $y - \hat{y}$ between the observed y-values and the y-values predicted by line. The residuals measure the vertical distances between the observed data points and the corresponding points on the regression line.

The sum of the squares of the residuals, $\Sigma(y - \hat{y})^2$, is called the SSE, or Sum of Squared Errors. The regression line is found by minimizing the SSE and so is also called *the least-squares line.* This line, usually denoted as $\hat{y} = a + bx$, is easily calculated on the TI-84 using LinRegTTest.

12.4 Testing the Significance of the Correlation Coefficient

The correlation coefficient r is a numerical measure of the strength and direction of the linear association between x and y:

- The correlation coefficient r is always between -1 and $+1$.
- When r is positive, the slope of the regression line is positive.
  When r is negative, the slope of the regression line is also negative.
- Values of r that are close to 0 indicate a weak correlation.
- Values of r that are close to $\underline{+}$ 1 indicate a strong correlation.

The value of r also shows on the output screens for LinRegTTest.
These properties are intuitive and easy to understand, but are somewhat subjective. So we also introduced a test to determine whether or not a correlation between two variables is *significant*. This test involves the correlation parameter ρ, and the hypotheses are:

$$H_0: \rho = 0 \quad \text{vs.} \quad H_a: \rho \neq 0$$

This is a two-tailed t-test with test statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$. We can get both the test statistic and p-value from the calculator using **LinRegTTest**. We also can make the decision about Ho by comparing r directly to an appropriate critical value. More importantly

Rejecting H_0 at the .05 significance level provides evidence that the correlation is significant.

There are several important assumptions that we make when calculating a regression equation, and judging how well it fits the data:

1. There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular x-value lies on a straight line in the population.

2. The y-values corresponding to a *fixed* x-value are normally distributed about the line. Assumption (1) states that these normal distributions are centered on the regression line, so for a fixed x-value, the mean of the normal distribution of corresponding y-values is the y-value predicted by the regression equation.

3. The standard deviations of the normal distributions described in (2) are *all equal*. That is, the population of y-values corresponding to a fixed x will have the same standard deviation σ, regardless of the value of x. The point estimate for this common standard deviation is called the **standard error of the estimate**; it is the standard deviation of the residuals, and so provides a measure of the spread of y-values about the regression line. The standard error of the estimate also appears on the output screens for LinRegTTest.

4. The residual errors are mutually independent (no pattern).

5. The data are obtained from a well-designed, random sample or randomized experiment.

12.5 Prediction

Once we determine that there is a significant correlation between x and y, we can use the least squares regression line to make predictions; that is, given a value of the independent variable x, we can use the equation $\hat{y} = a + bx$ to find the corresponding y-value. Some things to remember:

- If r is not significant, the regression equation should **not** be used for prediction.
- If r is significant then the regression equation 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 significant, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

12.6 Outliers

An outlier is a point that lies well outside the pattern of the observed data; specifically, it is an observed data point that is located at a vertical distance more than $2s$ from the regression line, where s is the standard error of the estimate. We showed both a graphical method and a numerical method for identifying outliers; see Example 12.12.

Table of 95% Critical Values for Pearson Correlation Coefficient

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	0.195

Exercises for Chapter 12

1. A specialty cleaning company charges an equipment fee and an hourly labor fee. The total amount of the fee the company charges for each session is given by the equation $y = 50 + 100x$.

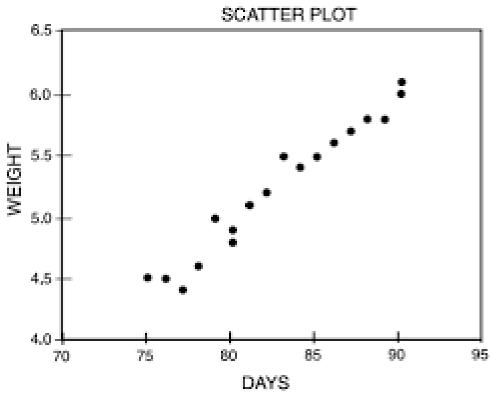
a. What are the independent and dependent variables?
b. What is the y-intercept? Interpret in the context of the problem.
c. What is the slope? Interpret in the context of the problem.

2. The price of a single issue of stock can fluctuate throughout the day. The price of stock for Shipment Express is $y = 15 - 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

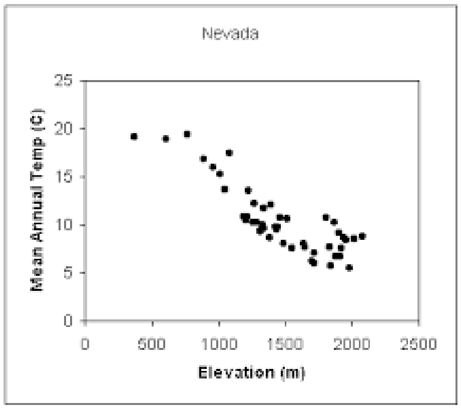
a. What are the slope and y-intercept? Interpret their meaning.
b. If you owned this stock, would you want a positive or negative slope? Why?

3. For each of the scatterplots, answer the following:
i. Does there appear to be a linear relationship between the variables?
ii. Is the linear relationship strong or weak?
iii. Is the relationdship positive or negative?

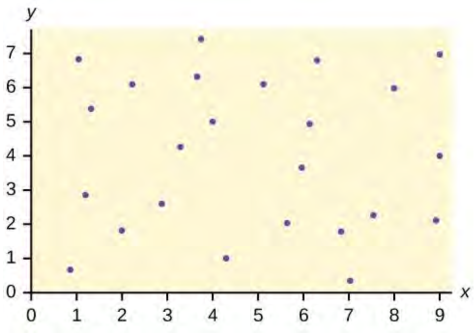
a.



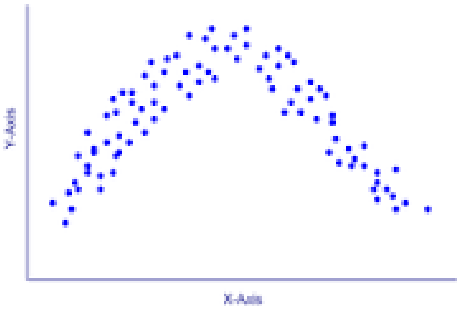
b.



c.



d.



4. What does an *r* value of zero mean?

5. A random sample of ten professional athletes produced the following data where *x* is the number of endorsements the player has and *y* is the amount of money made (in millions of dollars).

x	0	3	2	1	5	5	4	3	0	4
Y	2	8	7	3	13	12	9	9	3	10

a. Draw a scatter plot of the data.
b. Use regression to find the equation for the line of best fit.
c. Draw the line of best fit on the scatter plot.
d. What is the slope of the line of best fit? Interpret.
e. What is the *y*-intercept of the line of best fit? What does it represent?

6. When testing the significance of the correlation coefficient,
a. What is the null hypothesis?
b. What is the alternative hypothesis?

7. When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

8. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:
$\hat{y} = 1350 - 1.2x$ where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

a. How many acres will be left to mow after 20 hours of work.
b. How many acres will be left to mow after 100 hours of work?
c. How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

9. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where *x* is the day. The model can be written as follows $\hat{y} = 101.32 + 2.48x$ where $\hat{y}$ is in thousands of dollars.

a. What would you predict the sales to be on day 60?
b. What would you predict the sales to be on day 90?

10. The following table shows cell phone usage in the U.S. for the years 2003-2009:

Year	**Cellular Usage** (%)
2003	54.67
2005	74.19
2007	84.86
2009	90.82

a. Construct a scatterplot.
b. Find the line of best-fit.
c. Find the correlation coefficient.
d. Would the regression line provide a reliable estimate for cell phone usage in 2016? Explain.

11. Find the regression equation and correlation coefficient for the following data:

x	16.3	39.3	42.6	10.8	30.6
y	7	10	6	7	4

12. Find the regression equation and correlation coefficient for the following data:

x	57	53	59	61	53	56	60
y	156	164	163	177	159	175	151

13. The following data shows the number of years spent studying a foreign language and the score on a test of fluency.

Number of years (x)	Score (y)
3	57
4	78
4	72
2	58
5	89
3	63
4	73
5	84
3	75
2	48

Find the regression equation and correlation coefficient for the data.

14. In an area of the Midwest, records were kept on the relationship between the rainfall (in inches) and the yield of wheat (bushels per acre). The data is shown below:

Rain (inches)	10.5	8.8	13.4	12.5	18.8	10.3	7.0	15.6	16.0
Yield (bushels/acre)	50.5	46.2	58.8	59.0	82.4	49.2	31.9	76.0	78.8

Use this data to determine whether there is a significant correlation between the amount of rainfall and yield of wheat.

15. The data below are the final exam scores of 10 randomly selected statistics students and the number of hours they studied for the exam.

Hours	5	10	4	6	10	9
Score	64	86	69	86	59	87

Use this data to determine whether there is a significant correlation between the number of hours studied and the exam score.

The following data shows data for the first two decades of AIDS reporting. Use this data to answer questions #16-23.

Year	#AIDS Cases Diagnosed
1981	319
1982	1170
1983	3076
1984	6240
1985	11,776
1986	19,032
1987	28,564
1988	35,447
1989	42,674
1990	48,634
1991	59,660
1992	78,530
1993	78,834
1994	71,874
1995	68,505
1996	59,347
1997	47,149
1998	38,393
1999	25,174
2000	25,522
2001	25,643
2002	26,464

16. Create a scatterplot for the data, using YEAR as the independent variable.

17. Find the regression equation. Round the coefficients to whole numbers.

18. Identify each of the following from the regression output:

$r =$ _______ $r^2 =$ _______ $s =$ _______

19. Is there a significant correlation between the year and the number of AIDS cases diagnosed? Explain.

20. Interpret the slope of the regression line.

21. Calculate the following point predictions:

a. When $x = 1985$, $\hat{y} =$

b. When $x = 1990$, $\hat{y} =$

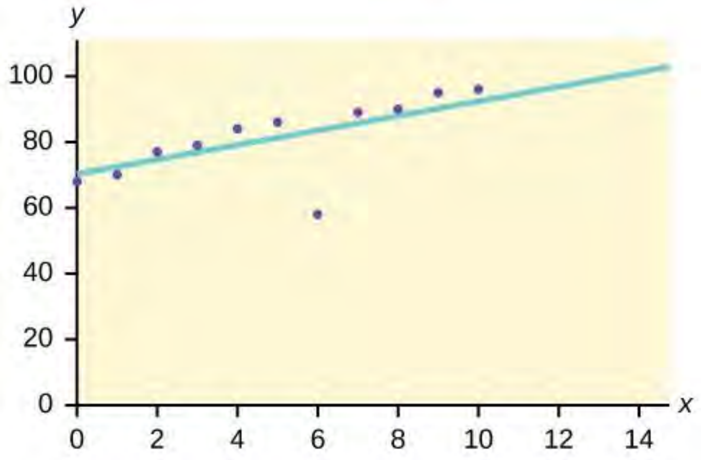
22. Would it be appropriate to use the regression line to estimate the number of diagnosed AIDS cases for the current year?

23. Does the line seem to fit the data? Why or why not?
Does your answer affect how you would use the regression equation for estimation?

24. a. The SSE (Sum of Squared Errors) for a data set of 18 numbers is 49. What is the standard error of the estimate?

b. The standard error for the estimate a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

25. The following scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.



a. Do there appear to be any outliers? If so, describe the point.

b. A point is removed, and the line of best fit is recalculated. The new correlation coefficient is *r* = 0.98. Does the point appear to have been an outlier? Why or why not?

c. What effect did the potential outlier have on the line of best fit?

d. Are you more or less confident in the predictive ability of the new line of best fit?

26. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The following table shows the GDP PPP of Cuba as compared to US dollars.

Year	Cuba's PPP
1999	1700
2000	1700
2001	2300
2002	2900
2003	3000
2004	3500
2005	4000
2006	11,000
2008	9500
2009	9700
2010	9900

a. Construct a scatter plot of the data.

b. Find the least-squares regression line

c. Find the correlation coefficient.

27. Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs.

School	**Mid-Career Salary (in thousands)**	**Yearly Tuition**
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

a. Construct a scatter plot of the data, using tuition as the independent variable.
b. Find the correlation coefficient *r*. Is there a significant correlation?
c. From the scatterplot, does there appear to be any outliers?
d. After removing the outlier(s), is there a significant correlation?

28. a. Explain what it means when a correlation has an r^2 of 0.72.
b. Can a coefficient of determination be negative? Why or why not?

29. Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	**Midpt .**	**Number of Driver Deaths (per 100,000)**
16-19	17.5	38
20-24	22	36
25-34	29.5	24
35-54	44.5	20
55-74 .	64.5	18
75+	80	28

For each age group, pick the midpoint of the interval for the *x* value; for the 75+ group, use 80.

a. Using “ages” as the independent variable, calculate the least squares (best–fit) line.
b. Find the correlation coefficient. Is it significant?
c. Based on this data, is there a linear relationship between age of a driver and driver fatality rate?

30. The table below shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Which variable should be the independent and which should be the dependent variable?
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line; put the equation in the form $\hat{y} = a + bx.$
d. Find the correlation coefficient. Is the correlation significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in the table?
g. Should we use the least squares line to find the estimated life expectancy for an individual born in 1850? Why or why not?
h. Interpret the slope of the regression line.

31. The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given below:

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

a. Decide which variable should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Should the regression equation be used to find the maximum value for the restaurants on p.70?

32. The table below gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	**Time** **(seconds)**
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

a. Decide which variable should be the dependent variable.

b. Draw a scatter plot of the data.

c. Does it appear from inspection that there is a relationship between the variables? Why or why not?

d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.

e. Find the correlation coefficient. Is the decrease in times significant?

f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.

g. Why are the answers from part f different from the chart values?

h. Does it appear that a line is the best way to fit the data? Why or why not?

i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Is this estimate reliable? Why or why not?

33. Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return:	74	66	81	52	73	62	52	45	62	46	60	46	38
Percent new:	5	6	8	11	12	15	16	17	18	18	19	20	20

a. Enter the data into your calculator and make a scatter plot.

b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.

c. Interpret the slope and intercept of the regression line in the context of the problem.

c. How well does the regression line fit the data? Explain your answer.

d. Which point has the largest residual? Is this point an outlier? An influential point? Explain.

e. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

34. The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). The data is shown below:

Height (in feet)	**Stories**
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

a. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.

b. Does it appear from inspection that there is a relationship between the variables?

c. Calculate the least squares line. Put the equation in the form: $\hat{y} = a + bx$

d. Find the correlation coefficient. Is it significant?

e. Find the estimated heights for a 32 story building.

f. Find the height for a 94 story building.

g. Based on this data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?

h. Suppose we wanted to estimate the height of a building with six stories? Would the least squares line give an accurate estimate of height? Explain why or why not.

i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building

35. The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	**Heart Rate**
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

a. Enter the data into your calculator and make a scatter plot.
b. Find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Interpret the slope and y-intercept of the regression line.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

36. The following table shows data on average per capita wine consumption (in liters) and deaths from heart disease (per 100,000 residents) in a random sample of 10 countries:

Consumption	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Deaths	221	167	131	191	220	297	71	172	211	300

a. Enter the data into your calculator and make a scatter plot.
b. Find the regression equation and add to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Do the data provide convincing evidence that there is a linear relationship between the amount of wine consumed and the heart disease death rate? Carry out an appropriate test at a the .05 significance level to justify your answer.

37. The percentage of female wage and salary workers who are paid hourly rates for the years 1979 to 1992 are given in the following table:

Year	**% of workers paid hourly rates**
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

a. Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Explain.
c. Calculate the least-squares line. Put the equation in the form: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?

e. Find the estimated percentages for 1991 and 1988.
f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
g. Are there any outliers in the data?
h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

38. The average number of people in a family that received welfare for various years is given in the following table:

Year	**Welfare family size**
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

a. Using "year" as the independent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is the correlation significant?
d. Based on this data, is there a linear relationship between the year and the average number of people in a welfare family?
e. Using the least-squares line, estimate the welfare family size for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.

39. The following are advertised sale prices of color televisions at a major appliance store.

Size (inches)	**Sale Price** ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

Suppose that we wish to use the size of the television to predict the selling price.

a. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
b. Find the correlation coefficient. Is there a significant correlation between size and price?
c. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
d. Interpret the slope of the regression line.

40. According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	**Approx. Probate Fees and Taxes** ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

Suppose that we wish to investigate the relationship between these two variables.

a. Which would be the independent variable? Which would be the dependent variable?
b. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Find the estimated total cost for a next taxable estate of $1,000,000; why does this value differ from the value shown in the table?
e. Interpret the slope of the least-squares regression line.

f. List the residuals for each x value.

41. The following table shows data for the average heights of American boys of different ages. Suppose that we want to use a boy's age to predict his height.

Age (years)	**Height** (cm)
Birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Which variable should be the independent variable? Which should be the dependent variable?
b. Calculate the least-squares line. Put the equation in the form: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is the correlation significant? Explain.
d. Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
e. Interpret the slope of the least-squares line.

f. List the residuals for each x value.

REFERENCES

12.1 Linear Equations

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

12.5 Prediction

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Data from the United States Census Bureau. Available online at http://www.census.gov/compendia/statab/cats/ transportation/motor_vehicle_accidents_and_fatalities.html

Data from the National Center for Health Statistics.

12.6 Outliers

Data from the House Ways and Means Committee, the Health and Human Services Department. Data from Microsoft Bookshelf.

Data from the United States Department of Labor, the Bureau of Labor Statistics. Data from the Physician's Handbook, 1990.

Data from the United States Department of Labor, the Bureau of Labor Statistics.

13 | F DISTRIBUTION AND ONE-WAY ANOVA

plant
utrals bruschetta 101 personal palettes tulips gone wild
31 32 33 34 35 36 37 38
DE GRANDES IDI
EST PIZZAS → SINGLES → HARRY WEESE → REAL ESTATE
ure stars of Salone | Hob satisfaction | Retail directory | Hot new houses | Time 2
E MAGAZINE OF ARCHITECTURE AND DESIGN

Figure 13.1 One-way ANOVA is used to measure information from several groups.

Introduction

Chapter Objectives

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

- Interpret the F probability distribution as the number of groups and the sample size change.
- Discuss two uses for the F distribution: one-way ANOVA and the test of two variances.
- Conduct and interpret one-way ANOVA.
- Conduct and interpret hypothesis tests of two variances.

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

13.1 | Test of Two Variances/Standard Deviations

One use 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 an 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.

Unlike most other tests in this book, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher p-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here. We will again be using the 5 steps of hypothesis testing.

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:

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

F has the distribution similar to chi-square distribution but dependent on two degrees of freedom.

The F distribution has the following characteristics:

1.) All F values are greater than or equal to 0

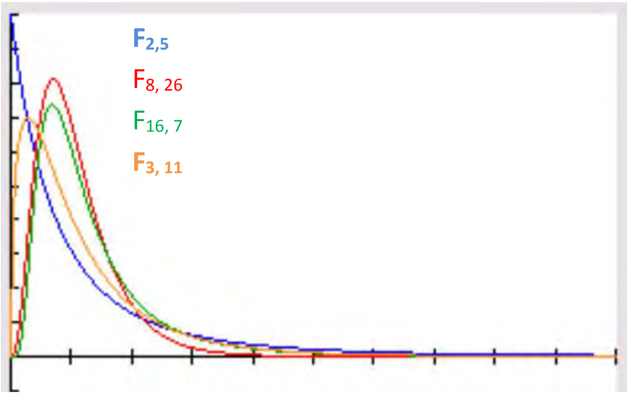
2.) There is a different F curve for each pair of degrees of freedom $n_1 - 1$, $n_2 - 1$ (Figure below).

3.) The curve is nonsymmetrical and skewed to the right

4.) There is 100% under the curve

5.) The notation is $F \sim F(n_1 - 1, n_2 - 1)$ where $n_1 - 1$, are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

6.) $\mu = \dfrac{d.f.N.}{d.f.D. - 1}$



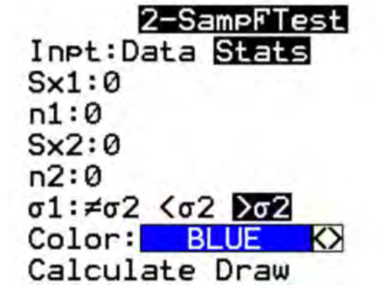
Since the null hypothesis will have equality $\sigma_1^2 = \sigma_2^2$ (or $\sigma_1^2 \geq \sigma_2^2$ or $\sigma_1^2 \leq \sigma_2^2$) then the F Ratio becomes

$$F = \frac{s_1^2}{s_2^2}$$

Using the TI-83, 83+, 84, 84+ Calculator

To calculate F test value

Press STAT, TESTS, 2SampFTest

 2-SampFTest Inpt:Data Stats Sx1:0 n1:0 Sx2:0 n2:0 σ1:≠σ2 <σ2 >σ2 Color: BLUE Calculate Draw	NOTE: When using the 2-SampFtest, standard deviation is requested instead of variance. Make sure you know which is given in the problem.

Critical Values of the F-distribution

Recall, critical value is a value of the distribution that separates the confidence area from the non-confidence area. We will be using the F- distribution table from Statistics Online Computational Resource site (http://www.socr.ucla.edu/Applets.dir/F_Table.html) to find the critical values. The same rules apply in the previous distributions (Z, t, and χ^2) when it applies to alpha (α): if the test is two-tailed you will split α.

To find F_R, the right critical value:

1. Scroll to the correct α page
2. Look up the degrees of freedom of numerator (d.f.1) at the top of the chart and look up the degrees of freedom of the denominator (d.f.2) on the side of the chart.
3. Round to the nearest degrees of freedom if necessary.

To find F_L, the left critical value:

1. Switch the degrees of freedom: d.f.N. will now be d.f.D. and vice versa
2. Look up the new d.f.N. at the top of the chart and the new d.f.D on the side of the chart
3. $F_L = \dfrac{1}{\textit{chart value}}$

Example 13.1

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

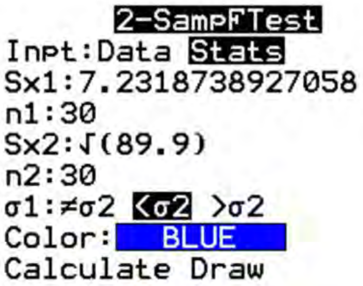
Solution 13.1 using p-value

Claim: $\sigma_1^2 < \sigma_2^2$ (first instructor variance is smaller than 2nd instructor's variance).

