Business Statistics I -- MGMT 2262 -- Mt Royal University -- Version 2016 Revision A

Collection edited by: Claude Laflamme
Content authors: Claude Laflamme, Alexander Holmes, and OpenStax
Based on: Business Statistics -- BSTA 200 -- Humber College -- Version 2016RevA -- DRAFT 2016-04-04 <**http://legacy.cnx.org/content/col11969/1.5**>.
Online: <**http://legacy.cnx.org/content/col11990/1.5**>

Collection structure revised: 2016/08/28
PDF Generated: 2016/09/18 11:39:27

Table of Contents

PREFACE-- MGMT 2262 -- MT ROYAL UNIVERSITY -- VERSION 2016REVA

About *Introductory Statistics*

This version of *Introductory Statistics* MGMT 2262 has been adapted specifically for Mount Royal University, and is designed for students majoring in fields other than math or engineering. This text assumes students have been exposed to intermediate algebra, and it focuses on the applications of statistical knowledge rather than the theory behind it.

Coverage and Scope

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 The Normal Distribution
Chapter 6 The Central Limit Theorem
Chapter 7 Confidence Intervals
Chapter 8 Hypothesis Testing with One Sample
Chapter 9 Linear Regression and Correlation

Pedagogical Foundation and Features

- **Examples** are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics; these include examples about college life and learning, health and medicine, retail and business, and sports and entertainment.
- **Practice, Homework, and Bringing It Together** problems give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Credits

Adaptation for Mount Royal University

Editorial Group, Lyryx Learning

Authors of original collection

Barbara Illowsky De Anza College
Susan Dean De Anza College

Adaptation at University of Oklahoma

Alexander Holmes Regent's Professor of Economics University of Oklahoma
Kevin Hadley Analyst, Federal Reserve Bank of Kansas City
Mathew Price Research Assistant, University of Oklahoma

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 OBJECTIVE

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

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

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

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

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

1.1 | Definitions of Statistics, Probability, and Key Terms - - MtRoyal - Version2016RevA

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

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

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

Probability

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

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

Key Terms

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

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

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

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

A **variable**, or random variable, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical

variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

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

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

NOTE

Though the words " **mean**" and " **average**" are sometimes used interchangeably, they do not necessarily mean the same thing. In general, “average” is any centre of location and “mean” is a specific type of centre. Many people use average and mean as the same, but not always. For example, when people talk about average housing price, they are usually referring to the median house price.

Example 1.1

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

Solution 1.1

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

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

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

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

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

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

Try It

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

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

Solution 1.2

1. f; 2. g; 3. e; 4. d; 5. b; 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. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Table 1.1

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** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

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

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

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

The **data** are either: yes, had head injury, or no, did not.

Example 1.4

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 1.4

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

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

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

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

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

1.2 | Data, Sampling, and Variation -- MtRoyal - Version2016RevA

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values.

Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data (also called qualitative data) are the result of categorizing or describing attributes of a population. Hair colour, 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 colour might be black, dark brown, light brown, blonde, grey, 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 or colour or blood type.

There are two types of categorical data: nominal and ordinal. **Nominal data** is categorical data that cannot be ordered in a meaningful way. For example, the colour of a car is categorical, but the order of the colours are not meaningful.

Ordinal data is categorical data that can be ordered in a meaningful way. For example, the level of satisfaction someone has with their experience at a restaurant from not at all satisfied to completely satisfied.

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

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

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring time, distance, area, and so on; anything that can be subdivided and then subdivided again and again is a continuous variable. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example 1.5 Data Sample of Quantitative Discrete Data

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

Try It

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?

Example 1.6 Data Sample of Quantitative Continuous Data

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

Try It

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

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 two pounds (32 ounces chocolate chip cookies).

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

Solution 1.7

One Possible Solution:

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

Try to identify additional data sets in this example.

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.

Try It

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

NOTE

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

Try It

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

Example 1.9

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 1.1**. What type of data does this graph show?

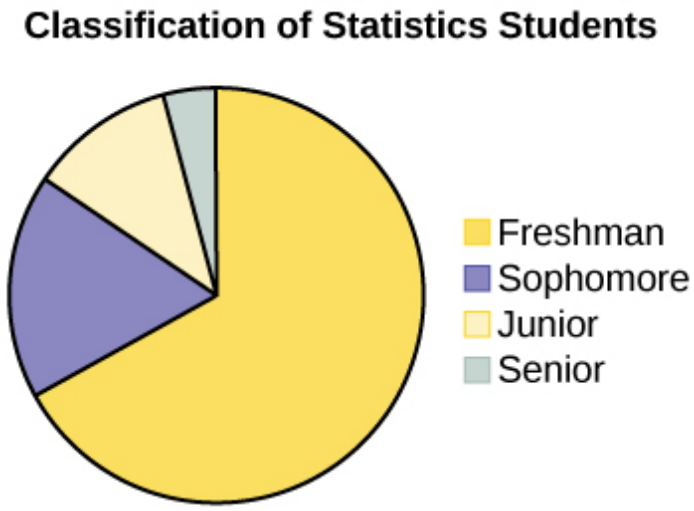


Figure 1.2

Solution 1.9

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

Try It

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

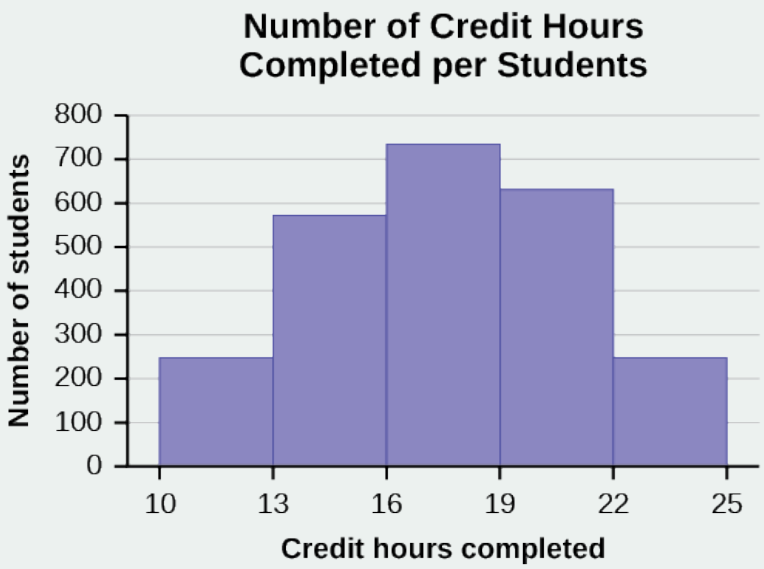


Figure 1.3

What type of data does this graph show?

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

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

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

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

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

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

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

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

If you are sampling **without replacement**,

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

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

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

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

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

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

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

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

Example 1.10

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

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.

b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.

c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.

d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.

e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Solution 1.10

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

Example 1.11

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

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

b. A pollster interviews all human resource personnel in five different high tech companies.

c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.

d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution 1.11

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

Try It

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

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

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

Example 1.12

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

Suppose we take two different samples.

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

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

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

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

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?

Solution 1.12

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are,

more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

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

Solution 1.12

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

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

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

c. Is the sample biased?

Solution 1.12

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

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

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

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

Variation in Data

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

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

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

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

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide

to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

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

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

Size of a Sample

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

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

Critical Evaluation

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

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

1.3 | Experimental Design and Ethics -- MtRoyal - Version2016RevA

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

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

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

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

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

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

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

Example 1.13

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 1.13

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.

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.

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.

KEY TERMS

Average also called mean or arithmetic mean; 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; 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 used in statistical research can be put into two groups: **categorical** (an attribute whose value is a label) or **quantitative** (an attribute whose value is indicated by a number). Categorical data can be separated into two subgroups: **nominal** and **ordinal**. Data is nominal if it cannot be meaningfully ordered. Data is ordinal if the data can be meaningfully ordered. 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

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

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

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

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

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

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

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

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 Definitions of Statistics, Probability, and Key Terms -- MtRoyal - Version2016RevA

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

1.2 Data, Sampling, and Variation -- MtRoyal - Version2016RevA

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 -- MtRoyal - Version2016RevA

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

"An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule."[2] 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.

HOMEWORK

1.1 Definitions of Statistics, Probability, and Key Terms -- MtRoyal - Version2016RevA

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.

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

2. 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.

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

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

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

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

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

8. 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.

9. 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

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

10. Consider the following:

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.

11. 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.

1.2 Data, Sampling, and Variation -- MtRoyal - Version2016RevA

For the following exercises, 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.