H_0: $\sigma_1^2 \geq \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

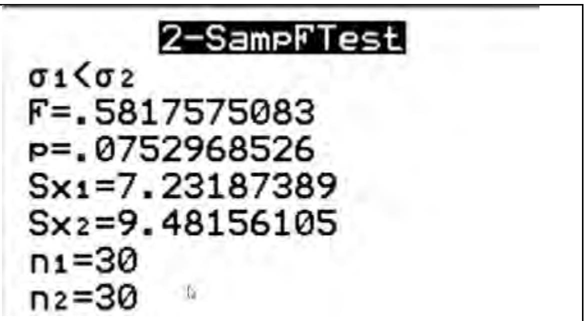
$\alpha = 0.05$; left-tailed test

F = 2sampFtest = .5818



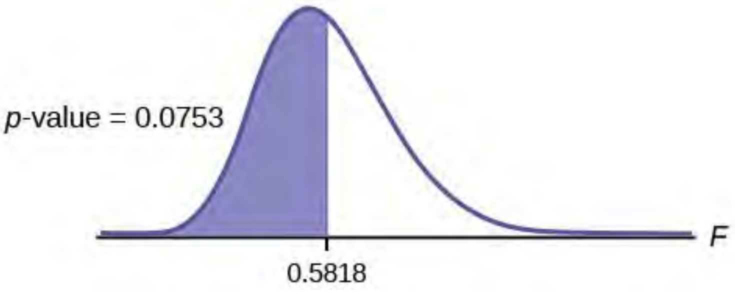
```
   2-SampFTest
 Inpt:Data Stats
 Sx1:7.2318738927058
 n1:30
 Sx2:√(89.9)
 n2:30
 σ1:≠σ2 <σ2 >σ2
 Color: BLUE
 Calculate Draw
```

NOTE: *you can enter the square root within the 2-SampFTest function*



```
        2-SampFTest
 σ1<σ2
 F=.5817575083
 p=.0752968526
 Sx1=7.23187389
 Sx2=9.48156105
 n1=30
 n2=30
```

Draw the graph labeling and shading appropriately.



p-value = P(F < 0.5818) = 0.0753

Since p-value ≥ α, Do Not Reject H_0.

With a 5% level of significance, from the data, there is not sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Solution 13.1 using critical value

Claim: $\sigma_1^2 < \sigma_2^2$ (first instructor variance is smaller than 2nd instructor's variance).

H_0: $\sigma_1^2 \geq \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

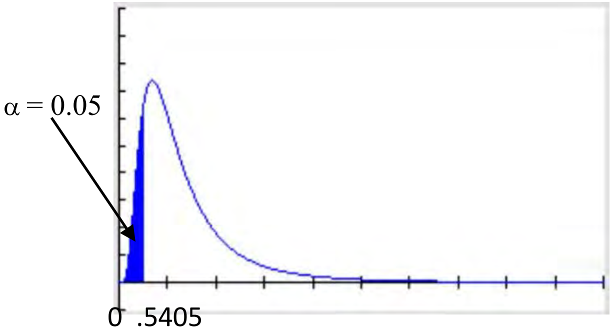
α = 0.05; left-tailed test

F = 2sampFtest = 0.5818

Distribution for the test is $F_{29,29}$ where d.f.N. = $n_1 - 1 = 29$ and d.f.D. = $n_2 - 1 = 29$.

Since it is left-tailed, we are looking for the left critical value, $F_{L.}$ Here, when we switch up the degrees of freedom, it looks the same since in this example d.f.N. = d.f.D. On the chart, there is no 29 at the top of the chart; therefore, we go to the closest degrees of freedom which is 30. However, on the degrees of freedom on the side of the chart, we use 29 since there is one.

$$F_L = \frac{1}{\textit{chart value}} = \frac{1}{1.85} = .5405$$



Since the test statistic, 0.5818 is not in the critical region (0.5818 > 0.6173), Do Not Reject H_0.

With a 5% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Example 13.2

The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, and Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the standard deviations of the heights of singers in each of these two groups (Tenor1 and Bass2) are different using 10% level of significance? Use the table below.

Tenor 1:	Bass 2:
69, 72, 71, 66, 76, 74, 71, 66, 68, 67, 70, 65, 72, 70, 68, 64, 73, 66, 68, 67, 64	72, 75, 67, 75, 74, 72, 72, 74, 72, 72, 74, 70, 66, 68, 75, 68, 70, 72, 67, 70, 70, 69, 72, 71, 74, 75

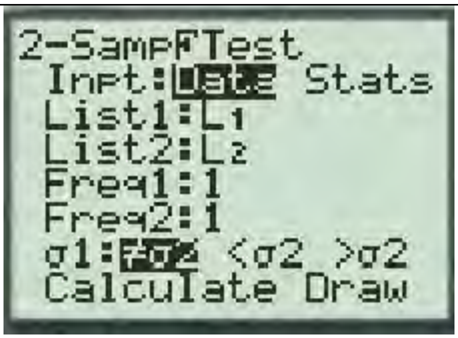
Solution 13.2 using p-value

H₀: $\sigma_1 = \sigma_2$

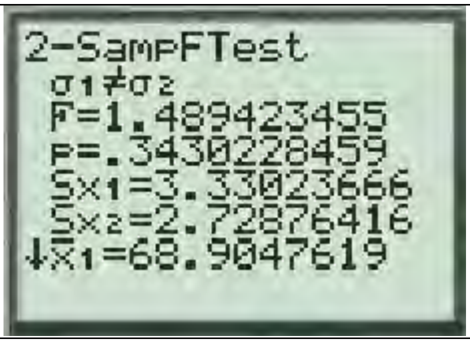
Hₐ: $\sigma_1 \neq \sigma_2$ (claim)

$\alpha = 0.10$; two-tailed test

F = 2-SampFTest = 1.489



```
2-SampFTest
 Inpt:Data Stats
 List1:L1
 List2:L2
 Freq1:1
 Freq2:1
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```



```
2-SampFTest
 σ1≠σ2
 F=1.489423455
 p=.3430228459
 Sx1=3.33023666
 Sx2=2.72876416
↓x̄1=68.9047619
```

p-value = .343; therefore, Do Not Reject H₀.

There is not enough evidence at 10% level of significance to support the claim standard deviations of the heights of singers in each of these two groups (Tenor1 and Bass2) are different.

Solution 13.2 using critical value

H₀: $\sigma_1 = \sigma_2$

Hₐ: $\sigma_1 \neq \sigma_2$ (claim)

$\alpha = 0.10$; two-tailed test

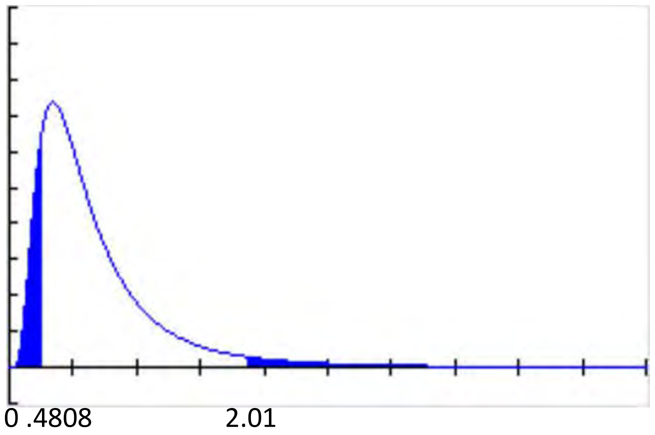
F = 2-SampFTest = 1.489

Since it is two-tailed, we split α. We look up the critical values using the 0.05 chart of the F-distribution chart (http://www.socr.ucla.edu/Applets.dir/F_Table.html).

Distribution for the test is $F_{20, 25}$ where d.f.N. = $n_1 - 1 = 20$ and d.f.D. = $n_2 - 1 = 25$.

$F_R = 2.01$

$F_L = \frac{1}{\textit{chart value}} = \frac{1}{2.08} = .4808$ NOTE: d.f.N. and d.f.D. switch for left critical value.



Do Not Reject H_0 because the test statistic, 1.489 is not in neither tail.

There is not enough evidence at 10% level of significance to support the claim standard deviations of the heights of singers in each of these two groups (Tenor1 and Bass2) are different.

Try It

The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, and Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have enough evidence that the standard deviation of Tenor 1 heights is larger than the standard deviation of the heights of Bass 2 singers using 5% level of significance? Use the table below.

Tenor 1:	Bass 2:
69, 72, 71, 66, 76, 74, 71, 66, 68, 67, 70, 65, 72, 70, 68, 64, 73, 66, 68, 67, 64	72, 75, 67, 75, 74, 72, 72, 74, 72, 72, 74, 70, 66, 68, 75, 68, 70, 72, 67, 70, 70, 69, 72, 71, 74, 75

13.2 | One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. There must be at least three groups. The test actually uses variances to help determine if the means are equal or not. In order to perform a one- way ANOVA test, there are five basic assumptions to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).
4. The factor is a categorical variable.
5. The response is a numerical variable.

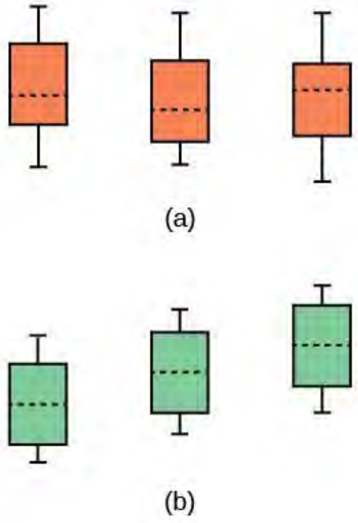
The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k independent samples:

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
H_a: At least one of the means is different

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), H_0: $\mu 1 = \mu 2 = \mu 3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).



NOTE: (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

To calculate the F ratio for the ANOVA test value, two estimates of the variance are made.

1. **Variance between samples**: An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called variation due to treatment (factor) or explained variation. This variance is also known as **Mean Square Between** (MS_B).

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. This variance is known as **Mean Square Within** (MS_W).

Test Statistic (Value): $F = \dfrac{MS_B}{MS_W}$

Always **right-tailed test**

$SS_{between}$ (SS_B) = the sum of squares that represents the variation *among* the different samples

SS_{within} (SS_W) = the sum of squares that represents the variation *within* samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in Descriptive Statistics.

Calculation of Sum of Squares:

- k = the number of independent samples (groups)
- n_j = the size of the jth group
- s_j = the sum of the values in the jth group
- N = total sample size: $\sum n_j$
- x = one value $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Total Sum of Squares: $SS_{total} = \sum x^2 - \dfrac{\left(\sum x^2\right)}{n}$
- Explained variation: $SS_{between} = \sum \left[\dfrac{\left(s_j\right)^2}{n_j}\right] - \dfrac{\left(\sum s_j\right)^2}{n}$
- Unexplained variation: $SS_{within} = SS_{total} - SS_{between}$

Calculation of Mean Square:

- Degrees of freedom of numerator: d.f.N. = k – 1
- Degrees of freedom of denominator: d.f.D. = N – k
- Mean Square Between: $\mathrm{MS_B} = \dfrac{\mathrm{SS_B}}{\mathrm{d.f.N}}$
- Mean Square Within: $\mathrm{MS_W} = \dfrac{\mathrm{SS_W}}{\mathrm{d.f.D}}$

The one-way ANOVA test depends on the fact that $\mathrm{MS_B}$ can be influenced by population differences among means of the several groups. Since $\mathrm{MS_W}$ compares values of each group to its own group mean, the fact that group means might be different does not affect $\mathrm{MS_W}$.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least one of the means of the sample groups come from populations with different normal distributions. If the null hypothesis is true, $\mathrm{MS_B}$ and $\mathrm{MS_W}$ should both estimate the same value.

> NOTE: 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.

Test Statistic (Value):

$$\mathrm{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$$

If $\mathrm{MS_B}$ and $\mathrm{MS_W}$ estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, $\mathrm{MS_B}$ consists of the population variance plus a variance produced from the differences between the samples. $\mathrm{MS_W}$ is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, $\mathrm{MS_B}$ will generally be larger than $\mathrm{MS_W}$. Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that $\mathrm{MS_W}$ will be larger in a given sample.

ANOVA Summary Table

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	$\mathrm{SS_B}$	d.f.N = k - 1	$\mathrm{MS_B} = \dfrac{\mathrm{SS_B}}{\mathrm{d.f.N}}$	$\mathrm{F} = \dfrac{\mathrm{MS_B}}{\mathrm{MS_W}}$	Fcdf by calculator
Within (Error)	$\mathrm{SS_W}$	d.f.D = N – k	$\mathrm{MS_W} = \dfrac{\mathrm{SS_W}}{\mathrm{d.f.D}}$		
Total		N - 1			

Using the TI-83, 83+, 84, 84+ Calculator

To calculate p-value of F-distribution

p-value = Fcdf(Test Value, 10^99, d.f.N., d.f.D.)

Example 13.3

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in the summary table:

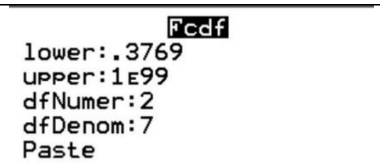
Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	**2.2458**	**2**			
Within (Error)	**20.8542**	**7**			
Total					

a. Complete the ANOVA summary table

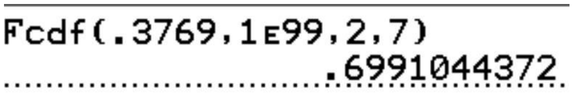
b. Using 5% level of significance, test the claim that the mean weight loss for the three different plans are the same.

Solution 13.3 part a

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	**2.2458**	**2**	$MS_B = \frac{2.2458}{2}$ $MS_B = 1.1229$	$F = \frac{1.1229}{2.9792}$ $F = .3769$	.6991
Within (Error)	**20.8542**	**7**	$MS_W = \frac{20.8542}{7}$ $MS_W = 2.9792$		
Total	23.1	9			



```
Fcdf
lower:.3769
upper:1E99
dfNumer:2
dfDenom:7
Paste
```



```
Fcdf(.3769,1E99,2,7)
......................6991044372
```

Solution 13.3 part b

H$_0$: $\mu_1 = \mu_2 = \mu_3$ (claim)
H$_a$: At least one of the means is different

$\alpha = 0.05$; right-tailed test

$$F = \frac{MS_B}{MS_W} = .3769$$

p-value = .6991; therefore Do Not Reject H$_0$

There is not sufficient evidence at 5% level of significance to reject the claim that the mean weight loss for the three different plans are the same.

Using the TI-83, 83+, 84, 84+ Calculator
To calculate test statistic given the data

If we are given the data of 3 or more groups, we can use the List feature of the calculator to enter the data. We can also use the ANOVA(L1, L2, L3, …) function to find the test statistic.

Example 13.4

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in the following table:

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

a. Using a significance level of 1%, is there a difference in the mean grades among the sororities?
b. Create an ANOVA summary table.

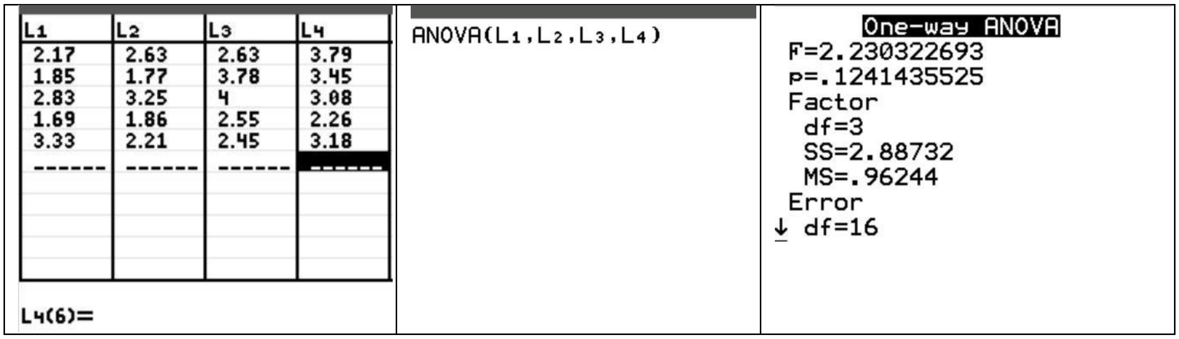
Solution 13.4 part a

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: At least one of the means is different (claim)

$\alpha = 0.01$; right-tailed test

F = ANOVA(L1, L2, L3, L4) = 2.23

p-value = 0.124; therefore Do Not Reject H_0 since p-value $\geq \alpha$.



```
L1     L2     L3     L4
2.17   2.63   2.63   3.79
1.85   1.77   3.78   3.45
2.83   3.25   4      3.08
1.69   1.86   2.55   2.26
3.33   2.21   2.45   3.18
------ ------ ------ ------

L4(6)=
```

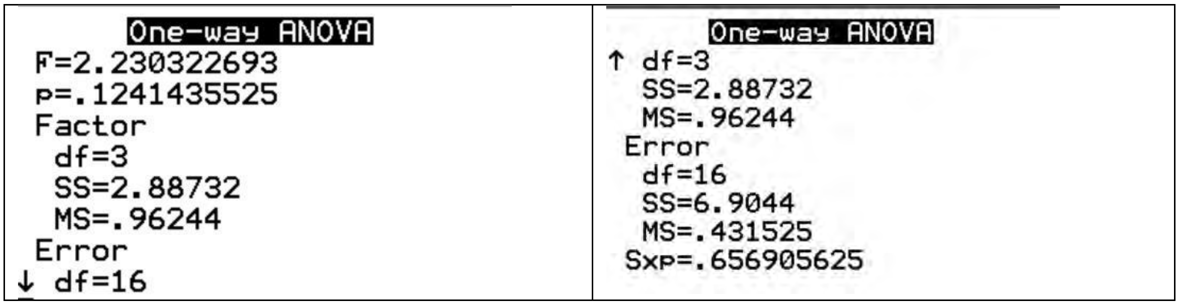
```
ANOVA(L1,L2,L3,L4)
```

```
   One-way ANOVA
 F=2.230322693
 p=.1241435525
 Factor
   df=3
   SS=2.88732
   MS=.96244
 Error
↓  df=16
```

There is not sufficient evidence at 5% level of significance to support the claim that the mean weight loss for the three different plans are the same.

Solution 13.4 part b

The calculator will help us fill in the ANOVA Summary table