12. number of tickets sold to a concert

13. percent of body fat

14. favorite baseball team

15. time in line to buy groceries

16. number of students enrolled at Evergreen Valley College

17. most-watched television show

18. brand of toothpaste

19. distance to the closest movie theatre

20. age of executives in Fortune 500 companies

21. number of competing computer spreadsheet software packages

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

22. "Number of times per week" is what type of data?
a. nominal qualitative ordinal
b. quantitative discrete
c. quantitative continuous
d. categorical nominal
e. categorical ordinal

23. "Duration (amount of time)" is what type of data?
a. nominal qualitative ordinal
b. quantitative discrete
c. quantitative continuous
d. categorical nominal
e. categorical ordinal

24. 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.

25. 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.

26. 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.

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

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

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

30. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

31. 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 Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

32. 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.

33. 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 percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are 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.

34. 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: categorical nominal, categorical ordinal, 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?

35. 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 nonsampling 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?

36. 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 **m54036 (http://legacy.cnx.org/content/m54036/latest/)** could explain this connection?

37. 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?”[3]

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?

38. 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.”[4]

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.”[5]

a. What are some 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.

---

3. 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).
4. Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (http://poq.oxfordjournals.org/content/70/5/759.full) (accessed May 1, 2013).
5. 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 Definitions of Statistics, Probability, and Key Terms -- MtRoyal - Version2016RevA

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

1.2 Data, Sampling, and Variation -- MtRoyal - Version2016RevA

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 May 1, 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 -- MtRoyal - Version2016RevA

"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).

SOLUTIONS

2

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

4

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

6

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

8

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

10 a

12 quantitative discrete, 150

14 qualitative, Oakland A's

16 quantitative discrete, 11,234 students

18 qualitative, Crest

20 quantitative continuous, 47.3 years

22 b

24

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.

26 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.

28 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.

30 b

32 convenience; cluster; stratified ; systematic; simple random

34

a. categorical nominal

b. quantitative discrete

c. quantitative discrete

d. categorical nominal

36 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.

38

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.

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 OBJECTIVE

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

- Display data graphically and interpret graphs: stemplots, histograms and box plots.
- Recognize, describe, calculate, and interpret measures of location: quartiles and percentiles.
- Recognize, describe, calculate, and interpret measures of centre: mean, median and mode.
- Recognize, describe, calculate, and interpret measures of variation: variance, standard deviation, range, interquartile range and coefficient of variation.

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.

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.

2.1 | Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA

Categorical Data Descriptions

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

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

Table 2.1 Fall Term 2007 (Census day)

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

In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

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

Look at **Figure 2.2** and **Figure 2.3** 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.

Subfigure (a) (fig-ch01_patchfile_01.jpg)
(a)

Subfigure (b) (fig-ch01_patchfile_02.jpg)
(b)

Figure 2.2

Figure (fig-ch01_patchfile_03.jpg)

Figure 2.3

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

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

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

Table 2.2 De Anza College Spring 2010

Figure (fig-ch01_patchfile_04.jpg)

Figure 2.4

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%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 2.3 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

Figure (fig-ch01_patchfile_05.jpg)

Figure 2.5

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 2.6** can be difficult to understand visually. The graph in **Figure 2.7** is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

Bar Graph with Other/Unknown Category (fig-ch01_patchfile_06.jpg)

Figure 2.6 Bar Graph with Other/Unknown Category

Pareto Chart With Bars Sorted by Size (fig-ch01_patchfile_07.jpg)

Figure 2.7 Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The chart in **Figure 2.8b** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in **Figure 2.8a**.

Subfigure (a) (fig-ch01_patchfile_08.jpg)
(a)

Subfigure (b) (fig-ch01_patchfile_09.jpg)
(b)

Figure 2.8

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

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

Example 2.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem	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 2.4 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% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Try It

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

Example 2.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:
1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

The leaves are to the right of the decimal.

Solution 2.2

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 2.5

Try It

2.2 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 2.3

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 2.7** and **Table 2.8** show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution 2.3

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

Table 2.6

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 2.7 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90

Table 2.8 Presidential Age at Death

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

Table 2.8 Presidential Age at Death

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in **Example 2.4**, 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 2.4

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 2.9** and in **Figure 2.9**.