```
   One-way ANOVA
 F=2.230322693
 p=.1241435525
 Factor
   df=3
   SS=2.88732
   MS=.96244
 Error
↓  df=16
```

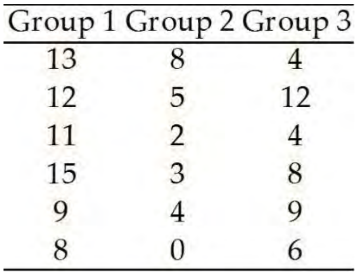
```
   One-way ANOVA
↑  df=3
   SS=2.88732
   MS=.96244
 Error
   df=16
   SS=6.9044
   MS=.431525
 Sxp=.656905625
```

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	2.88732	3	.96244	F = 2.23	0.124
Within (Error)	6.9044	16	.431525		
Total	9.7916	19			

To recap, $MS_B = .96244$ is an estimate of σ^2 that is based on the variability among the sample means. $MS_W = .431525$ is an estimate of σ^2 that is based on the sample variances.

Example 13.5

A medical researcher wishes to try three different techniques to lower blood pressure of patients with high blood pressure. The subjects are randomly selected and assigned to one of three groups. Group 1 is given medication, Group 2 is given an exercise program, and Group 3 is assigned a diet program. At the end of six weeks, each subject's blood pressure is recorded, with data as shown below. Using p-value and the level of significance of 1%, test the claim that the mean is the same for the different groups.



Group 1	Group 2	Group 3
13	8	4
12	5	12
11	2	4
15	3	8
9	4	9
8	0	6

Solution 13.5

H_0: $\mu_1 = \mu_2 = \mu_3$ (claim); H_a: At least one of the means is different

$\alpha = 0.01$; right-tailed test

F = ANOVA(L1, L2, L3) = 11.09; p-value = 0.0011

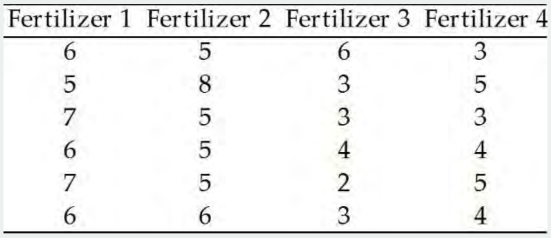
Reject H0 because $0.0011 < \alpha$

There is enough evidence at 1% level of significance to reject the claim that the mean is the same for the different groups.

Try It Σ

Four different types of fertilizers are used on raspberry plants. The number of raspberries on each randomly selected plant is given below.

A suitable hypothesis test will be conducted to test the claim that the type of fertilizer makes no difference in the mean number of raspberries per plant. Use $\alpha = .05$ to test the claim.



Fertilizer 1	Fertilizer 2	Fertilizer 3	Fertilizer 4
6	5	6	3
5	8	3	5
7	5	3	3
6	5	4	4
7	5	2	5
6	6	3	4

Key Terms

Analysis of Variance also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population. The test statistic for analysis of variance is the F-ratio.

Review of Tests

You have seen the F test statistic used in two different circumstances. The following bulleted list is a summary that will help you decide which F test is the appropriate one to use.

- **Testing 2 Variances**:

  Test statistic: $F = \frac{s_1^2}{s_2^2}$ (calculator: 2-SampFTest)

  Degrees of freedom of numerator: d.f.N. = $n_1 - 1$

  Degrees of freedom of denominator: d.f.D. = $n_2 - 1$

  Test may be left-, right-, or two-tailed

- **Testing 3 or more Means:**

  Test statistic (value): F = $\frac{\text{MS}_\text{B}}{\text{MS}_\text{W}}$ (calculator: ANOVA(L_1, L_2, …L_k))

  P-value: Fcdf(Test Value, 10^99, d.f.N., d.f.D.)

  Mean Square Between: $\text{MS}_\text{B} = \frac{\text{SS}_\text{B}}{\text{d.f.N}}$

  Mean Square Within: $\text{MS}_\text{W} = \frac{\text{SS}_\text{W}}{\text{d.f.D}}$

  Total Sum of Squares: $SS_{total} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$

  Explained variation: $SS_B = \sum\left[\frac{\left(s_j\right)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$

  Unexplained variation: $SS_W = SS_{total} - SS_B$

  Degrees of freedom of numerator: d.f.N. = k – 1

  Degrees of freedom of denominator: d.f.D. = N – k

Exercises for Chapter 13

1. State the two assumptions that must be true in order to perform an *F*- test of two variances.

2. Suppose that we test the hypotheses H$_0$: $\sigma_1^2 = \sigma_2^2$ vs. Ha: $\sigma_1^2 \neq \sigma_2^2$ using the following data:

$s_1^2 = 9.4$ $s_2^2 = 10.6$
$n_1 = 12$ $n_2 = 11$

Find the test statistic and *p*-value for the test.

3. Suppose that we test the hypotheses H$_0$: $\sigma_1^2 = \sigma_2^2$ vs. Ha: $\sigma_1^2 \neq \sigma_2^2$ using the following data:

$s_1^2 = 22.5$ $s_2^2 = 31.3$
$n_1 = 19$ $n_2 = 23$

Find the test statistic and *p*-value for the test.

4. Two coworkers commute from the same building. They are interested in comparing the variability of their driving times to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker claims that his commuting time is less variable than that of his colleague. Test this claim at the 10% level.

a. State the null and alternative hypotheses.
b**.** What test should be used in this problem?
c. What is the *F* statistic?
d. What is the *p*-value?
e. Using $\alpha = .10$, what is the decision about H$_0$?
f. Does this data support the first worker's claim?

5. Two students are interested in comparing the amount of variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are less variable. Assuming that the populations of test scores are normally distributed, we will test her claim at the $\alpha = .05$ significance level.

a. State the null and alternative hypotheses.
b**.** What test should be used in this problem?
c. What are the *F* statistic and *p*-value?
d. At the 5% significance level, would we reject the null hypothesis?
e. What does this say about the student's claim?

6. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8 and the second cyclist has a variance of 32.1. Use this data along with a .05 significance level to test the claim that there is a difference in the variances.

7. List the five basic assumptions that must be fulfilled in order to perform a one-way ANOVA test.

8. State the hypotheses for a one-way ANOVA test if there are three groups.

9. State the hypotheses for a one-way ANOVA test if there are four groups.

10. Groups of men from three different areas of the country are to be tested for mean weight. The entries in the table are the weights for the different groups. The one-way ANOVA results are shown in the table:

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

a. What is the Sum of Squares Factor?
b. What is the Sum of Squares Error?
c. What is the degrees of freedom for the numerator?
d. What is the degrees of freedom for the denominator?
e. What is the Mean Square Factor?
f. What is the Mean Square Error?
g. What is the *F* statistic?

11. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in the table are the goals per game for the different teams. The one-way ANOVA results are shown in the table below:

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

a. What is SS_{between}?
b. What is the degrees of freedom for the numerator?
c. What is MS_{between}?
d. What is SS_{within}?
e. What is the degrees of freedom for the denominator?
f. What is MS_{within}?
g. What is the *F* statistic?
h. Judging by the *F* statistic, do you think it is likely that you will reject the null hypothesis?

12. Five basketball teams each took a random sample of players regarding how high each player can jump (in inches). The results are shown in the table below:

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

.

a. What are the Sum of Squares and Mean Squares Factors?
b. What are the Sum of Squares and Mean Squares Errors?
c. What is the *F* statistic?
d. What is the *p*-value?
e. At the 5% significance level, is there a difference in the mean jump heights among the teams?

13. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in the table:

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

This data is used to test whether there is a significant difference among the means for each group.

a. What are the hypotheses for the test?
b. What are the SS_{between} and MS_{between}?|
c. What are the SS_{within} and MS_{within}?
d**.** What is the *F* Statistic?
e**.** What is the *p*-value?
f. At the 10% significance level, is there evidence that the mean scores among the different groups are different?

14. Three different traffic routes are tested for mean driving time. The entries in the table are the driving times in minutes on the three different routes. The data from three random samples are shown in the table below:

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

Use this data to test the claim that the routes have the same mean driving time.

15. Three students, Linda, Tuan, and Javier, are each given five laboratory rats for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10%, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's rats	**Tuan's rats**	**Javier's rats**
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

16. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in the table below. Using a 5% significance level, test the hypothesis that the three mean commuting mileages are the same.

Working Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

17. The table below lists the number of pages in four different types of magazines. Use this data and a 5% level of significance to test the hypothesis that the four magazine types have the same mean number of pages.

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

18. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. The table below shows the results of a study.

CNN	FOX	LOCAL
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a .05 level of significance.

19. Are the means for the final exams the same for all statistics class delivery types? The table below shows the scores on final exams from randomly selected classes using the different delivery types.

Online	Hybrid	Face-to-face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

Assume that all distributions are normal, the population standard deviations are approximately the same, and the data were collected independently and randomly. Use a .05 level of significance.

20. Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics and Asians? Suppose that the table below shows the results of a study:

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Use this data, along with a .05 significance level, to test the claim that the mean is the same for all four populations. Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly.

21. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that the data below shows the results of a study:

Powder	Machine Made	Hard Packed
1210	2107	2846
1080	1149	1638
1537	862	2019
941	1870	1178
	1528	2233
	1382	

Use this data and a .05 significance level to test the claim that the mean number of visitors is the same for all three types of snow conditions. Assume that all distributions are normal, the three population standard deviations are approximately the same, and the data were collected independently and randomly.

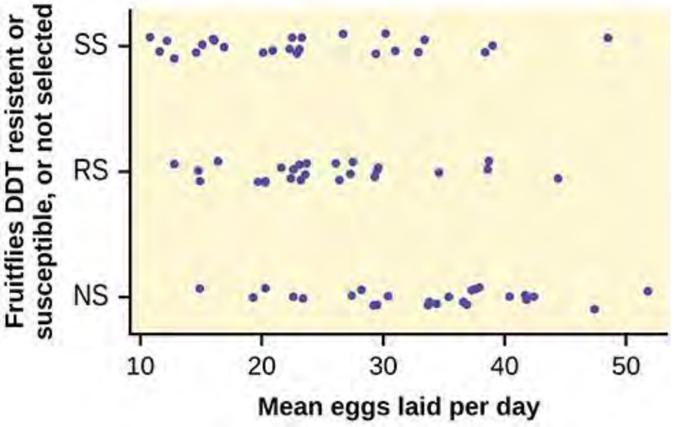
22. DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective, but persisted in the environment and over time became seen as harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). Finally there was a control line of non-selected or typical fruitflies (NS). Here are the data:

RS	**SS**	**NS**	**RS**	**SS**	**NS**
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
16.4	26.7	28.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
14.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1% level of significance, are the mean rates of egg selection for the three strains of fruitfly different? If so, in what way? Specifically, the researchers were interested in whether or not the selectively bred strains were different from the non-selected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:



23. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's rats	**Tuan's rats**	**Javier's rats**
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Determine whether or not there is a significant difference in the variance among Javier's and Linda's rats. Test at a significance level of 10%.

24. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

Working Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Determine whether or not the variance in mileage driven is statistically the same among the working class and professional (middle income) groups. Use a 5% significance level.

25. The following table lists the number of pages in four different types of magazines:

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Use this data to test whether there is a significant difference in the variance for home decorating magazines and news magazines.

26. Is the variance for the amount of money (in dollars) that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that the table shows the results of a study.

Saturday	**Sunday**	**Saturday**	**Sunday**
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Use this data, along with a 5% significance level to test the claim that the variances are equal.

27. Are the variances for incomes on the East Coast and the West Coast the same? Suppose that the table below shows the results of a study; income is shown in thousands of dollars. Use this data to test the hypothesis that the variances are the same. Assume that both distributions are normal, and the samples are randomly and independently selected.

East	**West**
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

28. Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of ten, with each receiving one of three doses of caffeine: 0 mg, 100 mg, 200 mg. This is approximately the amount in no, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?

Here are the data:

0 mg	**100 mg**	**200 mg**	**0 mg**	**100 mg**	**200 mg**
242	248	246	245	246	248
244	245	250	248	247	252
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

29. King Manuel I, Komnenus ruled the Byzantine Empire from Constantinople (Istanbul) during the years 1145 to 1180 A.D. The empire was very powerful during his reign, but declined significantly afterwards. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from a fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	**Second Coinage**	**Third Coinage**	**Fourth Coinage**
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

Did the silver content of the coins change over the course of Manuel's reign? Here are the means and variances of each coinage. The data are unbalanced.

	First	**Second**	**Third**	**Fourth**
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

30.

Four different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in the summary table:

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	**60.37**	**3**			
Within (Error)	**101.81**	**53**			
Total					

a. Complete the ANOVA summary table
b. Using 5% level of significance, test the claim that the mean weight loss for the four different plans are the same.

31.

We are interested in testing the mean compression strength of four different box types. The one-way ANOVA results are shown in the summary table:

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	**32.138**	**3**			
Within (Error)	**32.901**				
Total		**23**			

a. Complete the ANOVA summary table
b. Using 1% level of significance, test the claim that the mean compression strength for the four different box types is different.

REFERENCES

13.1 Test of Two Variances

“MLB Vs. Division Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true.

Tomato Data, Marist College School of Science (unpublished student research)

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. A Handbook of Small Datasets: Data for Fruitfly

13.2 ANOVA

Data from a fourth grade classroom in 1994 in a private K – 12 school in San Jose, CA.

Fecundity. London: Chapman & Hall, 1994.

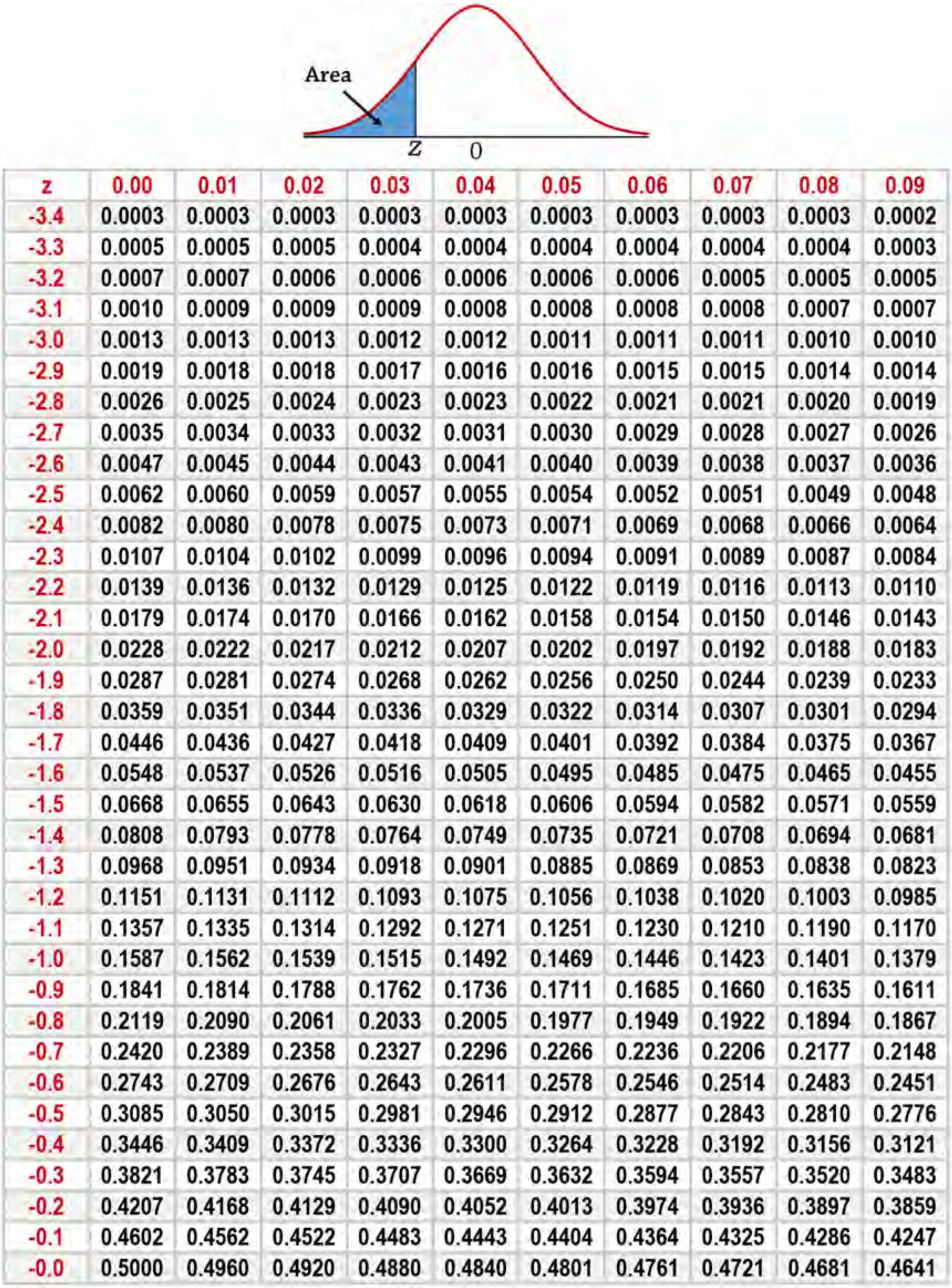
Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets.* London: Chapman & Hall, 1994, pg. 50.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. A Handbook of Small Datasets. London: Chapman & Hall, 1994, pg. 118.

“MLB Standings – 2012.” Available online at http://espn.go.com/mlb/standings/_/year/2012.

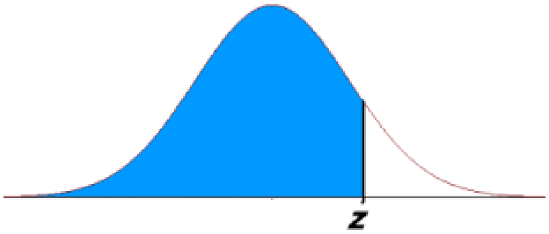
Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

NORMAL DISTRIBUTION TABLES



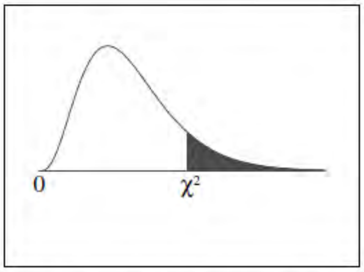
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