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

Table 2.9

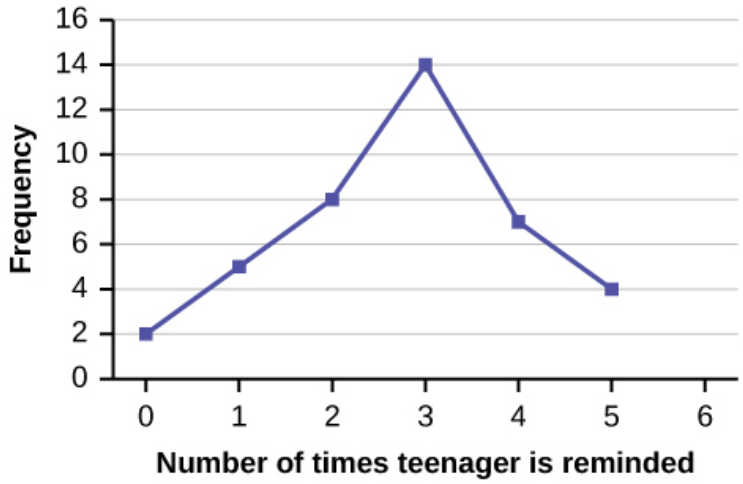


Figure 2.9

Try It

2.4 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 2.10**. Construct a line graph.

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

Table 2.10

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in **Example 2.5** has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.5

By the end of 2011, Facebook had over 146 million users in the United States. **Table 2.10** shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

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

Table 2.11

Solution 2.5

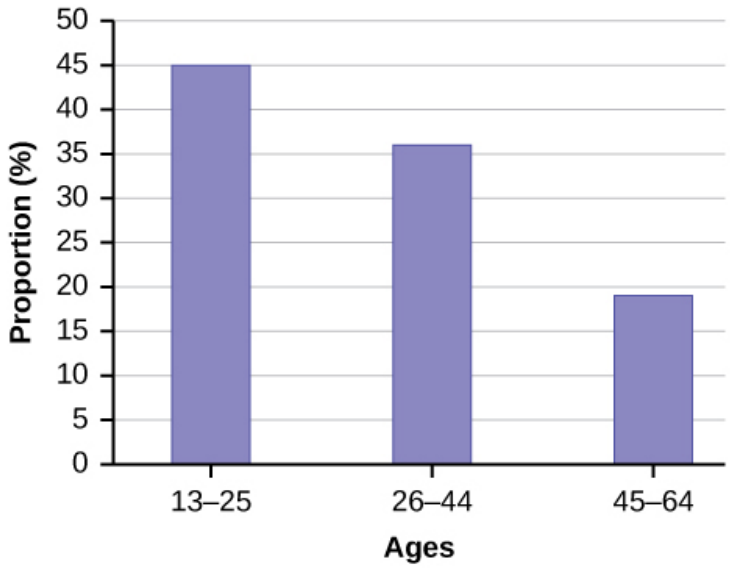


Figure 2.10

Try It

2.5 The population in Park City is made up of children, working-age adults, and retirees. **Table 2.12** shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Table 2.12

Example 2.6

The columns in **Table 2.12** contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

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

Table 2.13

Solution 2.6

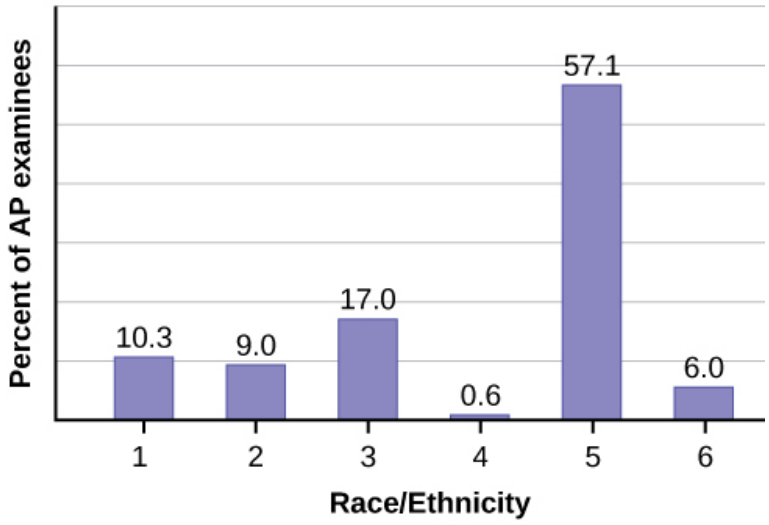


Figure 2.11

Try It

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

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

Table 2.14

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table 2.15 lists the different data values in ascending order and their frequencies.

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

Table 2.15 Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to **Table 2.15**, 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.