NORMAL DISTRIBUTION TABLE



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table of 95% Critical Values for Pearson Correlation Coefficient

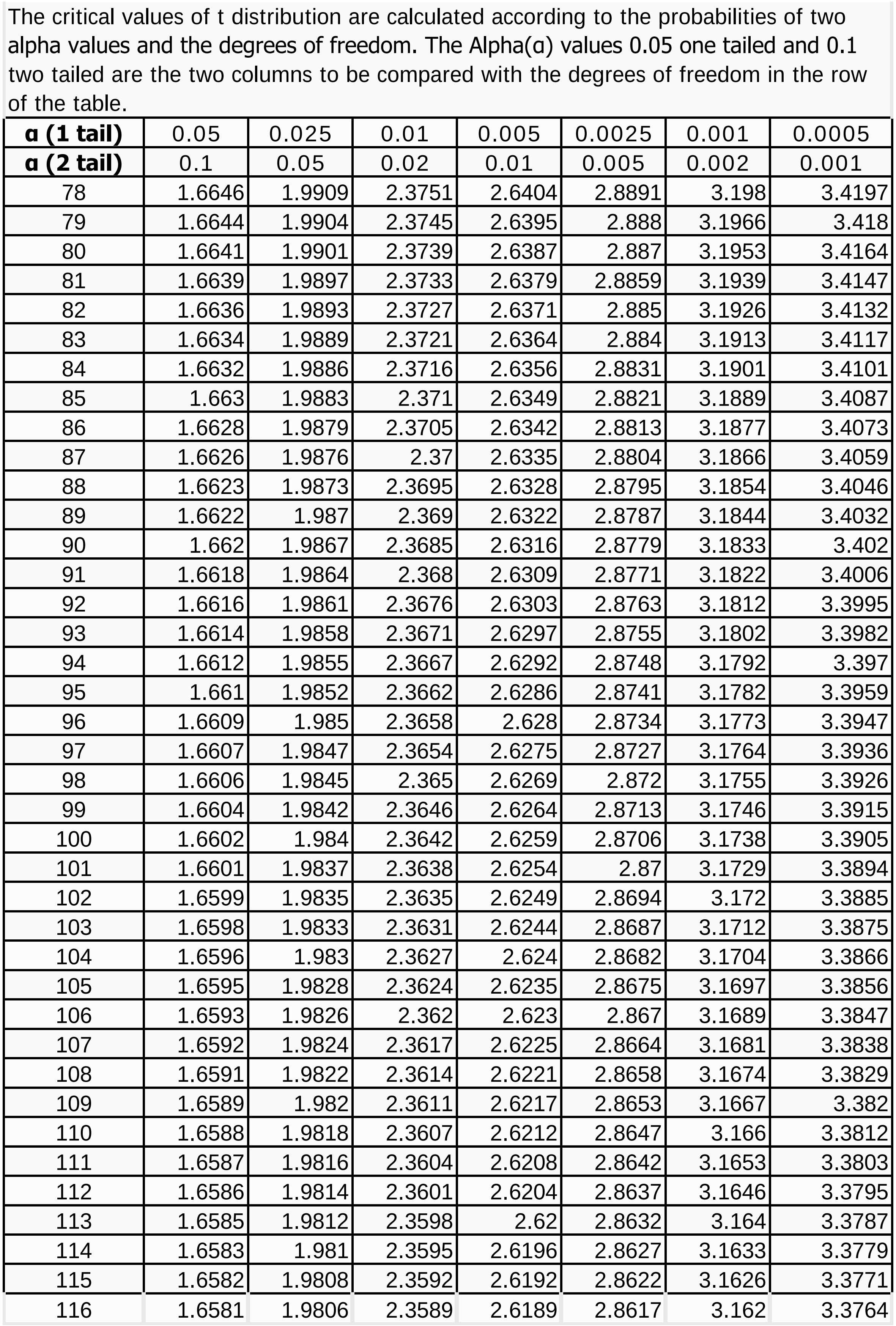
Degrees of Freedom: *n* - 2	**Critical Values: (+and-)**
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

The critical values of t distribution are calculated according to the probabilities of two alpha values and the degrees of freedom. The Alpha(ɑ) values 0.05 one tailed and 0.1 two tailed are the two columns to be compared with the degrees of freedom in the row of the table.

ɑ (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
ɑ (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df 1	6.3138	12.7065	31.8193	63.6551	127.345	318.493	636.045
2	2.92	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.747	4.6041	5.5976	7.1732	8.6103
5	2.015	2.5706	3.365	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.998	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.306	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.201	2.7181	3.1058	3.4966	4.0247	4.4369
12	1.7823	2.1788	2.681	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.852	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
15	1.753	2.1314	2.6025	2.9467	3.286	3.7328	4.0728
16	1.7459	2.1199	2.5835	2.9208	3.252	3.6861	4.015
17	1.7396	2.1098	2.5669	2.8983	3.2224	3.6458	3.9651
18	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	1.7291	2.093	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.7247	2.086	2.528	2.8454	3.1534	3.5518	3.8495
21	1.7207	2.0796	2.5176	2.8314	3.1352	3.5272	3.8193
22	1.7172	2.0739	2.5083	2.8188	3.1188	3.505	3.7921
23	1.7139	2.0686	2.4998	2.8073	3.104	3.485	3.7676
24	1.7109	2.0639	2.4922	2.797	3.0905	3.4668	3.7454
25	1.7081	2.0596	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.7056	2.0555	2.4786	2.7787	3.0669	3.435	3.7067
27	1.7033	2.0518	2.4727	2.7707	3.0565	3.4211	3.6896
28	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.6991	2.0452	2.462	2.7564	3.038	3.3962	3.6594
30	1.6973	2.0423	2.4572	2.75	3.0298	3.3852	3.6459
31	1.6955	2.0395	2.4528	2.744	3.0221	3.3749	3.6334
32	1.6939	2.0369	2.4487	2.7385	3.015	3.3653	3.6218
33	1.6924	2.0345	2.4448	2.7333	3.0082	3.3563	3.6109
34	1.6909	2.0322	2.4411	2.7284	3.0019	3.3479	3.6008
35	1.6896	2.0301	2.4377	2.7238	2.9961	3.34	3.5912
36	1.6883	2.0281	2.4345	2.7195	2.9905	3.3326	3.5822
37	1.6871	2.0262	2.4315	2.7154	2.9853	3.3256	3.5737
38	1.6859	2.0244	2.4286	2.7115	2.9803	3.319	3.5657
39	1.6849	2.0227	2.4258	2.7079	2.9756	3.3128	3.5581
40	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.551
41	1.6829	2.0196	2.4208	2.7012	2.967	3.3013	3.5442
42	1.682	2.0181	2.4185	2.6981	2.963	3.2959	3.5378

The critical values of t distribution are calculated according to the probabilities of two alpha values and the degrees of freedom. The Alpha(ɑ) values 0.05 one tailed and 0.1 two tailed are the two columns to be compared with the degrees of freedom in the row of the table.

ɑ (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
ɑ (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
43	1.6811	2.0167	2.4162	2.6951	2.9591	3.2909	3.5316
44	1.6802	2.0154	2.4142	2.6923	2.9555	3.2861	3.5258
45	1.6794	2.0141	2.4121	2.6896	2.9521	3.2815	3.5202
46	1.6787	2.0129	2.4102	2.687	2.9488	3.2771	3.5149
47	1.6779	2.0117	2.4083	2.6846	2.9456	3.2729	3.5099
48	1.6772	2.0106	2.4066	2.6822	2.9426	3.2689	3.5051
49	1.6766	2.0096	2.4049	2.68	2.9397	3.2651	3.5004
50	1.6759	2.0086	2.4033	2.6778	2.937	3.2614	3.496
51	1.6753	2.0076	2.4017	2.6757	2.9343	3.2579	3.4917
52	1.6747	2.0066	2.4002	2.6737	2.9318	3.2545	3.4877
53	1.6741	2.0057	2.3988	2.6718	2.9293	3.2513	3.4838
54	1.6736	2.0049	2.3974	2.67	2.927	3.2482	3.48
55	1.673	2.0041	2.3961	2.6682	2.9247	3.2451	3.4764
56	1.6725	2.0032	2.3948	2.6665	2.9225	3.2423	3.473
57	1.672	2.0025	2.3936	2.6649	2.9204	3.2394	3.4696
58	1.6715	2.0017	2.3924	2.6633	2.9184	3.2368	3.4663
59	1.6711	2.001	2.3912	2.6618	2.9164	3.2342	3.4632
60	1.6706	2.0003	2.3901	2.6603	2.9146	3.2317	3.4602
61	1.6702	1.9996	2.389	2.6589	2.9127	3.2293	3.4573
62	1.6698	1.999	2.388	2.6575	2.911	3.2269	3.4545
63	1.6694	1.9983	2.387	2.6561	2.9092	3.2247	3.4518
64	1.669	1.9977	2.386	2.6549	2.9076	3.2225	3.4491
65	1.6686	1.9971	2.3851	2.6536	2.906	3.2204	3.4466
66	1.6683	1.9966	2.3842	2.6524	2.9045	3.2184	3.4441
67	1.6679	1.996	2.3833	2.6512	2.903	3.2164	3.4417
68	1.6676	1.9955	2.3824	2.6501	2.9015	3.2144	3.4395
69	1.6673	1.995	2.3816	2.649	2.9001	3.2126	3.4372
70	1.6669	1.9944	2.3808	2.6479	2.8987	3.2108	3.435
71	1.6666	1.9939	2.38	2.6468	2.8974	3.209	3.4329
72	1.6663	1.9935	2.3793	2.6459	2.8961	3.2073	3.4308
73	1.666	1.993	2.3785	2.6449	2.8948	3.2056	3.4288
74	1.6657	1.9925	2.3778	2.6439	2.8936	3.204	3.4269
75	1.6654	1.9921	2.3771	2.643	2.8925	3.2025	3.425
76	1.6652	1.9917	2.3764	2.6421	2.8913	3.201	3.4232
77	1.6649	1.9913	2.3758	2.6412	2.8902	3.1995	3.4214



The critical values of t distribution are calculated according to the probabilities of two alpha values and the degrees of freedom. The Alpha(α) values 0.05 one tailed and 0.1 two tailed are the two columns to be compared with the degrees of freedom in the row of the table.

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
78	1.6646	1.9909	2.3751	2.6404	2.8891	3.198	3.4197
79	1.6644	1.9904	2.3745	2.6395	2.888	3.1966	3.418
80	1.6641	1.9901	2.3739	2.6387	2.887	3.1953	3.4164
81	1.6639	1.9897	2.3733	2.6379	2.8859	3.1939	3.4147
82	1.6636	1.9893	2.3727	2.6371	2.885	3.1926	3.4132
83	1.6634	1.9889	2.3721	2.6364	2.884	3.1913	3.4117
84	1.6632	1.9886	2.3716	2.6356	2.8831	3.1901	3.4101
85	1.663	1.9883	2.371	2.6349	2.8821	3.1889	3.4087
86	1.6628	1.9879	2.3705	2.6342	2.8813	3.1877	3.4073
87	1.6626	1.9876	2.37	2.6335	2.8804	3.1866	3.4059
88	1.6623	1.9873	2.3695	2.6328	2.8795	3.1854	3.4046
89	1.6622	1.987	2.369	2.6322	2.8787	3.1844	3.4032
90	1.662	1.9867	2.3685	2.6316	2.8779	3.1833	3.402
91	1.6618	1.9864	2.368	2.6309	2.8771	3.1822	3.4006
92	1.6616	1.9861	2.3676	2.6303	2.8763	3.1812	3.3995
93	1.6614	1.9858	2.3671	2.6297	2.8755	3.1802	3.3982
94	1.6612	1.9855	2.3667	2.6292	2.8748	3.1792	3.397
95	1.661	1.9852	2.3662	2.6286	2.8741	3.1782	3.3959
96	1.6609	1.985	2.3658	2.628	2.8734	3.1773	3.3947
97	1.6607	1.9847	2.3654	2.6275	2.8727	3.1764	3.3936
98	1.6606	1.9845	2.365	2.6269	2.872	3.1755	3.3926
99	1.6604	1.9842	2.3646	2.6264	2.8713	3.1746	3.3915
100	1.6602	1.984	2.3642	2.6259	2.8706	3.1738	3.3905
101	1.6601	1.9837	2.3638	2.6254	2.87	3.1729	3.3894
102	1.6599	1.9835	2.3635	2.6249	2.8694	3.172	3.3885
103	1.6598	1.9833	2.3631	2.6244	2.8687	3.1712	3.3875
104	1.6596	1.983	2.3627	2.624	2.8682	3.1704	3.3866
105	1.6595	1.9828	2.3624	2.6235	2.8675	3.1697	3.3856
106	1.6593	1.9826	2.362	2.623	2.867	3.1689	3.3847
107	1.6592	1.9824	2.3617	2.6225	2.8664	3.1681	3.3838
108	1.6591	1.9822	2.3614	2.6221	2.8658	3.1674	3.3829
109	1.6589	1.982	2.3611	2.6217	2.8653	3.1667	3.382
110	1.6588	1.9818	2.3607	2.6212	2.8647	3.166	3.3812
111	1.6587	1.9816	2.3604	2.6208	2.8642	3.1653	3.3803
112	1.6586	1.9814	2.3601	2.6204	2.8637	3.1646	3.3795
113	1.6585	1.9812	2.3598	2.62	2.8632	3.164	3.3787
114	1.6583	1.981	2.3595	2.6196	2.8627	3.1633	3.3779
115	1.6582	1.9808	2.3592	2.6192	2.8622	3.1626	3.3771
116	1.6581	1.9806	2.3589	2.6189	2.8617	3.162	3.3764

The critical values of t distribution are calculated according to the probabilities of two alpha values and the degrees of freedom. The Alpha(α) values 0.05 one tailed and 0.1 two tailed are the two columns to be compared with the degrees of freedom in the row of the table.

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
117	1.658	1.9805	2.3586	2.6185	2.8612	3.1614	3.3756
118	1.6579	1.9803	2.3583	2.6181	2.8608	3.1607	3.3749
119	1.6578	1.9801	2.3581	2.6178	2.8603	3.1601	3.3741
120	1.6577	1.9799	2.3578	2.6174	2.8599	3.1595	3.3735
121	1.6575	1.9798	2.3576	2.6171	2.8594	3.1589	3.3727
122	1.6574	1.9796	2.3573	2.6168	2.859	3.1584	3.3721
123	1.6573	1.9794	2.3571	2.6164	2.8585	3.1578	3.3714
124	1.6572	1.9793	2.3568	2.6161	2.8582	3.1573	3.3707
125	1.6571	1.9791	2.3565	2.6158	2.8577	3.1567	3.37
126	1.657	1.979	2.3563	2.6154	2.8573	3.1562	3.3694
127	1.657	1.9788	2.3561	2.6151	2.8569	3.1556	3.3688
128	1.6568	1.9787	2.3559	2.6148	2.8565	3.1551	3.3682
129	1.6568	1.9785	2.3556	2.6145	2.8561	3.1546	3.3676
130	1.6567	1.9784	2.3554	2.6142	2.8557	3.1541	3.3669
131	1.6566	1.9782	2.3552	2.6139	2.8554	3.1536	3.3663
132	1.6565	1.9781	2.3549	2.6136	2.855	3.1531	3.3658
133	1.6564	1.9779	2.3547	2.6133	2.8546	3.1526	3.3652
134	1.6563	1.9778	2.3545	2.613	2.8542	3.1522	3.3646
135	1.6562	1.9777	2.3543	2.6127	2.8539	3.1517	3.3641
136	1.6561	1.9776	2.3541	2.6125	2.8536	3.1512	3.3635
137	1.6561	1.9774	2.3539	2.6122	2.8532	3.1508	3.363
138	1.656	1.9773	2.3537	2.6119	2.8529	3.1503	3.3624
139	1.6559	1.9772	2.3535	2.6117	2.8525	3.1499	3.3619
140	1.6558	1.9771	2.3533	2.6114	2.8522	3.1495	3.3614
141	1.6557	1.9769	2.3531	2.6112	2.8519	3.1491	3.3609
142	1.6557	1.9768	2.3529	2.6109	2.8516	3.1486	3.3604
143	1.6556	1.9767	2.3527	2.6106	2.8512	3.1482	3.3599
144	1.6555	1.9766	2.3525	2.6104	2.851	3.1478	3.3594
145	1.6554	1.9765	2.3523	2.6102	2.8506	3.1474	3.3589
146	1.6554	1.9764	2.3522	2.6099	2.8503	3.147	3.3584
147	1.6553	1.9762	2.352	2.6097	2.85	3.1466	3.3579
148	1.6552	1.9761	2.3518	2.6094	2.8497	3.1462	3.3575
149	1.6551	1.976	2.3516	2.6092	2.8494	3.1458	3.357
150	1.6551	1.9759	2.3515	2.609	2.8491	3.1455	3.3565
151	1.655	1.9758	2.3513	2.6088	2.8489	3.1451	3.3561
152	1.6549	1.9757	2.3511	2.6085	2.8486	3.1447	3.3557
153	1.6549	1.9756	2.351	2.6083	2.8483	3.1443	3.3552
154	1.6548	1.9755	2.3508	2.6081	2.8481	3.144	3.3548
155	1.6547	1.9754	2.3507	2.6079	2.8478	3.1436	3.3544

The critical values of t distribution are calculated according to the probabilities of two alpha values and the degrees of freedom. The Alpha(α) values 0.05 one tailed and 0.1 two tailed are the two columns to be compared with the degrees of freedom in the row of the table.

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
156	1.6547	1.9753	2.3505	2.6077	2.8475	3.1433	3.354
157	1.6546	1.9752	2.3503	2.6075	2.8472	3.143	3.3536
158	1.6546	1.9751	2.3502	2.6073	2.847	3.1426	3.3531
159	1.6545	1.975	2.35	2.6071	2.8467	3.1423	3.3528
160	1.6544	1.9749	2.3499	2.6069	2.8465	3.1419	3.3523
161	1.6544	1.9748	2.3497	2.6067	2.8463	3.1417	3.352
162	1.6543	1.9747	2.3496	2.6065	2.846	3.1413	3.3516
163	1.6543	1.9746	2.3495	2.6063	2.8458	3.141	3.3512
164	1.6542	1.9745	2.3493	2.6062	2.8455	3.1407	3.3508
165	1.6542	1.9744	2.3492	2.606	2.8452	3.1403	3.3505
166	1.6541	1.9744	2.349	2.6058	2.845	3.14	3.3501
167	1.654	1.9743	2.3489	2.6056	2.8448	3.1398	3.3497
168	1.654	1.9742	2.3487	2.6054	2.8446	3.1394	3.3494
169	1.6539	1.9741	2.3486	2.6052	2.8443	3.1392	3.349
170	1.6539	1.974	2.3485	2.6051	2.8441	3.1388	3.3487
171	1.6538	1.9739	2.3484	2.6049	2.8439	3.1386	3.3483
172	1.6537	1.9739	2.3482	2.6047	2.8437	3.1383	3.348
173	1.6537	1.9738	2.3481	2.6046	2.8435	3.138	3.3477
174	1.6537	1.9737	2.348	2.6044	2.8433	3.1377	3.3473
175	1.6536	1.9736	2.3478	2.6042	2.843	3.1375	3.347
176	1.6536	1.9735	2.3477	2.6041	2.8429	3.1372	3.3466
177	1.6535	1.9735	2.3476	2.6039	2.8427	3.1369	3.3464
178	1.6535	1.9734	2.3475	2.6037	2.8424	3.1366	3.346
179	1.6534	1.9733	2.3474	2.6036	2.8423	3.1364	3.3457
180	1.6534	1.9732	2.3472	2.6034	2.842	3.1361	3.3454
181	1.6533	1.9731	2.3471	2.6033	2.8419	3.1358	3.3451
182	1.6533	1.9731	2.347	2.6031	2.8416	3.1356	3.3448
183	1.6532	1.973	2.3469	2.603	2.8415	3.1354	3.3445
184	1.6532	1.9729	2.3468	2.6028	2.8413	3.1351	3.3442
185	1.6531	1.9729	2.3467	2.6027	2.8411	3.1349	3.3439
186	1.6531	1.9728	2.3466	2.6025	2.8409	3.1346	3.3436
187	1.6531	1.9727	2.3465	2.6024	2.8407	3.1344	3.3433
188	1.653	1.9727	2.3463	2.6022	2.8406	3.1341	3.343
189	1.6529	1.9726	2.3463	2.6021	2.8403	3.1339	3.3428
190	1.6529	1.9725	2.3461	2.6019	2.8402	3.1337	3.3425
191	1.6529	1.9725	2.346	2.6018	2.84	3.1334	3.3422
192	1.6528	1.9724	2.3459	2.6017	2.8398	3.1332	3.3419
193	1.6528	1.9723	2.3458	2.6015	2.8397	3.133	3.3417
194	1.6528	1.9723	2.3457	2.6014	2.8395	3.1328	3.3414
195	1.6527	1.9722	2.3456	2.6013	2.8393	3.1326	3.3411
196	1.6527	1.9721	2.3455	2.6012	2.8392	3.1323	3.3409

Solutions to Odd Numbered Problems

CHAPTER 1:

1) AIDS patients; **3** The average length of time (in months) AIDS patients live after treatment.

5) X = the length of time (in months) AIDS patients live after treatment; **7)** b; **9)** a;

11) a. 0.5242 b. 0.03% c. 6.86% d. 823,088/823,856 e. quantitative discrete
f. quantitative continuous g. In both years, underwater earthquakes produced massive tsunamis.

13) systematic; **15)** simple random;

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

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

21) a. all clients for the fitness center b. a smaller, selected group of these clients
c. the population mean number of hours spent each week in the fitness center
d. the sample mean number of hours spent each week in the fitness center
e. X = the number of hours spent in the fitness center for a given client in a given week.
f. values for X, such as 2, 1.7, 3.5, …

23) a. all patients of the doctor b. a smaller, selected group of these patients
c. the mean recovery period for all patients d. the mean recovery time for the sample patients
e. X = the recovery time of a single patient f. values for X, such as , 9 months, 4.3 months,…

25) a. all voters in the district b. a smaller, selected group of these voters
c. the proportion of all her voters in the district who approve of the politician's performance.
d. the proportion of the sample who approve of the politician's job performance.
e. X = whether or not a voter approves of the politician's job performance f. yes, no

27) a. all voters in the region (county, state or nation) b. a smaller, selected group of these voters
c. the proportion of all voters in the region who will vote for the cause.
d. the proportion of the sample who will vote for the cause.
e. X = whether or not a voter will vote for the cause f. yes, no

29) a **31)** c **33)** quantitative continuous; e.g. 21.2%

35) quantitative discrete; e.g. 11,234 students **37)** qualitative; e.g. Crest, Colgate

39) quantitative continuous; e.g. 51 yrs, 63.5 yrs

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

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

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

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

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

51) Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate. Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

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

CHAPTER 2:

1) 1.5 – 2.5

3) n = ∑f = 40

5) frequency distribution

7) midpoints: 34, 39, 44, 49, 54, 59, 64

# of potholes	**f**
1	12
2	4
3	5
4	5
5	3
6	3
7	3

9) r.f: 2.9%, 8.6%, 14.3%, 31.4%, 20%, 20%, 2.9%

11) n = ∑f = 35

13) midpoints: 74, 83, 92, 101, 110, 119, 128

15) r.f: .147, .158, .168, .189, .232, .074, .032

17) n = ∑f = 95

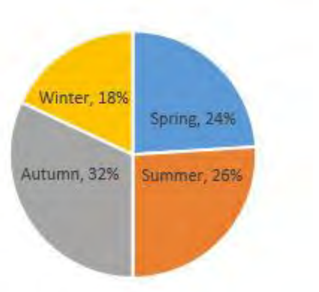
19) grouped frequency distribution with class width = 18

Classes	Frequency
236 – 253	3
254 – 271	5
272 – 289	11
290 – 307	17
308 – 325	11
326 – 343	9
344 – 361	1

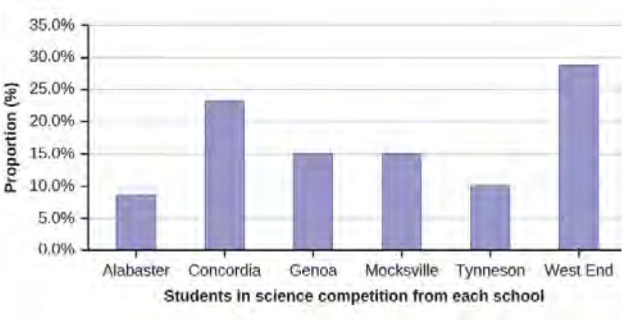
21) grouped frequency distribution with class width = 86

Classes	Frequency	Midpoint	Rel. Freq.	Cum. Freq.
30 – 115	4	72.5	14.3%	4
116 – 201	5	158.5	17.9%	9
202 – 287	5	244.5	17.9%	14
288 – 373	3	330.5	10.7%	17
374 – 459	5	416.5	17.9%	22
460 – 545	6	502.5	21.4%	28

23)



25)



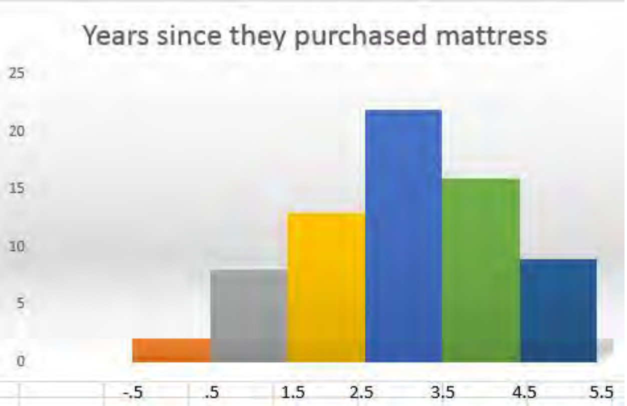
27)

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

29)

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

31)

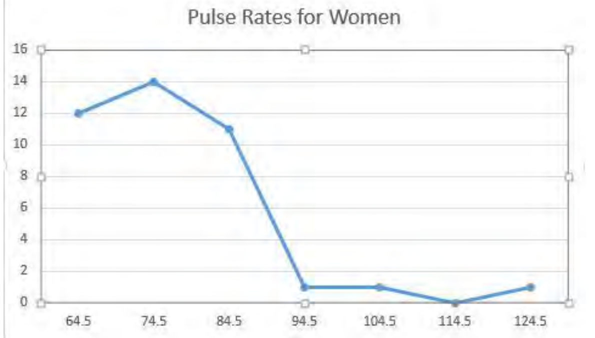


33)

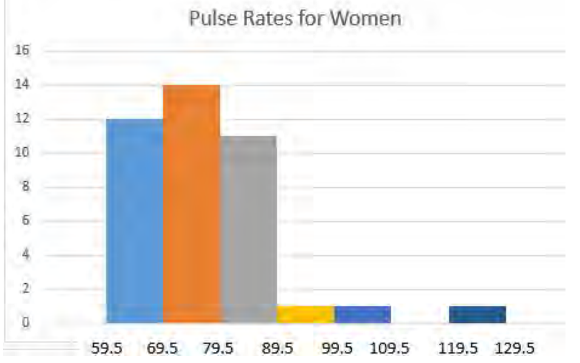
Cars	Freq.	Rel Freq	Cum. Freq
3	14	21.5%	14
4	19	29.2%	33
5	12	18.5%	45
6	9	13.8%	54
7	11	16.9%	65

35)

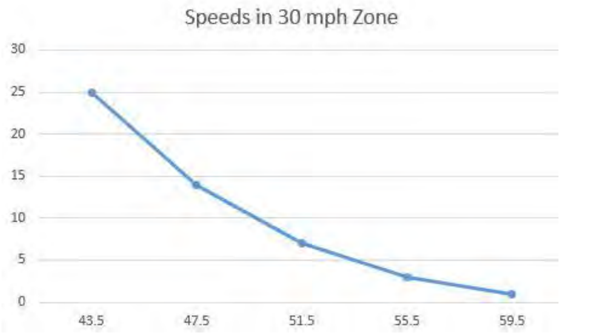
a. Frequency Polygon

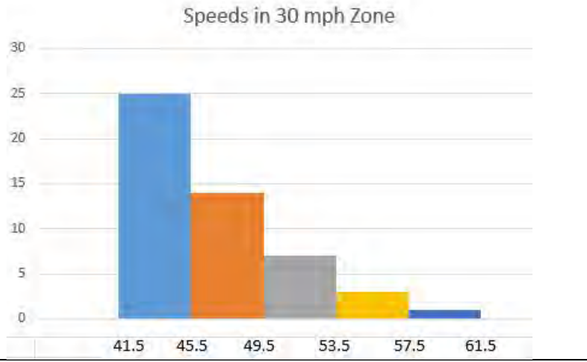


Histogram

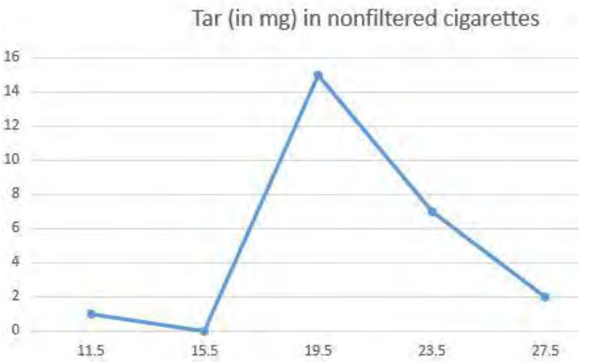


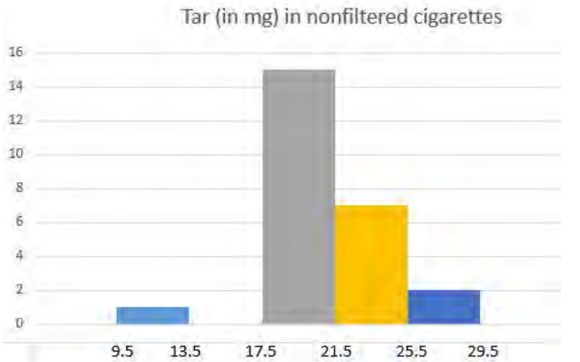
b.



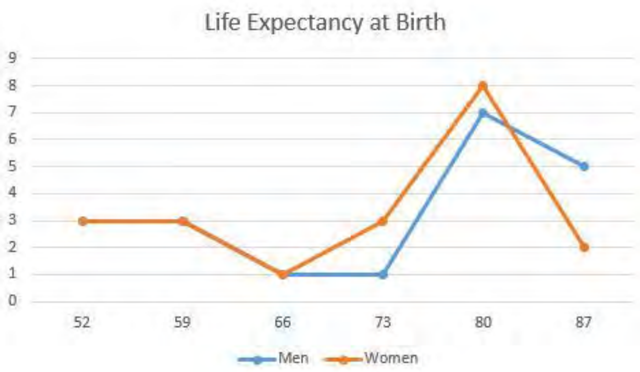


c.



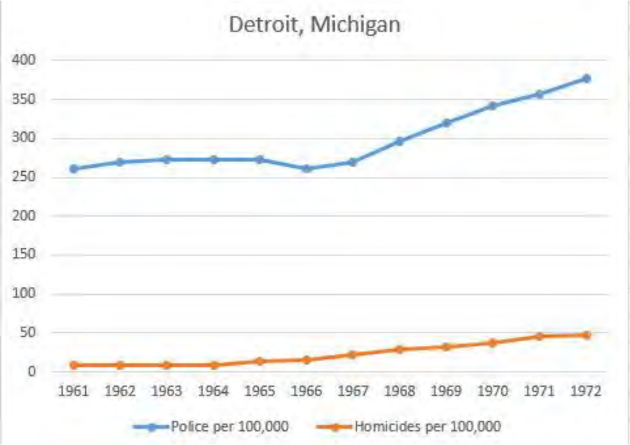


37)



Both skewed left. Similar life expectancy pattern for both men and women

39) Police variable increased faster. The number of police did not impact the number of homicides. Homicides continued to increase.



41) $\bar{x} = 11.9$, med = 12, s = 6.71, min = 2, Q1 = 5, Q2 = 12, Q3 = 17, max = 23

43) $\bar{x} = 33.1$, med = 32, s = 14.2, min = 2, Q1 = 24, Q2 = 32, Q3 = 44, max = 57

45) $\bar{x} = 4.8$, med = 4, mode = 4

47) They are the same

49) mean

51) The mean reflects skewing the most since it can be "pulled" to the right or left when extreme values are either to the right or left respectively.

53) s = 34.5, $\bar{x}$ - s = 49.6

55) $\bar{x} = 41.3$, med = 39, s = 8.4

57) $\bar{x} = 72$, med = 71.5, s = 6.8

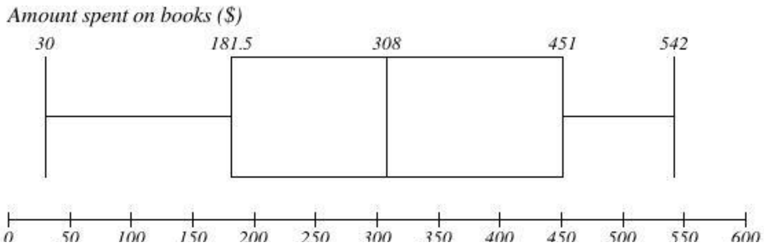
59) **a.)** P40 = 37 years old, **b.)** P78 = 70 years old, **c.)** 40th percentile, **d.)** 84th percentile

61) High percentile

63) Li's salary is higher than 78% of recent graduates; she should be pleased.

65) 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more. So you can afford 34% of houses and 66% of the houses are too expensive for your budget.

67)



69) d

71) a.) 17 or younger is 25% while 65 and older is less than 25% b.) 62.4%

73) a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.

b. The BMW 3 series is most likely to have an outlier. It has the longest whisker.

c. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.

d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.

e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.

f. *IQR* ~ 17 years

g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.

h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

75) a.) $\bar{x}$ = 23.3 b.) above average c.) s = 12.9 d.) It is within 1 standard deviation therefore not unusual

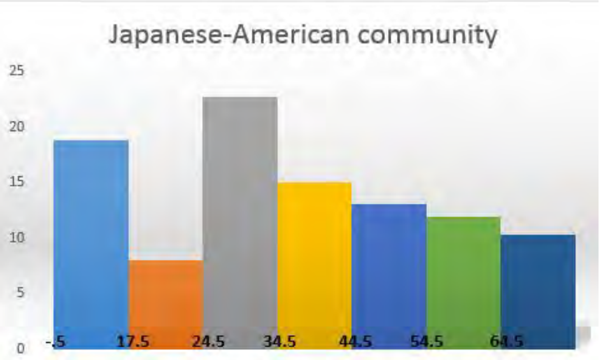
77) Z_{piano} = -.4; Z_{guitar} = .25; Z_{drum} = -1. Drums cost the lowest. Guitar cost the highest.

79) Z_{Rachel} = -1 ; Z_{Kenjii} = -.25 ; Z_{Nedda} = .25.

a.) Kenjii ran faster than the average.
b.) Rachel is the fastest runner with respect to her class because she had the lowest time with respect to the mean.

81) IQR = 8; mode = 19

83)



a. The first class has almost double the width of the majority of classes. The second class has the smallest class width. This gives a distortion of the data.

b. 49.7% c. first boxplot (a)

85) a. true since median = 3 for all three

b. false, mean of third graph would be skewed to right more than first graph

c. true, graph two the data is more spread out to the tails while graph one is concentrated in the middle.

d. false, the third graph the 3rd quartile is larger than 6 while the other two has no data greater than 5.

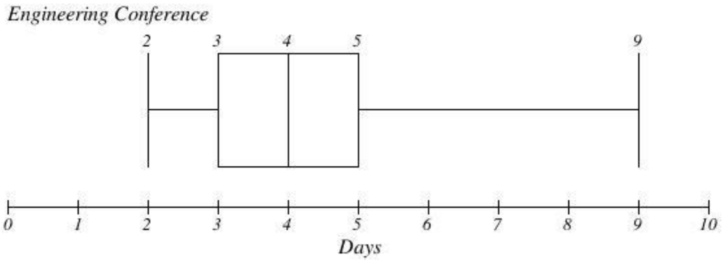
87)

a.

Days	Freq	R.F.
2	4	4.8%
3	36	42.9%
4	18	21.4%
5	19	22.6%
6	4	4.8%
7	1	1.2%
8	1	1.2%
9	1	1.2%

b. Q1 = 3 days, Q2 = 4 days, Q3 = 5 days

c. P65 = 4 days **d.** 3 to 5 days **e.**



f. P10 = 3 **g.** $\bar{x}$ = 3.9 days **h.** s = 1.3 days **i.** mode = 3 days

j. choose mode = 4 days since it is also close to mean = 3.9 days.

k. answers vary

89) a. between 63.7 and 74.9 inches **b**. between 66.5 and 72.1 inches

c. Z = 3.1 **d**. 2.5%

91) a. 11.8% **b.** -10.7% and 16.3% **c.** 7.2%

93) a. at least 75% **b.** between -.5% and 12.1%

CHAPTER 3:

1) S = {BB, BG, BP, GG, GB, GP, PP, PB, PG}

3) S = {AB, AC, AD, AE, AF, AG, BA, BC, BD, BE, BF, BG, CA, CB, CD, CE, CF, CG, DA, DB, DC, DE, DF, DG, EA, EB, EC, ED, EF, EG, FA, FB, FC, FD, FE, FG, GA, GB, GC, GD, GE, GF}

5) P(H) = 12/42 = 2/7 P(N) = 15/42 = 5/14 P(F) = 10/42 = 5/21 P(C) = 5/42

7) P(A) = 22/97 P(E) = 47/194 P(F) = 27/97 P(N) = 23/194 P(O) = 7/97 P(S) = 6/97

9) P(even) = ½ P(prime) = ½

11) a. P(O’) **b.** P(O or H) **c.** P(I and H’) **d.** P(H | I)

13) Probability of an event knowing that a previous event already occurred

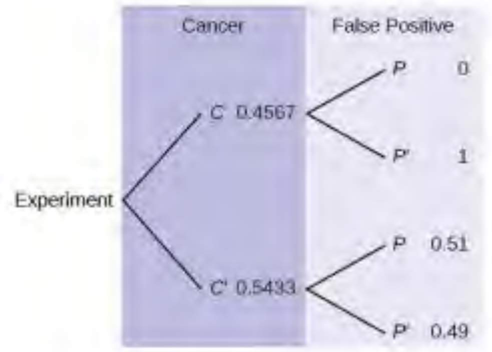
15) P(E) + P(E’) = 1

17) P(E | F) = P (E and F)/ P(F) = 0/.5 = 0

19) a. 0 **b**. 0 **c**. 0.63

21) a. P(C) = .48 **b.**P(L) = .376 **c.** P(C | L) = .55 **d**. Choosing a Californian registered voter who prefer life in prison…knowing that they are Latino Californian. **e.** P(L and C) = .376(.55) = .207
f. a person who is Latino Californian and who is a registered voter that prefers life in prison …
g. L and C are not independents because P(C | L) ≠ P(C)
h. P(L or C) = .376 + .48 - .207 = .649 **i.** a person who is Latino Californian or who is registered voter that prefers life in prison… **j.** L and C are not mutually exclusive because P(L and C) ≠ 0

23)



25) a. Answers vary (i.e. individual is male, individual is between 18-34, ..)
b. 40% of total approve of Mayor Ford’s action in office
c. 60% of total disapprove …
d. 30% of the 18-34 years old approve of Mayor …
e. 45.7% **f.** 63% **g.** 40% **h.** 44% **i.** 78.9% **j.** 30%

27) a. 1/7 **b.** 0 **c.** ½ **d.** 5/14 **e.** 2/7 **f.** 3/14 **g.** 5/7 **h.** Physician **i.** Service **j.** 4.1 **k.** 81.3 **l.** ½

29) a. 1046 **b.** 58% **c.** 439 **d.** 57% **e.** 60%

31) a. 6/38 **b.** 3/38 **c.** 1/38 **d.** 4/38 **e.** 2/38 **f.** 5/38 **g.** 3/38

33) a. S = {G1, G2, G3, G4, G5, Y1, Y2, Y3} **b.** P(G) = 5/8 **c**. P(G | E) = 2/3

d. P(G and E) = ¼ **e.** P(G or E) = ¾ **f.** Not mutually exclusive since P(G and E) ≠ 0

35) a. S = {GH, GT, BH, BT, RH, RT} **b.** 3/10*1/2 = 3/20 c. Yes mutually exclusive since they can't happen at the same time.

37) a. S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT} **b.** P(A) = ½

c. Not mutually exclusive because can't happen at the same time.

39) a. P(Y or Z) = P(Y) + P(Z) – P(Y)*P(Z) **b**. 0.5

41) a. iii **b.** i **c.** iv **d.** ii

43) a. P(R) = .44 **b**. P(R | E) = .56 **c**. P(R | O) = .31 **d.** Not independent because P(R | E) ≠ P(R) **e.** No because P(R | E) > P(R)

45) a. .52 = .43 + .15 – P(O and Rh-); P(O and Rh-) = .06 **b.** complement of part a: .96

47) a. P(C or N) = .36 + .12 - .08 = .4 **b**. complement of part a: .6

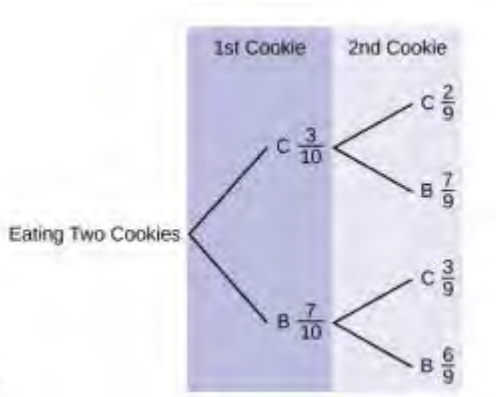
49) a. P(independent) = 2/67 **b.** 34/67 **c.** 10/67 **d.** 8/67 **e.** 10/34

f. 8/32 **g.** *iv* **h.** *i*

51) a. *ii* **b.** *ii*

53) a. 26/106 **b.** 33/106 **c.** 21/106 **d.** 38/106 **e**. 21/33

55) a.



b. Not independent since there is no replacement (eaten)

c. P(CC) = 6/90; P(CB) = 21/90; P(BC) = 21/90 P(BB) = 42/90

d. 48/90

e. 42/90

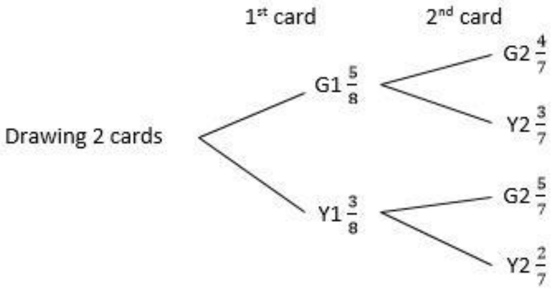
f. 63/90

57) a. .4567 **b.** 0 base on tree diagram from problem 23 **c.** 0.51 based on tree diagram

d. No because at least 1 false positive happens over 50% of time.

59) a. 0.1 **b.** 0.90 **c.** P(J or K) = .45

61) a.



b. 5/14

c. 25/28

d. 4/7

e. Not independent because

P(G2 | G1) = P(G)

63) a. 5140

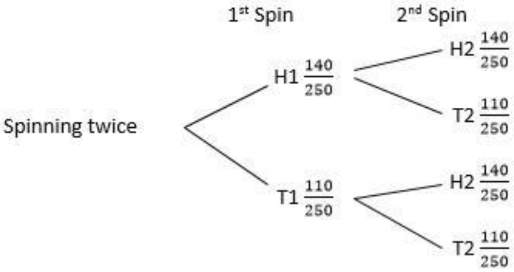
b.

	< 20	20 – 64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

c. .4857

65) a. P(H) = 140/250 P(T) = 110/250

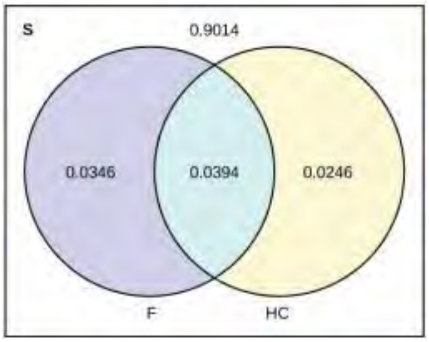
b.



c. 308/625

d. 504/625

67) a. 0.074 **b.** 0.064 **c.** 0.6163 **d.**



69) 32 ballots

71) a. 17,576,000 **b.** 15,600,000

73) 12!

75) a. $\frac{{}_{10}C_5}{{}_{22}C_5}$ **b.** $\frac{{}_{12}C_3 \cdot {}_{10}C_2}{{}_{22}C_5}$ **c.** $\frac{{}_{10}C_4 \cdot {}_{12}C_1}{{}_{22}C_5} + \frac{{}_{10}C_5}{{}_{22}C_5}$

CHAPTER 4:

1)

x	P(x)
0	0.12
1	0.18
2	0.30
3	0.15
4	**0.10**
5	0.10
6	0.05

3) P($x \geq 5$) = 0.10 + 0.05 = 0.15 **5)** 1 **7)** P($x > 1$) = 0.35 + 0.40 + 0.10 = 0.85
9) E(X) = 1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45

11)

x	P(x)
0	0.03
1	0.04
2	0.08
3	0.85

13) Let X = the number of events Javier volunteers for each month.

15)

x	P(x)
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

17) P($x > 0$) = 1 – P($x = 0$) = 0.95.
19) **a**. X = number of years needed to complete B.S. degree.
b. No students completed B.S. degrees in two years or less.
c. E[x] = 4.85 years
21) X = the number of years a physics major will spend doing post-graduate research.
23) 0.15
25) E(X) **=** 1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 2.6 years
27) P($x \geq 5$) = .30 + .20 + .10 = 0.60.

29) P($x < 4$) = 0.10 + 0.05 + 0.10 = 0.25
31) The sum of the probabilities sum to one because it is a probability distribution.
35) E(X) = 30(12/52) – 2(40/52) = $5.38.
37) X = the number that reply “yes”

39) x = 0, 1, 2, 3, 4, 5, 6, 7, 8
41) μ = (8)(.713) = 5.7

43) $P(x \leq 5)$ = binomcdf(8, .713, 5) = 0.4151.

45) E[x] = 6(6/52) + 2(6/52) – 2(40/62) = -$.62 No you should not play this game to win money.

47) a.

Software company

x	P(x)
5,000,000	.10
1,000,000	.30
-1,000,000	.60

Hardware company

x	P(x)
3,000,000	.20
1,000,000	.40
-1,000,000	.40

Biotech company

x	P(x)
6,000,000	.10
1,000,000	.70
-1,000,000	.20

b. E[x] for software = $200,000

E[x] for hardware = $600,000

E[x] for Biotech = $400,000

c. 3rd investment because it has the lowest probability of loss

d. 1st investment because it has the highest probability of loss

e. 2nd investment

49) b

51) Let X = the amount of money to be won on a ticket. The following table shows the PDF for X.

X	P(x)
0	0.969
5	0.025
25	0.005
100	0.001

E(X) = 0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35. So a fair price for a ticket is $0.35. Any price over $0.35 will enable the lottery to raise money.

53) X = the number of patients calling in claiming to have the flu, who actually have the flu.
Possible values of X are x = 0, 1, 2, ...25   **55)** 0.0165

57) a. X = the number of DVDs a Video to Go customer rents   b. $P(x = 3) = 0.12$
c. $P(x \geq 4) = 0.07 + 0.04 = 0.11$ d. $P(x \leq 2) = 0.77$

59) 4.43   **61)** c

63) X = number of questions answered correctly; $X \sim B(32, 0.333)$.
We want the probability that *more than* 75% of 32 questions correct; so we want $P(x > 24)$.
The event "more than 24" is the complement of "less than or equal to 24."
So $P(x \geq 24) = 1 - P(x \leq 24)$ = 1 – binomcdf(32, .333, 24) ≈ 0.00000026.

65) a. X = the number of college and universities that offer online offerings.

b. Possible values are x = 0, 1, 2, …, 13. c. $X \sim B(13, 0.96)$ d. 12.48 e. 0.013

f. $P(x = 12) = 0.3186$; $P(x = 13) = 0.5882$. So it is more likely to get 13.

67) a. X = the number of fencers who do **not** use the foil as their main weapon

b. x = 0, 1, 2, 3,... 25 c. $X \sim B(25, 0.40)$ d. 10 e. 0.0442

f. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

69) a. X = the number of audits in a 20-year period b. x = 0, 1, 2, …, 20 c. $X \sim B(20, 0.02)$

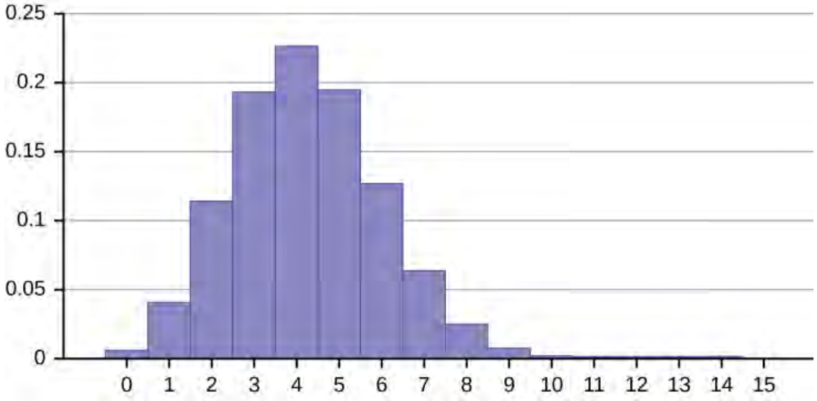
d. 0.4 e. 0.6676 f. 0.0071

71) a. X = the number of matches b. 0, 1, 2, 3 c. $X \sim B(3, 1/6)$ d. In dollars: $x = -1, 1, 2, 3$

e. ½ f. Multiply each Y value by the corresponding X probability from the PDF table to get -0.0787. You lose about eight cents, on average, per game.

g. The house has the advantage.

73) a. $X \sim B(15, 0.281)$



b. $\mu = np = 15(0.281) = 4.215$, $\sigma = \sqrt{npq} = \sqrt{15(0.281)(719)}$ =1.7409

c. $P(x > 5) = 1 - P(x \leq 5)$ = 1 – binomcdf(15, 0.281, 5) = 1 – 0.7754 = 0.2246

$P(x = 3)$ = binompdf(15, 0.281, 3) = 0.1927

$P(x = 4)$ = binompdf(15, 0.281, 4) = 0.2259

It is more likely that four people are literate than three people are.

75) 1.4

77) a. X = the number of adults surveyed until one says he or she will watch the Super Bowl.

b. $X \sim G(0.40)$ c. $\mu = 2.5$ d. 0.0187 e. 0.2304

79) a. X = the number of pages that advertise footwear b. X takes on the values 0, 1, 2, ..., 20

c. $X \sim B(20, 29/192)$ d. 3.02 e. No f. 0.9997

g. X = the number of pages we must survey until we find one that advertises footwear.

$X \sim G(29/192)$ h. $P(x \leq 3)$ = geometcdf(29/192, 3) = 0.3881

i. μ = 192/29 = 6.6207 pages.

81) a. $X \sim G(0.25)$ μ = 1/0.25 = 4. c. $P(x = 10)$ = geometpdf(0.25, 10) = 0.0188

d. $P(x = 20)$ = geometpdf(0.25, 20) = 0.0011 e. $P(x \leq 5)$ = geometcdf(0.25, 5) = 0.7627.

83) X = the number of U.S. teens who die from motor vehicle injuries per day. **85)** 0, 1, 2, 3, ...

87) No **89)** X = number of customers per day; $X \sim P(120)$.

91) If the store averages 120 customers per day, then it will average 10 customers per hour, or 40 per each four hours. So $P(x = 35)$ = poissonpdf(40, 35) = 0.0485

93) The store averages 120 customers per day, so it will average 10 customers per hour, or 20 per each two hours. So P(x < 12) = P($x \leq$ 11) = poissonpdf(20, 11) = 0.0214.

95) $X \sim P(5.5)$ a. μ = 5.5 and σ = 5.5 b. P($x \leq$ 6) = poissoncdf(5.5, 6) ≈ 0.6860
c. P(x = 6) = poissonpdf(5.5, 6). There is a 15.7% probability that the law staff will receive more calls than they can handle.
d. $P(x > 8) = 1 - P(x \leq 8)$ = 1 – poissoncdf(5.5, 8) ≈ 1 – 0.8944 = 0.1056

97) a. X = the number of children for a Spanish woman; b. 0, 1, 2, 3,.. ; .c. $X \sim P(1.47)$
d. P(x = 0) = poissonpdf(1.47, 0) = 0.2299; e. P(x < 1.47) = poissoncdf(1.47, 1) = 0.5679
f. P(x > 1.47) = 1 – P(x $\leq$ 1.47) = 1 – 0.5679 = 0.4321

99) a. X = the number of people audited in one year b. 0, 1, 2, ..., 100 c. $X \sim P(2)$ d. 2
e. P(x = 0) = poissonpdf(2, 0) = 0.1353;
f. P($x \geq$ 3) = 1 – P(x < 2) = 1 – poissoncdf(2, 2) = 0.3233.

101) a. X~B(10, .15) **b.** .95 **c.** 0.00138

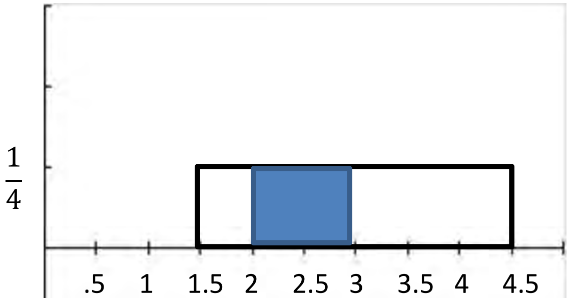
CHAPTER 5:

1) X ~ U(3, 8) **3)** P(2 < x < 5) **5)** 0 **7)** 0 **9)** 0

11) P(6 < x < 8) = 2(1/9) = 2/9 **13)** P(2.5 < x < 5.5) = 3(1/10) = 0.3

15) a. 1/3 b. a = 1.5, b = 4.5

c. Graph:



d. P(2 < x < 3) = base*height = 1(1/3) = 1/3

e. P(x < 3.5 | x < 4) = $\frac{P(x<3.5\ and\ x<4)}{P(x<4)} = \frac{P(x<3.5)}{P(x<4)} = \frac{2*1/3}{2.5*1/3}$ = 0.8

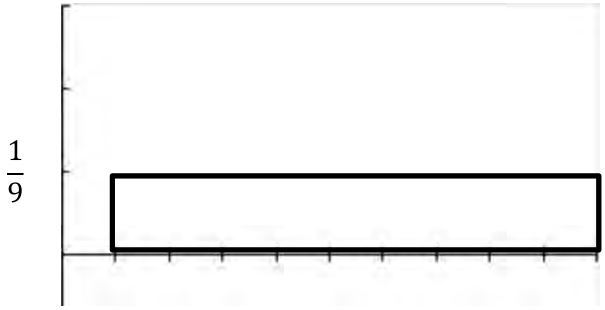
f. P(x = 1.5) = 0 g. P_{90} = k , where 0.90 = (k – 1.5)(1/3) => k = 4.2

h. P(x > 3 | x > 2) = 0.6

17) a. Age of cars b. X = The age (in years) of cars in the staff parking lot

c. continuous d. 0.5 to 9.5 e. X ~ U(.5, 9.5) f. f(x) = 1/9

g. Graph:



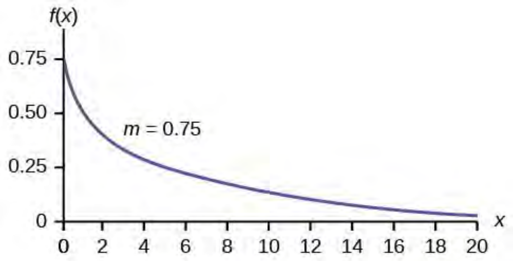
h. P(x < 4) = .389 i. P(x < 4 | x < 7.5) = 0.5

j. The second problem had a given situation

k. P_{75} = k, where .75 = (k - .5)(1/9) => k = 7.25.

19) a. m = .75 b. f(x) = -.75$e^{-.75x}$ c. P(x < k) = 1 – $e^{-.75k}$

d. Graph:



e. $P(x < 4) = 1 – e^{-.75(4)} = .95$ f. $k = \ln(.70)/-.75 = .48$

g. median = $\ln(.50)/-.75 = .92$ h. $\mu = 1.33$ which is larger than the median.

21) a. The number of nurses that said yes. b. The percentage of nurses that said yes.

23) a. X ~ U(1, 53) b. Graph is a uniform graph spanning from 1 to 52 with a height of 1/52.

c. $f(x) = 1/9$ d. $\mu = 54/2 = 27$ e. $\sigma = 52/\text{sqrt}(12) = 15.01$

f. $P(x = 19) = 0$ g. $P(2 < x < 31)$ = base*height = 29(1/52) = .558

h. $P(x > 40) = 13\ (1/52) = .25$

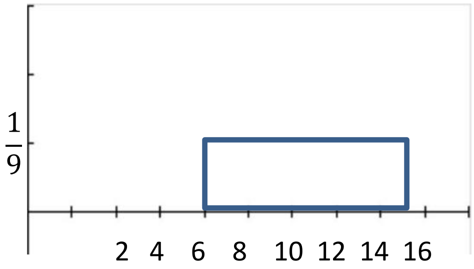
i. $P(12 < x \mid x < 28) = P(12 < x < 28)/ P(x < 28) = .308/.519 =.593$

j. $P_{70} = k$, where $0.70 = (k – 1)(1/52)$ => $k = 37.4$

k. $Q_3 = k$, where $0.75 = (k – 1)(1/52)$ => $k = 40$.

25) a. X = weight loss in a month b. X ~U(6, 15)

c. Graph:



d. $f(x) = 1/9$ e. $\mu = 21/2 = 10.5$ f. $\sigma = 9/\text{sqrt}(12) = 2.6$

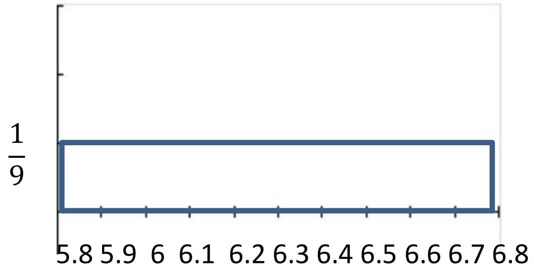
g. $P(x > 10)$ = base*height = 5(1/9) = .556

h. $P(x < 12 \mid x > 10) = P(10 < x < 12) / P(x > 10) = .222/.556 = .399$

i. $P(7< x < 13 \mid x > 9) = P(9 < x < 13) / P(x > 9) = .444/.667= .666$

27) a. X = age of first grade on Sept. 1st b. X ~U(5.8, 6.8)

c. Graph:



d. $f(x) = 1$ e. $\mu = 12.6/2 = 6.3$ f. $\sigma = 1/\text{sqrt}(12) = .289$

g. $P(x > 6.5)$ = base*height = .3(1) = 0.3 h. $P(4 < x < 6) =P(x < 6) =.2(1) = 0.2$

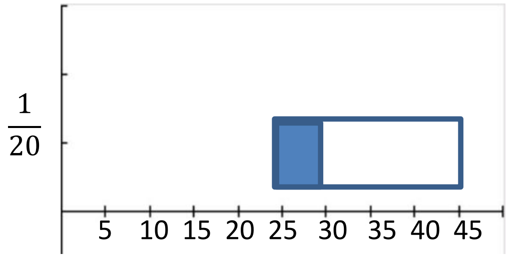
h. $P_{70} = k$, where $.70 = (k – 5.8)(1)$ => $k = 6.5$

29) a. X = minutes until the next bus departs b. X ~U(25, 45)