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

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

Table 2.16 Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of **Table 2.16** is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in **Table 2.17**.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15

Table 2.17 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

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

Table 2.17 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

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

NOTE

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

Table 2.18 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
60–61.99	5	$\frac{5}{100}$ = 0.05	0.05
62–63.99	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08
64–65.99	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
66-67.99	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
68–69.99	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
70–71.99	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
72–73.99	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99
74–75.99	1	$\frac{1}{100}$ = 0.01	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Table 2.18 Frequency Table of Soccer Player Height

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

- 60 to 61.99 inches
- 62 to 63.99 inches
- 64 to 65.99 inches
- 66 to 67.99 inches
- 68 to 69.99 inches
- 70 to 71.99 inches
- 72 to 73.99 inches
- 74 to 75.99 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example 2.7

From **Table 2.18**, find the percentage of heights that are less than 65.95 inches.

Solution 2.7

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

Try It

2.7 **Table 2.19** shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
3–4.99	6	$\frac{6}{50}$ = 0.12	0.12
5–6.99	7	$\frac{7}{50}$ = 0.14	0.12 + 0.14 = 0.26
7–9.99	15	$\frac{15}{50}$ = 0.30	0.26 + 0.30 = 0.56
10–11.99	8	$\frac{8}{50}$ = 0.16	0.56 + 0.16 = 0.72
12–12.99	9	$\frac{9}{50}$ = 0.18	0.72 + 0.18 = 0.90
13–14.99	5	$\frac{5}{50}$ = 0.10	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

Table 2.19

From **Table 2.19**, find the percentage of rainfall that is less than 9.01 inches.

Example 2.8

From **Table 2.18**, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution 2.8
Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Try It

2.8 From **Table 2.19**, find the percentage of rainfall that is between 6.99 and 13.05 inches.

Example 2.9

Use the heights of the 100 male semiprofessional soccer players in **Table 2.18**. Fill in the blanks and check your answers.

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

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

Solution 2.9

a. 29%

b. 36%

c. 77%

d. 87

e. quantitative continuous

f. get rosters from each team and choose a simple random sample from each

Example 2.10

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. **Table 2.20** was produced:

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

Table 2.20 Frequency of Commuting Distances

a. Is the table correct? If it is not correct, what is wrong?

b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.

c. What fraction of the people surveyed commute five or seven miles?

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

Solution 2.10

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.

b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.

c. $\frac{5}{19}$

d. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

Try It

2.10 **Table 2.19** represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Example 2.11

Table 2.21 contains the total number of deaths worldwide as a result of earthquakes for the period 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	823,356

Table 2.21

Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?

b. What percentage of deaths occurred after 2009?

c. What is the relative frequency of deaths that occurred in 2003 or earlier?

d. What is the percentage of deaths that occurred in 2004?

e. What kind of data are the numbers of deaths?

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

Solution 2.11

a. 97,118 (11.8%)

b. 41.6%

c. 67,092/823,356 or 0.081 or 8.1 %

d. 27.8%

e. Quantitative discrete

f. Quantitative continuous

Try It Σ

2.11 **Table 2.22** contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

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

Table 2.22

Answer the following questions.

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

Histograms, Frequency Polygons, and Time Series Graphs

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

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

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

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

then:

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

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 $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

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

Example 2.12

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.
60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

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

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

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

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

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

NOTE

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

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95

- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

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

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

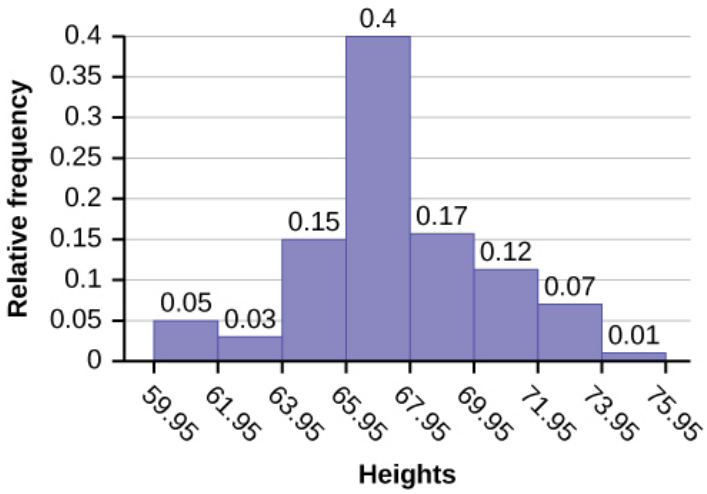


Figure 2.12

Try It

2.12 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

Example 2.13

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.
1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

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

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

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

Solution 2.13

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

Calculate the number of bars as follows:

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

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

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

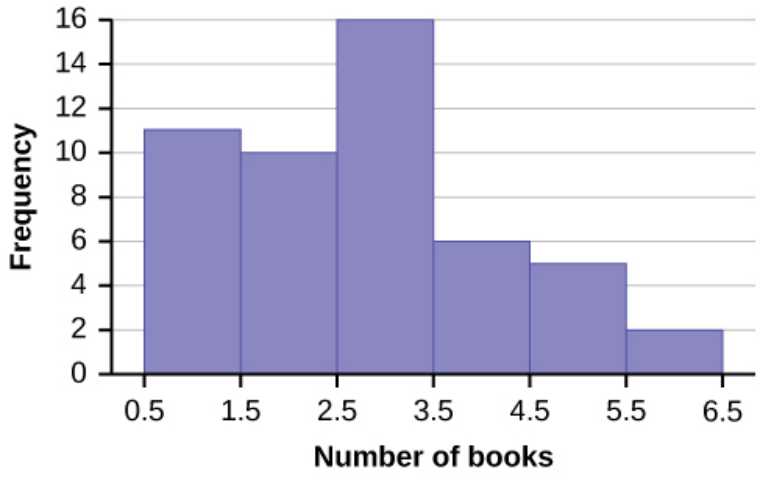


Figure 2.13

Example 2.14

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Table 2.23

Solution 2.14

Figure 2.14

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

Frequency Polygons

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

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 2.15

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.24

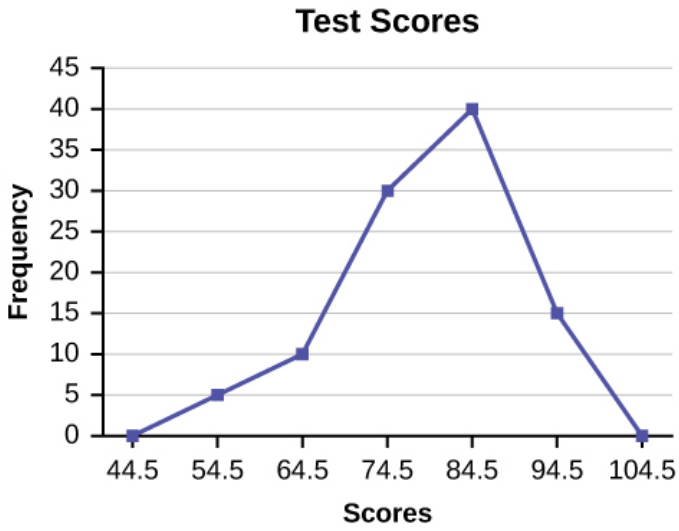


Figure 2.15

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.

Try It

2.15 Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in **Table 2.25**.

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

Table 2.25

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

Example 2.16

We will construct an overlay frequency polygon comparing the scores from **Example 2.15** with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.26

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

Table 2.27

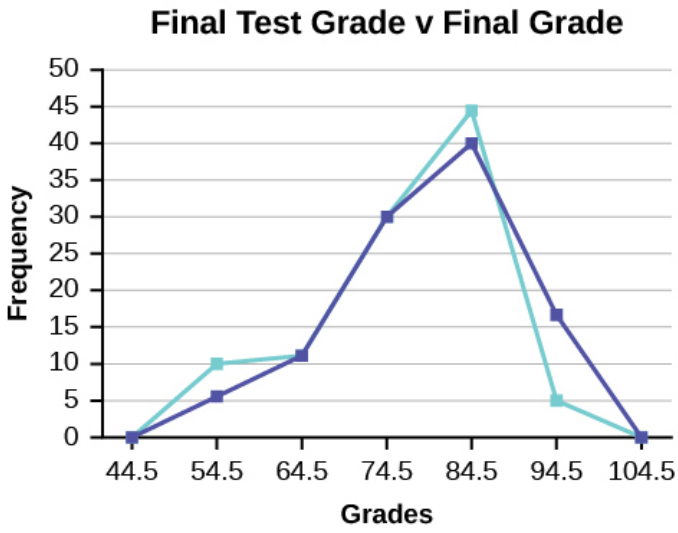


Figure 2.16

How NOT to Lie with Statistics

It is important to remember that the very reason we develop a variety of methods to present data is to develop insights into the subject of what the observations represent. We want to get a "sense" of the data. Are the observations all very much alike or are they spread across a wide range of values, are they bunched at one end of the spectrum or are they distributed evenly and so on. We are trying to get a visual picture of the numerical data. Shortly we will develop formal mathematical measures of the data, but our visual graphical presentation can say much. It can, unfortunately, also say much that is distracting, confusing and simply wrong in terms of the impression the visual leaves. Many years ago Darrell Huff wrote the book *How to Lie with Statistics*. It has been through 25 plus printings and sold more than one and one-half million copies. His perspective was a harsh one and used many actual examples that were designed to mislead. He wanted to make people aware of such deception, but perhaps more importantly to educate so that others do not make the same errors inadvertently.