c. Graph: a rectangle that spans from 25 to 45 on x –axis with a height of 1/20

d. Uniform. It is continuous. e. $\mu = 70/2 = 35$ f. $\sigma = 20/\text{sqrt}(12) = 5.77$

g. $P(x \le 30)$ = base* height = 5(1/20) = .25



h. $P(30 < x < 40)$ = base*height = 10(1/20) = .50 i. $P(25 < x < 55) = P(x > 25) = 1$

j. $P_{90} = k$, where $.90 = (k - 25)(1/20) \Rightarrow k = 43$.

k. $P_{75} = k$, where $.75 = (k - 25)(1/20) \Rightarrow k = 40$

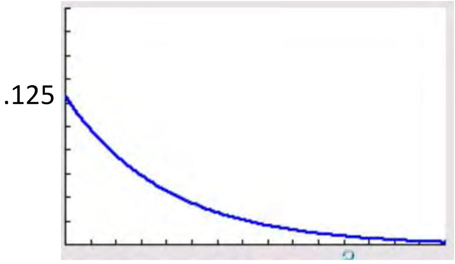
l. $P(x > 40 \mid x \geq 30) = P(x > 40) / P(x \geq 30) = .25/.75 = 1/3$

31) a. $\mu = 3$ b. $P(x > 4) = .25$

33) a. X = minutes on a long distance phone call b. X ~ exp(.125)

c. Continuous d. $\mu = 8$ (given in the problem) e. $\sigma = 8$

f. Graph:



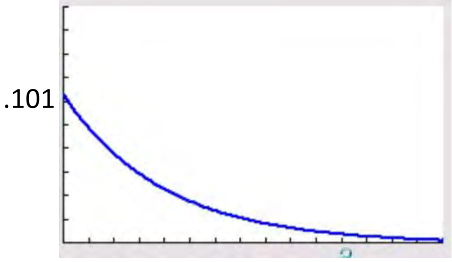
g. $P(x < 9) = 1 - e^{(-.125*9)} = .675$ h. $P(x > 9) = e^{(-.125*9)} = .325$

i. $P(7 < x < 9) = e^{(-.125*7)} - e^{(-.125*9)} = .0922$ j. 25*8 = 200 minutes since average is 8 minutes.

35) a. X = percent of persons in each state who speak a language at home other than English

b. Continuous c. X ~ exp(.101) d. $\mu = 9.848$ (given in the problem) e. $\sigma = 9.848$

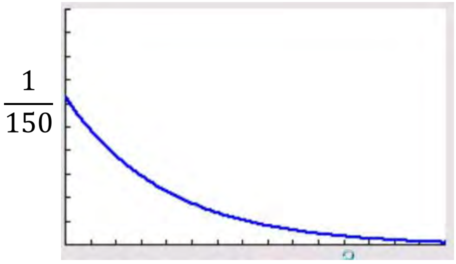
g. Graph:



h. $P(x < 12) = 1 - e^{(-.101*12)} = .702$ i. $P(8 < x < 14) = e^{(-.101*8)} - e^{(-.101*14)} = .203$

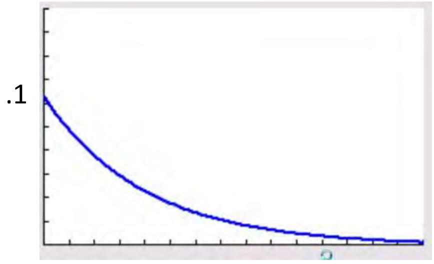
j. The number is different because the population is different. It makes it higher because of the younger age children are included.

37) a. X = Cost of car during first year b. X ~ exp(1/150)
c. $\mu = 150$ (given in the problem) d. $\sigma = 150$
e. Graph:

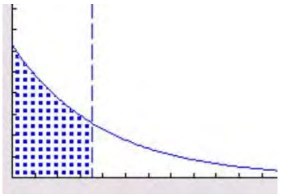


f. $P(x > 300) = e^{(-2)} = .135$

39) a. Decay rate = 0.1 b. $\mu = 10$
c. Graph:



d. $P(x < 6) = 1 - e^{(-.10*6)} = .451$ e. $P(3 < x < 6) = e^{(-.10*3)} - e^{(-.10*6)} = .192$
f. $P(x < 7) = 1 - e^{(-.10*7)} = .503$



g. $P_{40} = \ln(.60)/-.1 = 5.108$ h. Average = μ = 10.

41) a. $\mu = 2.5$ b. $P(x > 3)$ c. $P_{90} = \ln(.10)/-.5) = 4.605$
d. $P(x < 3 \mid x > 2) = P(2 < x < 3)/P(x > 2) = \frac{e^{-.5*2}-e^{-.5*3}}{e^{-.5*2}}$=.393
e. Poisson distribution; poissoncdf(30, 19) = .0219

43) a. $P(x < 3) = 1 - e^{-.483(3)} = .765$
b. $P(x < 9 \mid x > 6) = P(6 < x < 9)/ P(x > 6) = \frac{e^{-.483*6}-e^{-.483*9}}{e^{-.483*6}}$=.765
c. $P(x = 0) = 0$ d. $P(x \geq 2) = .381$

45) a. $P(x > 10) = e^{-.2(10)}$ =.135. m = 12/60 = .2 b. $P_{75} = \ln(.25)/-.2 = 6.93$ minutes
c. $P(x < 25 \mid x > 20) = P(20 < x < 25)/P(x > 20) = .632$
d. $P(x < 7$ visits) = poissoncdf(12, 6) = .0458

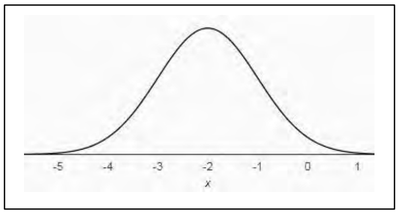
CHAPTER 6:

1) $X \sim N(12.05,\ 0.01)$ **3)** $\sigma = 2$ **5)** Median = μ = -4 **7)** μ = -2

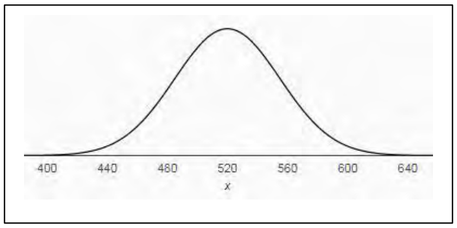
9) Z = -1.5 **11)** z = 1.67 **13)** z = -1.67 **15)** $z \approx -0.33$ **17)** x = 6.5

19) x = 1 **21)** x = 1.97

23)



25)



27) 0.67, right **29)** 3.14, left

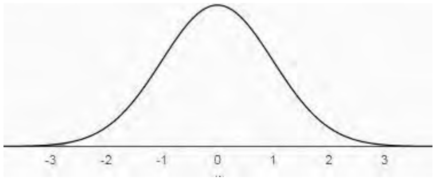
31) about 68% **33)** about 4% **35)** between –5 and –1 **37)** about 50% **39)** about 27%

41) $P(x < 1)$

43) Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$.

45) $P(x > 3)$ or $1 - P(x < 3)$

47) $P(z < 2.16) = .00272$



49) 1 – 0.543 = 0.457 **51)** 0.0013 **53)** 56.03 **55)** 0.1186

57) **a.** Check student's solution. **b.** X~N(7.5, 1.3) **c.** 0.0272 **d.** 0.0272, unusual

59) a. Check student's solution. **b.** X~N(5, .85) **c.** 0.0093 **d.** 0.380

61) z = 2.35

63) a. Use the *z*-score formula to get $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations *below* the mean. An NBA player whose height is 77 inches is shorter than average.
b. Use the *z*-score formula to get $z = 1.5424$. The height 85 inches is 1.5424 standard deviations *above* the mean. An NBA player whose height is 85 inches is taller than average.
c. Height = 79 + 3.5(3.89) = 90.67 inches, which is over 7.7 feet tall.
There are very few NBA players this tall so the answer is no, not likely.

65) a. iv b. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

67) a. $z = (720 - 520)/115 = 1.74$. This student scored 1.74 standard deviations above the mean.
b. $x = \mu + 1.5\sigma = 520 + 1.5(115) = 692.5$. This student scored 1.5 standard deviations above the mean.
c. Let X = an SAT math score and Y = an ACT math score.
=> The z-score for the SAT is $z = (700 - 514)/117 = 1.59$.
=>The z-score for the ACT is $z = (30 - 21)/5.3 = 1.67$.
The student who took the ACT did better relative to his/her peers (has the higher *z*-score).

69) 7.99 **71)** 0.0668

73) a. $X \sim N(66, 2.5)$ b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

75) a. $X \sim N(36, 10)$ b. $P(x > 40) =$ 0.3446.
c. Approximately 25% of people consume less than 29.26% of their calories as fat.

77) a. X = # of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. $X \sim N(3, 1.5)$ c. $P(x < 1) = 0.0918$. d. $P(x > 10) =$ 0.0000015. e. 2.21 hours

79) a. X = the number of days a particular type of criminal trial will take b. $X \sim N(21, 7)$
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

81) a. mean = 5.51, $s = 2.15$ e. $X \sim N(5.51, 2.15)$ f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the actual data.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

83) μ = n*p = 100*.2 = 20; σ = sqrt(npq) = sqrt(100*.2*.8) = 4
a. 16 and 24 **b**. 95% **c.** 99.7%

85) **a**. 0.631 **b**. 0.0228 **c**. 30.8 oz and 32.4

87) a. 0.191 **b**. 0.434 **c**. 0.257 **d**. -3.4%

CHAPTER 7:

1) $\mu = 4$ hours; $\sigma = 1.2$ hours; $n = 16$.
3) a. Check student's solution. b. $P(3.5 < x < 4.25) = 0.2441$
5) The fact that the two distributions are different accounts for the different probabilities.
7) 0.3345 **9)** 7,833.92 **11)** 0.0089 **13)** 7,326.49 **15)** 77.45%
17) 0.4207 **19)** $\Sigma x = 3{,}888.5$ **21**) 0.8186 **23**) $\sigma_{\Sigma x} = 5$ **25**) 0.9772
27) The sample size n gets larger. **29**) 49 **31)** 26.00 **33**) 0.1587
35) $\mu_{\Sigma x} = 1{,}000$
37) a. $X \sim U(24, 26)$; $\mu = 25$ $\sigma = 0.5774$ b. $\overline{X} \sim N(25, 0.0577)$ c. 0.0416
39) 0.0003 **41**) 25.07 **43)** a. $\Sigma X \sim N(2{,}500, 5.7735)$ b. 0 **45**) 2,507.40
47) $\sigma = 10$; $m = 1/10 = 0.10$. **51)** 0.7799 **53**) 1.69 **55**) 0.0072
57) 391.54 **59)** 405.51
61) a. X = amount of change students carry b. $X \sim E(0.88, 0.88)$
c. $\bar{x}$ = mean amount of change for sample of 25 students. $\overline{X} \sim N(0.88, 0.176)$
e. 0.0819 f. 0.1882
g. The distributions are different. Part a is exponential and part b is normal.
63) a. X = length of time for an individual to complete IRS form 1040, in hours.
b. $\overline{X}$ = mean length of time for a sample of 36 taxpayers to complete IRS form 1040
c. $\overline{X} \sim N(10.53, 0.333)$
d. Yes, we would be surprised, since $P(\bar{x} > 12)$ is approximately 0.
e. No; would not be totally surprised because the probability is $P(x > 12) = 0.2312$. So about 23% of individuals take more than 12 hours to complete the form.
65) a. X = the length of a song, in minutes, in the collection b. $X \sim U(2, 3.5)$
c. $\overline{X}$ = the average length, in minutes, of the songs from a sample of five albums from the collection d. $\overline{X} \sim N(2.75, 0.066)$
e. Q_1 = invNorm(.25, 2.75, 0.066) = 2.71 minutes f. 0.09 minutes

67) a. True. By the CLT, the mean of a sampling distribution of the means is approximately the mean of the data distribution for large n.
b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
c. False. The standard deviation of the sampling distribution of the means decreases as the sample size increases; so $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ will be much smaller than σ.

69) a. X = the yearly income of someone in a third world country
b. $\overline{X}$ = the average salary from samples of 1,000 residents of a third world country
c. $\overline{X} \sim N(2000, 252.98)$
d. Very wide differences in data values can have averages smaller than standard deviations.
e. The distribution of the sample mean will have higher probabilities closer to the population mean.
71) b **73)** b **75)** a

77) a. 0 **b.** 0.101 **c.** 0.0139 **d.** 0.000395 **e.** 0.268

79) a. $52,330 b. $46,634

81) We are given that $\mu = 17$, $\sigma = 0.8$, $\bar{x} = 16.7$, and $n = 30$. To calculate the probability, we use $P(\bar{x} < 16.7) = \text{normalcdf}(-10\text{^}99, 16.7, 17, \frac{0.8}{\sqrt{30}}) = 0.020$. If the process is working properly, then there is only about a 2% probability that a sample of 30 batteries would have an average lifetime of 16.7 hours or less. So the sample data appears to be incompatible with the company's claim. Therefore, the class was justified to question the claim.

83) We are given that $\mu = 1$, $\sigma = 1$, $\bar{x} = 1.1$, and $n = 70$. We calculate the probability $P(\bar{x} < 1.1) = \text{normalcdf}(-10\text{^}99, 1.1, 1, \frac{1}{\sqrt{70}}) = 0.7958$. So there is an 80% chance that the average service time will be less than 1.1 hours. It would be wise to schedule more time since there is a 20% chance that the mean maintenance time will be greater than 1.1 hours.

85) Since $P(5.111 < \bar{x} < 5.291) = \text{normalcdf}(5.111, 5.91, 5.201, \frac{0.065}{\sqrt{280}}) \approx 1$, we can conclude that virtually all the coins are within the limits. We would not expect any coins to be rejected in a sample of 280.

87) 0.329

CHAPTER 8:

1) b. $1 - \alpha = .94$, $\alpha = .06$, $\alpha/2 = .03$; The critical values are $\pm z_{\alpha/2} = \pm\text{invnorm}(.03) = \pm 1.88$.

3) $1 - \alpha = .98$, $\alpha = .02$, $\alpha/2 = .01$; $t_{\alpha/2} = 2.55$; Margin of error is: $E = t_{\alpha/2}\frac{s}{\sqrt{n}} = .46$

5) a. $\bar{x} = 8.2$, $\sigma = 2.2$, $n = 200$

b. Using Zinterval: (7.944, 8.456); $\pm z_{\alpha/2} = \pm\text{invnorm}(.05) = \pm 1.645$; $E = .26$ minutes

c. Increase *n*

d. The level of confidence would decrease. The smaller the *n*, the larger the *E* but since E is staying the same, the confidence level would decrease.

e. The larger the confidence level, the larger *n* should be.

7). a. $\bar{x} = 30.4$, $\sigma = 15$, $n = 25$ b. Zinterval: (24.52, 36.28) c. μ d. $\alpha = .05$

e. $E = 36.28 - 30.4 = 5.88$ years

f. We are 95% confident that the true mean age for all Winger Foothill College students is between 24.52 and 36.28

g. The margin of error becomes smaller because as *n* increases the denominator in the formula for *E* formula becomes larger.

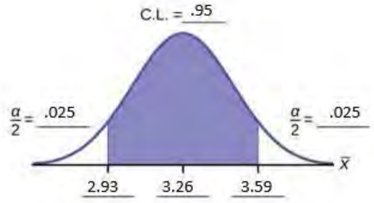
h. The margin of error for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean

9) a. (1.38, 1.62); $\pm t_c = \pm 1.995$; $E = 1.62 - 1.5 = .12$

b. 95% confident that the population mean for wait time is between 1.38 hours and 1.62 hours.

11) a. $\bar{x} = 3.26$; s = 1.02, n = 39 **b.** average number of colors on national flags

c. Tinterval (2.93, 3.59) **d.** 0.05 **e.**



f. 95% confidence that the population mean is between 2.93 and 3.59 colors for national flags.

g. Smaller, more accurate **h.** Smaller less confidence

13.) a. x = 120, n = 200, $\hat{p} = \frac{120}{200} = .6$ **b.** X number of households where women make the majority of the purchasing decisions. $\hat{p}$ is the proportion of household where women make the majority of the purchasing decisions. **c.** X ~ N(.6, .035) **d.** 1propZint: (.532, .668), E = .068 **e**. answers vary

15) a. 1propZint .627 < p < .673, E = .023 **b**. .623 < p < .677

17) a. Tinterval 70.19 < μ < 71.81 **b.** $\pm t_c = \pm 2.012$ **c.** the level of confidence would stay same but the interval would become closer together.

19) a. Z-distribution since σ is known **b.** Zinterval (22.45, 24.75) **c.** ±Zc = ±1.645

d. E = 1.15 hours **e.** Larger n **f.** Decrease **g**. n = 206

21) a. Zinterval (6.98, 8.82) **b**. Smaller margin of error, confidence interval is narrower.

23) a. $\bar{x}$ = $568,872.5 **b.** E = 281,758.5 **c.** ($287114, $850631) **d.** The population average of campaign contributions is between $287,114 and $850, 631

25) $n = \left(\frac{1.81*2.5}{1}\right)^2 = 21$

27) a. Tinterval (6243.4, 11014) **b.** E = 2385.3 **c.** $\bar{x}$ = 8628.7 **d.** $\pm t_c = \pm 2.03$

e. margin of error is smaller and confidence interval is closer together.

29) a. T distribution because sigma is unknown **b**. Tinterval (2.27, 2.76) **c.** $\pm t_c = \pm 2.31$

d. 95% confidence that the population mean lies within 2.27 hours and 2.76 hours.

31) a. $\bar{x}$ = $241,854.23 **b**. s = $521,130.4 **c**. Tinterval (47262, 456447)

33) a. $\bar{x}$ = 11.6, s = 4.1, n = 225 **b**. T distribution because sigma is unknown

c. Tinterval (11.12, 12.08)

35) a. Tinterval (7.64, 9.36) **b.** $\pm t_c = \pm 2.015$ **c.** E = .86 **d.** n is larger

37) a. Tinterval (12.32, 14.29) **b**. E = .99 oz

39) a. x = 320, n = 400, $\hat{p}$ = .80 **b.** 1propZint (.761, .839) **c.** E = 3.92% **d**. answer varies

41) a. X~N(.86, .0193) **b.** 1propZint (.823, .898) **c**. E = .038

43) a. X is the number of people over 50 who ran and died, $\hat{p}$ is the sample proportion of people over 50 who ran and died **b.** 1propZint (.00289, .0282) **c.** 97% chance the true proportion of those over 50 who ran and die is between .2% and 2.8%.

45) a. 1propZint 60% < p < 66% **b.** E = 3% **c**. ±3% represents the margin of error and that the true proportion of adult Americans who are worried a lot about the quality of education in our schools is between 60% and 66%.

47) a. 1propZint (.304, .373) **b.** $\hat{p}$ = .339

49) a. $\hat{p}$ = .52 E = .03 **b.** No because there is a chance that it can be 49% or 50%

c. 1propZint (.502, .538) **d.** yes

51) n = 385

53) a. 1propZint (.545, .596) **b.** n = 1068

CHAPTER 9:

1) a. The RV is $\bar{x}$ = the sample mean Internet speed in Megabits per second.

b. H_0: $\mu \leq 3$, Ha: $\mu > 3$

3) a. The RV is $\bar{x}$ = mean number of children from a random sample of American families.

b. H_0: $\mu = 2$, Ha: $\mu \neq 2$

5) a. H_0: $p = 0.42$ b. Ha: $p < 0.42$ **7)** a. H_0: $\mu = 15$ b. Ha: $\mu \neq 15$

9) Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

11) β = 1 – Power = 0.019 **13)** a. Power = 1 – β = 0.998 b. The Type II error

15) a normal distribution for a single population mean

17) It must be approximately normally distributed.

19) a. H_0: $\mu \geq 73$ Ha: $\mu < 73$ b. The *p*-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches. The data do support the claim.

21) a. The shaded region shows a low *p*-value.

b. The p-value is the area in the right tail, so it is a *right*-tailed test.

23) a. It is a test for a population mean. b. $\bar{x}$

c. the mean time spent in jail for 26 first time convicted burglars

d. $\sigma = 1.5$ e. $\bar{x} = 3$, $\sigma = 1.5$, $s = 1.8$, $n = 26$.

e. Since we know $\sigma = 1.5$, we should use it; using *s* instead would cause errors in the calculations. f. This will be a Z-test; $\bar{x} \sim N(2.5, 1.5/\text{sqrt}(26))$

25) This is a left-tailed test. **27)** This is a two-tailed test.

29) a right-tailed test **31)** a left-tailed test **33)** This is a left-tailed test.

35) a. H_0: $\mu = 34$; Ha: $\mu \neq 34$ b. H_0: $p \leq 0.60$; Ha: $p > 0.60$

c. H_0: $\mu \geq 100{,}000$; Ha: $\mu < 100{,}000$ d. H_0: $p = 0.29$; Ha: $p \neq 0.29$

e. H_0: $p = 0.05$; Ha: $p < 0.05$ f. H_0: $\mu \leq 10$; Ha: $\mu > 10$

g. H_0: $p = 0.50$; Ha: $p \neq 0.50$ h. H_0: $\mu = 6$; Ha: $\mu \neq 6$

i. H_0: $p \geq 0.11$; Ha: $p < 0.11$ j. H_0: $\mu \leq 20{,}000$; Ha: $\mu > 20{,}000$

37) c

39) a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years.

Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.

b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.

Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more than 60% do.

c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000.

Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.