Again, the goal is to enlighten with visuals that tell the story of the data. Pie charts have a number of common problems when used to convey the message of the data. Too many pieces of the pie overwhelm the reader. More than perhaps five or six categories ought to give an idea of the relative importance of each piece. This is after all the goal of a pie chart, what subset matters most relative to the others. If there are more components than this then perhaps an alternative approach would be better or perhaps some can be consolidated into an "other" category. Pie charts cannot show changes over time, although we see this attempted all too often. In federal, state, and city finance documents pie charts are often presented to show the components of revenue available to the governing body for appropriation: income tax, sales tax motor vehicle taxes and so on. In and of itself this is interesting information and can be nicely done with a pie chart. The error occurs when two years are set side-by-side. Because the total revenues change year to year, but the size of the pie is fixed, no real information is provided and the relative size of each piece of the pie cannot be meaningfully compared.

Histograms can be very helpful in understanding the data. Properly presented, they can be a quick visual way to present probabilities of different categories by the simple visual of comparing relative areas in each category. Here the error, purposeful or not, is to vary the width of the categories. This of course makes comparison to the other categories impossible. It does embellish the importance of the category with the expanded width because it has a greater area, inappropriately, and thus visually "says" that that category has a higher probability of occurrence.

Time series graphs perhaps are the most abused. A plot of some variable across time should never be presented on axes that change part way across the page either in the vertical or horizontal dimension. Perhaps the time frame is changed from years to months. Perhaps this is to save space or because monthly data was not available for early years. In either case this confounds the presentation and destroys any value of the graph. If this is not done to purposefully confuse the reader, then it certainly is either lazy or sloppy work.

Changing the units of measurement of the axis can smooth out a drop or accentuate one. If you want to show large changes, then measure the variable in small units, penny rather than thousands of dollars. And of course to continue the fraud, be sure that the axis does not begin at zero, zero. If it begins at zero, zero, then it becomes apparent that the axis has been manipulated.

Perhaps you have a client that is concerned with the volatility of the portfolio you manage. An easy way to present the data is to use long time periods on the time series graph. Use months or better, quarters rather than daily or weekly data. If that doesn't get the volatility down then spread the time axis relative to the rate of return or portfolio valuation axis. If you want to show "quick" dramatic growth, then shrink the time axis. Any positive growth will show visually "high" growth rates. Do note that if the growth is negative then this trick will show the portfolio is collapsing at a dramatic rate.

Again, the goal of descriptive statistics is to convey meaningful visuals that tell the story of the data. Purposeful manipulation is fraud and unethical at the worst, but even at its best, making these type of errors will lead to confusion on the part of the analysis.

2.2 | Box Plots -- MtRoyal - Version2016RevA

Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

NOTE

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

Consider, again, this dataset.

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

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

NOTE

See the calculator instructions on the **TI web site (http://education.ti.com/educationportal/sites/US/sectionHome/support.html)** or in the appendix.

Figure 2.17

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

NOTE

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

Example 2.17

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

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

Construct a box plot with the following properties; the calculator intructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70

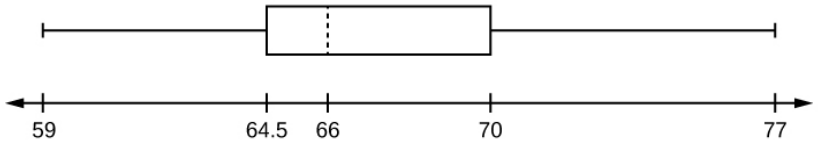


Figure 2.18

a. Each quarter has approximately 25% of the data.

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

c. Range = maximum value – the minimum value = 77 – 59 = 18

d. Interquartile Range: $IQR = Q3 – Q1 = 70 – 64.5 = 5.5$.

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

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

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

To find the minimum, maximum, and quartiles:

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

Put the data values into the list L1.

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

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4:Plotsoff. Press ENTER.

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

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

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

Try It

2.17 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

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

Figure 2.19

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

Example 2.18

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

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

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

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

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.

e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Solution 2.18

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

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

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

d.