d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%.

Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.

e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more.

Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.

f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10.
Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.

g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half.
Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.

h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks.
Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.

i. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%.
Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.

j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000.
Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

41) b **43)** d **45)** d

47) Hypotheses: H_0: $\mu \geq 50{,}000$, Ha : $\mu < 50{,}000$
Use a Z-test since $\sigma = 8000$ is known. From the TI-84, $z = -2.315$ and $p = 0.0103$.
Since $p < .05$, we reject the null hypothesis. There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

49) Hypotheses: H_0: $\mu = \$1.00$ b. Ha: $\mu \neq \$1.00$.
Use a Z-test since $\sigma = .20$ is known. From the TI-84, $z = -0.866$ and $p = 0.3865$.
Since $p > .01$, we do not reject the null hypothesis. There is not enough evidence to conclude that the mean cost of daily papers is different from $1. The mean cost could reasonably be $1.

51) Hypotheses: H_0: $\mu = 10$ Ha: $\mu \neq 10$
Use T-test because population SD is unknown. From TI-84, $t = -1.12$ and $p = 0.300$
Since $p > 0.05$, do not reject the null hypothesis. At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is different from 10.

53) Hypotheses: H_0: $p \geq 0.6$ Ha: $p < 0.6$
Use 1-PropZTest since testing a claim about a proportion.
From the TI-84, we have $z = 1.12$ and p-value = 0.1308.
Since $p > .05$, we do not reject Ho. There is not enough evidence to conclude that less than 60% of people who take Statistics feel more enriched.

55) Hypotheses: H_0: $\mu = 4$ Ha: $\mu \neq 4$
Use T-test because population SD is unknown. From TI-84, $t = 1.95$ and p-value = 0.076.
Since $p > 0.05$, we do not reject the null hypothesis. There is insufficient evidence to conclude that the average IQ of brown trout is different from four.

57) Hypotheses: H_0: $p \geq 0.13$ Ha: $p < 0.13$
Use 1-PropZTest since testing a claim about a proportion.
From the TI-84, we have $z = -2.688$ and p-value = 0.0036. Since $p < 0.05$, reject Ho.
There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

59) Hypotheses: H_0: $p = 0.40$ Ha: $p < 0.40$
Use 1-PropZTest since testing a claim about a proportion. From the TI-84, $z = -1.01$ and p-value = 0.1563. Since $p > 0.05$, do not reject Ho. There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.

61) Hypotheses: H_0: $p = 0.14$ Ha: $p < 0.14$
Use 1-PropZTest since testing a claim about a proportion. From the TI-84, $z = -0.2756$ and p-value = 0.3914. Since $p > 0.05$, do not reject Ho. There is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.

63) Hypotheses: H_0: $\mu = 69{,}110$ Ha: $\mu > 69{,}110$.
Use T-test because population SD is unknown. From TI-84, $t = 1.719$ and p-value = 0.0466. Since $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

65) c **67)** c

69) Hypotheses: H_0: $p = 0.488$, Ha: $p \neq 0.488$
Use 1-PropZTest since testing a claim about a proportion.
From the TI-84, $z = -1.01$ p-value = 0.0114. Since $p <$ 0.05, reject the null hypothesis. At the 5% level of significance, there is enough evidence to conclude that the percentage of families who own stocks is not 48.8%. So the survey does not appear to be accurate.

71) Hypotheses: H_0: $p = 0.517$, Ha: $p \neq 0.517$
Use 1-PropZTest since testing a claim about a proportion.
From the TI-84, p-value = 0.9203. Since $p > 0.05$, do not reject the null hypothesis.
At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively.

73) Hypotheses: H_0: $\mu \geq 11.52$, Ha: $\mu < 11.52$
Use T-test because population SD is unknown. From TI-84, p-value = 0.000002.
Since $p < .05$, Reject the null hypothesis. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average. Note that since the p-value is almost 0, we would reject Ho at *any* level of significance.

75) Hypotheses: H_0: $\mu \leq 5.8$ Ha: $\mu > 5.8$
Use T-test because population SD is unknown. From TI-84, p-value = 0.9987.
Since p is much larger than .05, we do not reject the null hypothesis. At the 5% level of significance, there is no evidence whatsoever to conclude that women visit their doctors an average of more than 5.8 times a year.

77) Hypotheses: H_0: $\mu \geq 150$ Ha: $\mu < 150$ (measuring in minutes)
Use T-test because population SD is unknown. From TI-84, p-value = 0.0622.
Since $p > 0.01$, do not reject the null hypothesis. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average. The student academic group's claim appears to be correct.

CHAPTER 10:

1) a. H_0: $p_1 = p_2$ Ha: $p_1 \neq p_2$ b. independent samples c. RV = $\hat{p}_1 - \hat{p}_2$
d. Two-proportion z-test.

3) a. H_0: $\mu_1 = \mu_2$ Ha: $\mu_1 \neq \mu_2$ b. independent samples c. RV = $\bar{x}_1 - \bar{x}_2$
d. Two-Sample z-test (because we know $\sigma_1 = \sigma_2$ = \$11,000)

5) a. H_0: $\mu_1 \leq \mu_2$ Ha: $\mu_1 > \mu_2$ (where μ_1 represents mean for teens)
b. independent samples c. RV = $\bar{x}_1 - \bar{x}_2$ d. Two-Sample t-test

7) a. H_0: $\mu_d \leq 0$ Ha: $\mu_d > 0$ (where d = # hrs sleep before – #hrs of sleep after)
b. dependent samples c. RV = $\bar{d}$ d. Matched pairs t-test

9) a. H_0: $\mu_1 = \mu_2$ Ha: $\mu_1 \neq \mu_2$ b. Use a t-test, because population SD's are unknown.
c. Using 2SampTTest, $t = 5.42$, $p = 0.00000023$ d. Reject Ho.
e. There is sufficient evidence to conclude that the mean life spans are different.

11) a. H_0: $\mu_1 \leq \mu_2$ Ha: $\mu_1 > \mu_2$ (where μ_1 represents mean for plants given the food)
b. Use a z-test, because population SD's are known.
c. Using 2SampZTest $z = 2.05$, $p = 0.0396$ d. Do not reject Ho since $p > .01$.
e. There is not enough evidence to conclude that food makes plants grow taller.

13) a. H_0: $p_1 = p_2$ Ha: $p_1 \neq p_2$ b. Two-Proportion z-test.
c. Using 2PropZTest $z = 1.28$, $p = 0.2016$ d. Do not reject Ho since $p > .05$.
e. There is not enough evidence to conclude that there is a difference in the crash rates.

15) a. The random variable is the sample mean of the differences between the number of failures before the patch and the number of failures after the patch: RV = $\bar{d}$
b. H_0: $\mu_d \leq 0$ Ha: $\mu_d > 0$ (where d = # failures before – #failures after)
c. This uses a t-test d. $t = 3.29$, $p = 0.0067$ e. Reject Ho since $p < .01$.
f. There is sufficient evidence to conclude that the average number of system failures is reduced after installing the patch.

17) a. H_0: $\mu_d \leq 0$ Ha: $\mu_d > 0$ (where d = pressure before meds – pressure after meds)
b. $t = 1.75$ c. $p = 0.0699$ d. Do not reject Ho since $p > .01$.
f. There is not enough evidence to conclude that the medication helped lower systolic blood pressure.

19) The hypotheses are H_0: $\mu_1 \leq \mu_2$, Ha: $\mu_1 > \mu_2$ (where μ_1 represents 4-yr schools)
Use 2-SampTTest, because population SD's are unknown.
From the calculator, $t = 0.248$, $p = 0.402$. Do not reject Ho since $p > .05$.
There is not enough evidence to conclude that the mean enrollment is higher for 4-yr schools.

21) The hypotheses are H_0: $\mu_1 \geq \mu_2$, Ha: $\mu_1 < \mu_2$ (where μ_1 represents mean salary for mechanical engineers). Use 2-SampTTest, because population SD's are unknown.
From the calculator, $t = -0.82$, $p = 0.2066$. Do not reject Ho since $p > .05$.
There is not enough evidence to conclude that the mean salary is lower for mechanical engineers.

23) b. **25)** b.

27) The hypotheses are H_0: $\mu_1 \geq \mu_2$ Ha: $\mu_1 < \mu_2$ (is the mean age for Canada).
Use 2-SampTTest. From the calculator, $t = -2.17$ and $p = 0.0157$.
Since $p > .01$, do not reject Ho. There is not enough evidence at the .01 level of significance to conclude that the mean age for entering prostitution is lower in Canada than in the U.S.

29) The hypothess are H_0: $\mu_1 \leq \mu_2$, Ha: $\mu_1 > \mu_2$ (where μ_1 is the mean for boys)
Use 2-SampZTest, since the population SD's are both known.
From the calculator, $z = 2.50$ and p-value: 0.0062. Since $p < 0.05$, Reject Ho.
At the 5% significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

31) The hypotheses are H_0: $\mu_1 \geq \mu_2$, Ha: $\mu_1 < \mu_2$ (where μ_1 is the mean for hybrid cars)
Use 2-SampZTest, since the population SD's are both known.
From the calculator, $z =$ 6.36 and $p = 0$. Since $p < .05$, we reject Ho.
At the 5% significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

33) The hypotheses are H_0: $\mu_d = 0$, Ha: $\mu_d < 0$ (d = wife's score – husband's score)
This is a matched-pairs t-test. From the calculator, $t = -1.86$ and $p = 0.0479$.
Since $p < .05$, we reject the null hypothesis. There sufficient evidence (just barely!) to conclude that the mean difference is negative. That is, at the .05 significance level, there is evidence that on average, husbands are happier with the division of child-care responsibilities.

35) The hypotheses are: H_0: $p_W = p_B$ Ha: $p_W \neq p_B$
We use 2-PropZTest; from the calculator we get $z = -\ 0.1944$ and p-value = 0.8458.
Do not reject Ho, since $p > .05$. There is absolutely evidence to conclude that the proportions of white and black female suicide victims, aged 15 to 24, are different.

37) The hypotheses would be Ho: $p_{2011} \leq p_{2010}$, Ha: $p_{2011} > p_{2010}$. Answer a.

39) The hypotheses are: H_0: $p_1 = p_2$, Ha: $p_1 \neq p_2$.
Use 2-PropZTest; from the calculator, $z = 4.29$ and p-value = 0.00002.
Since $p < .05$, we reject Ho. There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

41) The hypotheses are: H_0: $p_1 = p_2$, Ha: $p_1 \neq p_2$.
Use 2-PropZTest; from the calculator, $z = -2.94$ and p-value = 0.0033.
Since $p < .05$, we reject Ho. At the 5% level of significance, there is sufficient evidence to conclude that the proportion of eReader users aged 16 to 29 years is different from the proportion of eReader users 30 and older.

43) The hypotheses are: H_0: $p_1 \leq p_2$, Ha: $p_1 > p_2$ (where p_1 is the proportion for 16-29 yr olds)
Use 2-PropZTest; from the calculator, $z = -2.94$ and p-value = 0.2354.
Since $p > .01$, we do not reject Ho. At the 1% level of significance, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years and older.

45) The hypotheses are: H_0: $p_1 \geq p_2$, Ha: $p_1 < p_2$ (where p_1 is the proportion for male students)
Use 2-PropZTest; from the calculator, $z = -4.82$ and p-value = 0.0000007.
Since $p < .05$, we do reject Ho. At the 5% significance level, there is sufficient evidence to conclude that the proportions of male students with at least one pierced ear is less than the proportion for female students.

47) The hypotheses are H_0: $p_t \geq p_{ut}$, Ha: $p_t < p_{ut}$
(where p_t is the proportion of untreated patients who develop AIDS) Answer d.

49) The hypotheses are H_0: $\mu_d = 0$, Ha: $\mu_d > 0$ ($d =$ score before class – score after)
This is a matched-pairs t-test. From the calculator, $t = 1.19$ and $p = 0.8405$.
Since $p > .05$, we do not reject the null hypothesis. There is no evidence that the course helped improve scores.

51) The hypotheses are H_0: $\mu_d = 0$, Ha: $\mu_d \neq 0$ ($d =$ Hyatt price – Hilton price)
This is a matched-pairs t-test. From the calculator, $t = .41$ and $p = 0.6881$
Since $p > .05$, we do not reject the null hypothesis. There is no evidence that there is a difference in prices between the chains.

53) Let μ_1 = mean for girls, μ_2 = mean for boys. Using 2SampTInt, we are 95% confident that $.7224 < \mu_1 - \mu_2 < 2.2775$. So we are 95% confident that $.7224 + \mu_2 < \mu_1$.
There is evidence that the mean number for girls is at least .7224 more than the mean for boys.

55) Let μ_1 = mean for Cal State, μ_2 = mean for private universities. Using 2SampZInt, we are 95% confident that $-.1185 < \mu_1 - \mu_2 < .9185$. Since the interval contains both positive and negative values, this interval does not provide evidence that the mean time needed to graduate is higher for Cal State than it is for private schools.

57) Let p_1 = proportion for men, p_2 = women. Using 2PropZInt, we are 95% confident that $0.0022 < p_1 - p_2 < 0.2750$. Thus, there is evidence that $0.0022 + p_2 < p_1$. That is, we are 95% confident that the proportion for men is at least .0022 more than for women.

59) Let p_1 = proportion for millennials, p_2 = older generation. Ho: p1 = p2 (claim) Ha: p1 ≠ p2 Using 2PropZTest, test value Z = .92; p-value = .355, Do not Reject Ho. There is not enough evidence to reject the claim that millennials save at the same proportion as older generations.

CHAPTER 11:

1) $\mu = df = 25$; $s = \sqrt{2(df)} = \sqrt{50} \approx 7.071$.

3) The chi-square distribution approximates the normal distribution when df > 90.

5) $X^2_L = 0.7633$, $X^2_R = 36.191$.

7) a. Use a X^2 test for a single variance. b. Ho: $\sigma = 6$; Ha: $\sigma < 6$.
c. This is a left-tailed test.

9) a. Use a X^2 test for a single variance. b. $X^2 = 29(4.1)^2/(3.4)^2 = 42.17$
c. $p = X^2\text{cdf}(42.17, 10^9, 29) = 0.0542$. d. Since $p > .05$, do not reject Ho.
There is not quite enough evidence to conclude that the standard deviation of waiting times is greater than 3.4 minutes.

11) The hypotheses are Ho: $\sigma \leq 0.5$; Ha: $\sigma > 0.5$. Use a X^2 test for a single variance.
The test statistic is $X^2 = 83(.54)^2/(.5)^2 = 96.81$, and the p-value is
$p = X^2\text{cdf}(96.81, 10^9, 83) = 0.1426$. Since $p > .01$, we do not reject Ho.
There is not enough evidence to conclude that the machine needs recalibrating.

13) The hypotheses are Ho: $\sigma \leq 3$ (claim); Ha: $\sigma > 3$. Use a X^2 test for a single variance.
The test statistic is $X^2 = 4(3.194)^2/(3)^2 = 4.534$, and the p-value is
$p = X^2\text{cdf}(4.534, 10^9, 4) = 0.3385$. Since $p > .10$, we do not reject Ho.
There is not enough evidence to refute here claim.

15) The hypotheses are Ho: $\sigma \leq 0.75$; Ha: $\sigma > 0.75$. Use a X^2 test for a single variance.
The test statistic is $X^2 = 49(.789)^2/(.75)^2 = 54.23$, and the p-value is
$p = X^2\text{cdf}(54.23, 10^9, 49) = 0.2818$. Since $p > .01$, we do not reject Ho.
There is not enough evidence to conclude that the standard deviation is more than 0.75.

17) a. Round degrees freedom down to df = 40 to get $X^2_L = 26.509$ and $X^2_R = 55.758$.
Apply the formula to get the interval: $1.86 < \sigma < 2.69$. Since the lower bound of the interval is more than 1.5, this suggests that the standard deviation really is more than 1.5 oz as the manager suspected.
b. The hypotheses are Ho: $\sigma \leq 1.5$; Ha: $\sigma > 1.5$. Use a X^2 test for a single variance.
The test statistic is $X^2 = 48(2)^2/(1.5)^2 = 85.33$, and the p-value is
$p = X^2\text{cdf}(85.33, 10^9, 48) = 0.00073$. Since $p < .05$, we reject Ho.
There is sufficient evidence to conclude that the machine needs the standard deviation is greater than 1.5 oz.

19) Note that 5% of 150 is 7.5, so the hypotheses are Ho: $\sigma \leq 7.5$; Ha: $\sigma > 7.5$.
Use a X^2 test for a single variance; the test statistic is $X^2 = 14(10.43)^2/(7.5)^2 = 27.07$, and the p-value is $p = X^2\text{cdf}(27.07, 10^9, 14) = 0.0189$.
a. Since $p < .05$, we reject Ho. At the .05 level of significance, there is sufficient evidence to conclude that the standard deviation exceeds the weight tolerance.

b. Since $p > .01$, we **do not** reject Ho at the .01 level of significance. I.e. at the .01 level of significance, there is not enough evidence to conclude that the standard deviation exceeds the weight tolerance.

21) a. df = 3 (the number of categories minus 1)

b. Ho: $p_1 = .25,\ p_2 = .30,\ p_3 = .35,\ p_4 = .10$;
Ha: At least one of the proportions is different than stated.

c. According to Ho, the expected values are 5 A's, 6 B's, 7 C's and 2 D's, so $X^2 = 2.038$.

d. $p = X^2\text{cdf}(2.038, 10^9, 3) = 0.5645$. e. Do not reject Ho. There is not a significant difference between the observed distribution and the expected distribution.

23) Ho: p1 = .31, p2 = .41, p3 = .09, p4 = .04, p5 = .15; Ha: at least 1 proportion is different than stated. Test value: $X^2 = 32.43$; p-value = .00000156; Reject Ho; There is enough evidence to support the claim that the distribution has changed since 2016.

25) a. Ho: Distance and ticket class are independent. Ha: Distance and ticket class are dependent

b. df = (4)(2) = 8 **c**. 16.1 passengers expected **d**. 6.6 passengers expected

e. $X^2 = 15.92$ **f**. p-value = .0435 **g**. Reject Ho. There is enough evidence at 5% level of significance to reject the claim that distance and travel class are independent.

27) a. X^2 test, test for Homogeneity

b. Ho: class 1 and class 2 have the same distribution of test scores. Ha: class 1 and class 2 do not have the same distribution of test scores.

29) X^2 test, test for Homogeneity

31) X^2 test (test for independence)

33) Both use X^2 test with a contingency table

35) a. true **b.** false **c.** false (they are close in number but not exactly the same)

37) Expected frequency: 125.2, 224.4, 10, 40.4

39) Lake Tahoe frequency is expected values, Manhattan is observed values.

Ho: p1 = 9.2%, p2 = 8.3%, p3 = 73.6%, p4 = 5.6%, p5 = .8%, p6 = .6%, p7 = 1.7%
Ha: at least 1 proportion is not the same as stated
$X^2 = 2035$; p-value = 0; Reject Ho; There is enough evidence to reject that the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

41) a. $X^2 = 11.48$; p-value = .321; Do not reject Ho. There is not enough evidence to support that actual major fits the expected major for females. **b.** $X^2 = 7.18$; p-value = .709; Do not reject Ho. There is not enough evidence to support that actual major fits the expected major for males.

43) Ho: Surveyed obese fit the distribution of expected obese
Ha: Surveyed obese does not fit the distribution of expected obese

$X^2 = 54.01$; p-value = 0 Reject Ho. Surveyed obese does not fit the expected obese.

45) Ho: Family size and size of car are independent; $X^2 = 15.83$; p-value = .071. Do not reject Ho. There is not enough evidence to reject the claim that they are independent.

47) Ho: Location of honeymoon and Age are independent; $X^2 = 15.7$; p-value = .0734; Reject Ho. Location of honeymoon and Age are dependent.

49) Ho: type of fries and area of US sold are independent; $X^2 = 18.84$; p-value = .00445; Reject Ho. They are dependent.

51) Ho: Annual Salary and level of education are independent. $X^2 = 2558$; p = 0; Reject Ho. They are related.

53) Ho: Geographic location and favorite ice cream flavor are independent. $X^2 =14.06$; p-value = .521. They are independent.

55) Ho: Opinion Response and Ethnicity are independent; $X^2 =48.6$; p-value = 8.97E-9; Reject Ho. They are dependent.

57) Ho: Men and women have the same distribution for breakfast selection; $X^2 = 4.01$; p-value = .26; We can't reject that they have the same distribution.

59) Ho: Applicable reasons and Most important reasons have the same distribution; $X^2 = 234$; Reject Ho. They do not have the same distribution.

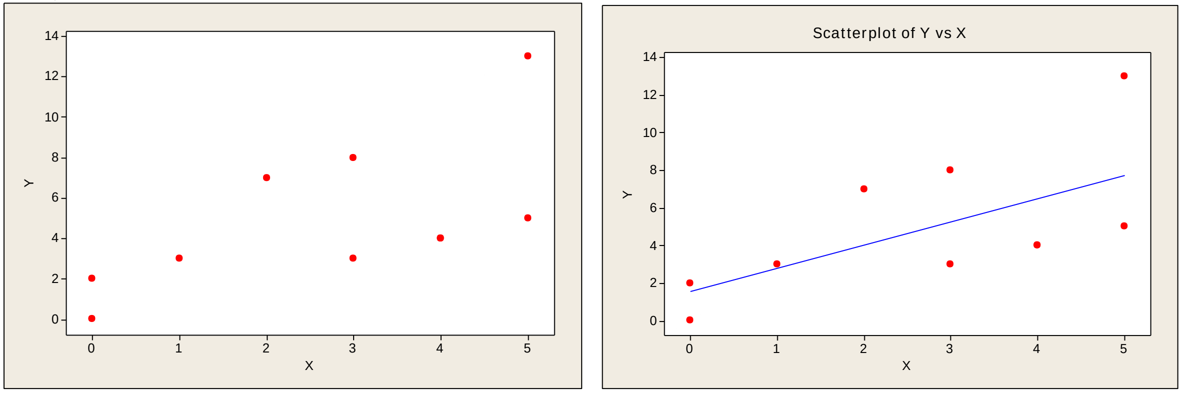
61) Ho: 2009 and 2013 distribution of cars is the same; $X^2 =6.65$; p-value = .248; Do not reject Ho

CHAPTER 12:

1) a. The independent variable is x = number of hours worked;
the dependent variable is y = total charge.
b. y-intercept is $a = 50$. This is the equipment fee (a fixed cost)
c. The slope is $b= 100$. The labor fee is \$100/hr.

3) a. Strong positive correlation b. Fairly strong negative correlation
c. No linear relationship d. Non-linear relationship.

5. a, c



b. The regression equation is: $y = 1.56 + 1.24\, x$.
d. The slope is $b = 1.24$. We expect athletes to earn about \$1.24 million per endorsement.
e. The intercept is $a = 1.56$. An athlete with no endorsements is expected to earn about \$1.56 million.

7) The critical value for $n = 100$ is CV = .205. Since $| r | = .89 >$ CV, the correlation is significant.

9) Given $\hat{y} = 101.32 + 2.48x$.
a. When $x = 60$, $y = 250.12$ b. When $x = 90$, $y = 324.52$

11) $\hat{y} = 6.39 + 0.015\, x$; $r = 0.094$ **13)** $\hat{y} = 31.53 + 10.90x$; $r = 0.911$

15) Hypotheses are: Ho: $\rho = 0$ Ha: $\rho \neq 0$. From the calculator, $t = 0.46$ and $p = 0.6693$.
Since $p > .05$, we do not reject Ho. The correlation is not significant.

17) $\hat{y} = -3{,}448{,}225 + 1750\, x$.

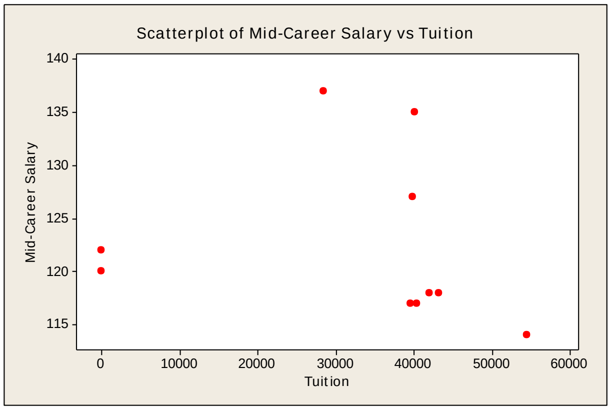
19) Hypotheses are Ho: $\rho = 0$ Ha: $\rho \neq 0$. From the calculator, $t = 2.27$ and $p = 0.0344$.
Since $p < .05$, we reject Ho. There is a significant correlation.
Alternatively, the correlation coefficient is $r = 0.4526$, and the critical value is CV = 0.423.
Since $r > 0.423$, the correlation is is significant.

21) a. When $x = 1985$, $\hat{y} = 25{,}525$ b. When $x = 1990$, $\hat{y} = 34{,}275$

23) Visually, the line does not appear to be a good fit to the data; however, the hypothesis test indicates a significant correlation, so using the regression line for point predictions is justified. However, some caution should be exercised, and the equation should be used *only* in the range of observed x-values.

25) a. Yes, there appears to be an outlier at (6, 58).
b. The fact that the correlation coefficient improved significantly when the point is removed suggests that this really was an outlier.
c. The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate is a predictor for the data.
d. We would be more confident in the predictive ability of the new regression line.

27) a.



b. r = -0.148. The critical value is CV = .632. Since | r | < .632, there is not a significant correlation between tuition and mid-career salary.

c. The two military institutions appear to be outliers (these have zero tuition)

d. Removing these two points, the new correlation coefficient is r = -0.715, and the new critical value is CV = .707. Since we now have | r | > .707, the correlation is significant.

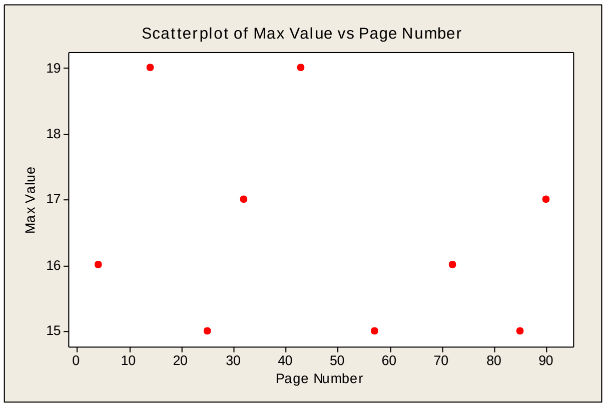
29) a. $\hat{y} = 35.58 - 0.192\,x$

b. r = –0.579; LinRegTTest gives p-value = 0.2288 so we would not reject the null hypothesis. I.e. the correlation is not significant.

c. There is not a significant linear relationship between deaths and age of driver.

31) a. Independent variable is x = page number; dependent variable is y = max value of discount.

b.



c. $\hat{y} = 17.218 - 0.014\,x$ d. r = – 0.275; from the table, CV = 0.666. Since $|r|$ < CV, there is not a significant correlation between page number and maximum value of discount.

e. Since there is not a significant correlation, we should not use the regression equation for estimation.

33) b. $\hat{y} = 31.93 - 0.304\,x$.

c. The slope of the regression line is -0.304 with a y-intercept of 31.93. The y-intercept indicates that when there are no returning sparrow hawks, there will be almost 32% new sparrow hawks; this doesn't make sense since because if there are no returning birds, then the new percentage would have to be 100% (this shows the dangers of extrapolation). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 0.304%.

d. The coefficient of determination is $r^2 = .560$ which means that about 56% of the total variation in the percent of new birds is explained by the model. and the correlation coefficient, $r = 0.71$, indicates a moderately strong correlation between returning and new percentages.

e. The ordered pair (66, 6) generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.87%. However, the standard error is $s = 3.66$, and so this residual is less than two standard errors from the predicted value; hence the point is not an outlier.

f. If there are 70% returning birds, we would expect to see $y = -0.304(70) + 31.93 = 10.65\%$ new birds in the colony.

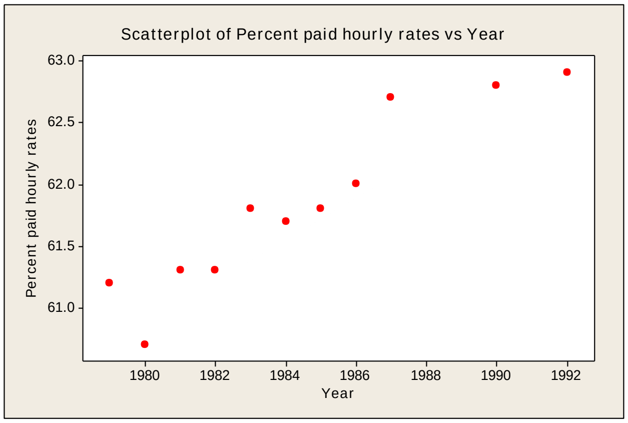
35) b. $\hat{y} = 193.88 - 1.495\,x$.

c. We have a slope of –1.495 with a y-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the y-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim 2,000 meters, the less effort the heart puts out), the y-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.

d. The coefficient of determination is $r^2 = .015$ which means that only 1.5% of the variation in heart rate is explained by this regression equation. So the model is a very poor fit to the observed data, and we conclude that there is not a linear relationship between swim time and heart rate.

e. The point (34.72, 124) generates the largest residual of –11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. The standard error is $s = 10.02$ – since the residual is less than two standard errors, this point is not considered an outlier. When this point is removed, the coefficient of determination increases to .098, which is better, but still indicates a poor fit to the data. So the point is not influential either.

37) a



b. Yes, the scatterplot exhibits a fairly strong linear pattern. c. $\hat{y} = -266.9 + 0.166\,x$

d. $r = 0.944$; the correlation is significant. e. When $x = 1991$, $y = 62.82$; when $x = 1988$, $y = 62.33$.

f. Yes, the scatterplot, correlation coefficient and hypotheses test all indicate a linear relationship.

g. Yes; the point (1987, 62.7) gives a residual of .906. The standard error is $s = .247$, so the observed y-value is more than three standard errors from the predicted y-value.

h. When $x = 2050$, we would get $y = 72.59$. But this would not be an accurate prediction, since $x = 2050$ is too far outside the range of observed data for a reliable prediction.
i. The slope is $b = 0.1656$. The percent of workers paid hourly rates tends to increase by 0.1656 (1/6 of 1%) each year.

39) a. $\hat{y} = -745.25 + 54.76\,x$ b. $r = 0.8944$; yes it is significant.
c. When $x = 32$, $y =$ \$1006.93; when $x = 50$, $y =$ \$1992.53.
d. For each additional inch, the price increases by \$54.76.

41) a. The independent variable is x = age; the dependent variable is y = height.
b. $\hat{y} = 65.09 + 7.09\,x$
c. The correlation coefficient is $r = 0.976$. Test the hypotheses Ho: $\rho = 0$ Ha: $\rho \neq 0$.
From the calculator, $t = 10.05$ and $p = 0.00017$. Since $p < .05$, we reject Ho.
The correlation is significant.
d. When $x = 1$, $y =$ 72.18 cm; when $x = 11$, $y = 143.8$ cm.
e. The slope is $b = 7.09$. On average, boys grow by about 7.09 cm per year.
f. Residuals: -14.29, 4.52, 5.03, 6.04, 4.55, 1.06, -6.92

CHAPTER 13:

1) i) The populations from which we are sampling are normally distributed.
ii) The samples are randomly selected and independent of one another.

3) From the calculator with 2SampFTest, $F = .72$ and $p = .4810$
Note: If the subscripts are switched, then we get $F = 1.39$, with the same p-value.

5) a. Ho: $\sigma_1 \leq \sigma_2$; Ha: $\sigma_1 > \sigma_2$. b. Use a two-sample F-test (test for two variances)
c. From the calculator, $F = 2.87$ and $p = 0.0291$. d. Since $p < .05$, reject Ho.
e. There is sufficient evidence to conclude that the second student's scores are less variable.

7) i) Each population from which we are sampling is normally distributed.
ii) All samples are random and independent of one another.
iii) The populations are assumed to have equal variances.
iv) The factor is a categorical variable
v) The response is a numerical variable.

9) Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; Ha: At least one of the means is different from the others.

11) a. $SS_{\text{between}} = 25.75$ b. $\text{df}_{\text{num}} = 3$ c. $MS_{\text{between}} = 8.58$ d. $SS_{\text{within}} = 13.2$
e. $\text{df}_{\text{denom}} = 16$ f. $MS_{\text{within}} = .825$ g. $F = 10.40$
h. The test is right-tailed, so a large F-statistic suggests that Ho would be rejected.

13) a. Ho: $\mu_1 = \mu_2 = \mu_3$; Ha: At least one of the means is different from the others.
b. $SS_{\text{between}} = 5700.4$ and $MS_{\text{between}} = 2850.2$
c. $SS_{\text{within}} = 9474$ and $MS_{\text{within}} = 789.5$ d. $F = 3.61$ e. $p = 0.0592$
f. Since $p < .10$, we reject Ho. There is sufficient evidence to conclude that there is a difference in the mean scores for the different groups.

15) Ho: $\mu_1 = \mu_2 = \mu_3$; Ha: At least one of the means is different from the others.
Use ANOVA F-test. From the calculator, $F = 0.67$ and $p = 0.5305$.
Since $p > .05$, we do not reject Ho. There is not enough evidence to conclude that there is a difference in the mean weight gains among the three groups of rats.

17) Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; Ha: At least one of the means is different from the others.
Use ANOVA F-test; from the calculator, $F = 8.69$ and $p = .0012$. Since $p < .05$, we reject Ho. There is sufficient evidence to conclude that there is a difference in the mean number of pages for different magazine types.

19) Ho: $\mu_1 = \mu_2 = \mu_3$; Ha: At least one of the means is different from the others.
Use ANOVA F-test; from the calculator, $F = .64$ and $p = .5437$. Since $p > .05$, we do not reject Ho. There is not enough evidence to reject the claim that the mean final exam scores are the same for all three delivery types.

21) Ho: $\mu_1 = \mu_2 = \mu_3$; Ha: At least one of the means is different from the others.
Use ANOVA F-test. From the calculator, $F = 3.10$ and $p = 0.082$.
Since $p > .05$, we do not reject Ho. There is not enough evidence to reject the claim that the mean number of visitors is the same for three types of snow conditions.

23) Ho: $\sigma_1^2 = \sigma_2^2$; Ha: $\sigma_1^2 \neq \sigma_2^2$.
Use 2SampFTest; from the calculator, we get $F = .33$ and $p = 0.3127$.
Since $p > .05$, do not reject Ho. There is not enough evidence to conclude that there is a difference in the variances among Javier's and Linda's rats.
NOTE: Here we used σ_1^2 to represent the variance for Linda's rats, and σ_2^2 to represent the variance for Javier's rats. Had these been entered in the opposite order, we would have gotten $F = 3.0$; however, because it is a two-tailed test, the p-value would be the same.

25) Ho: $\sigma_1^2 = \sigma_2^2$; Ha: $\sigma_1^2 \neq \sigma_2^2$.

Use 2SampFTest; from the calculator, we get $F = 12.69$ and $p = 0.0305$.

Since $p < .05$, do reject Ho. There is sufficient evidence to conclude that the variance is different for the two types of magazines.

Again, had we switched the subscripts, we would get a different F-statistic, $F = 0.079$. However, because it is a two-tailed test, the p – value would remain the same.

27) Ho: $\sigma_1^2 = \sigma_2^2$; Ha: $\sigma_1^2 \neq \sigma_2^2$.

Use 2SampFTest; from the calculator, we get $F = 0.812$ and $p = 0.7825$.

Since $p > .05$, do reject Ho. There is not enough evidence to reject the claim that the variances for incomes are the same on the two coasts.

29) Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$; Ha: At least one of the means is different from the others.

Use ANOVA F-test. From the calculator, $F = 26.27$ and $p = 0.0000001$.

Since $p < .05$, we reject Ho. There is very strong evidence that the mean silver content of the different coinages are not all the same.

31) a.

Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	Test Value (F)	p-value
Between (Factor)	32.138	3	10.713	6.512	.00298
Within (Error)	32.901	20	1.645		
Total	65.039	23			

b. Ho: μ1 = μ2 = μ3 = μ4; Ha: at least 1 mean is different; F = 6.512; p-value = .00298; Reject Ho. There is enough evidence to support claim that the means are different.

Back Cover