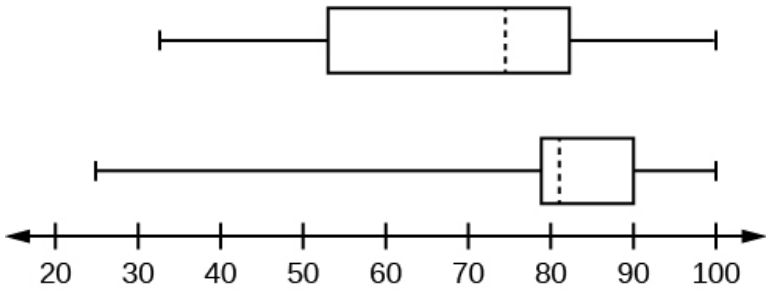


Figure 2.20

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

Try It

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

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74
The following data set shows the heights in inches for the girls in a class of 40 students.
61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69
Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Example 2.19

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

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

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

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

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

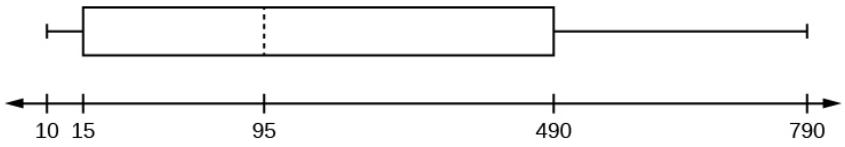


Figure 2.21

Try It

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

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

2.3 | Measures of the Location of the Data -- MtRoyal - Version2016RevA

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25^{th} percentile, and the third quartile, Q_3, is the same as the 75^{th} percentile. The median, M, is called both the second quartile and the 50^{th} percentile.

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

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

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

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.
1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1
Ordered from smallest to largest:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

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

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

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

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

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

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

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

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

Possible Quartile Positions

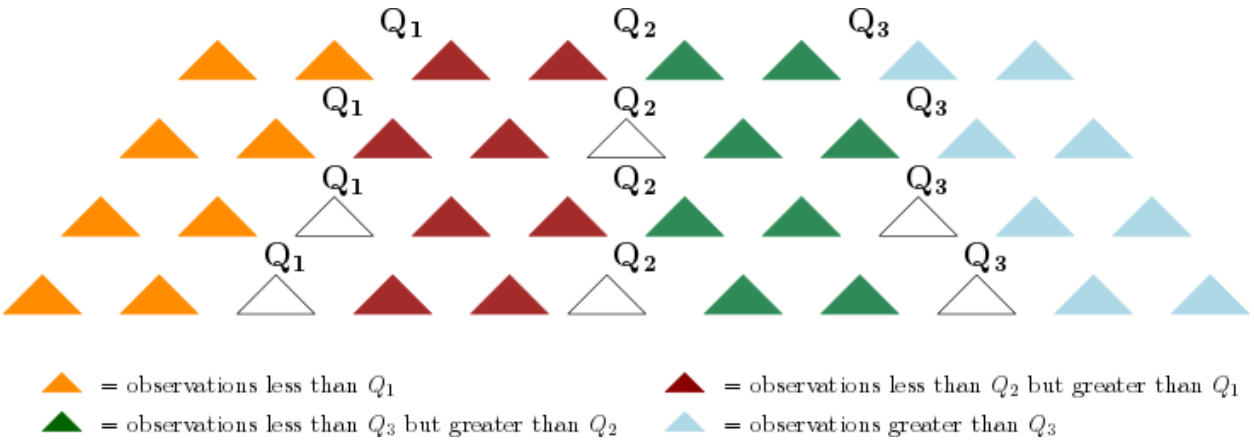


Figure 2.22

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

$IQR = Q_3 - Q_1$

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

NOTE

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

Example 2.20

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

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

Solution 2.20

Order the data from smallest to largest.
114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$M = 488{,}800$

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

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

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

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

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

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

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

Example 2.21

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

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Solution 2.21

The five number summary for the day and night classes is

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

Table 2.28

a. The IQR for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$
The IQR for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

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

b. Day class outliers are found using the IQR times 1.5 rule. So,
$Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$
$Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$
Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

$Q_1 - IQR\ (1.5) = 78 - 11(1.5) = 61.5$

$Q_3 + \text{IQR}(1.5) = 89 + 11(1.5) = 105.5$

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

Example 2.22

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

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

Table 2.29

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

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

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

Try It

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

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

Table 2.30

Example 2.23

Using **Table 2.29**:

a. Find the 80th percentile.

b. Find the 90th percentile.

c. Find the first quartile. What is another name for the first quartile?

Solution 2.23

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$

b. The 90th percentile will be the 45th data value (location is 0.90(50) = 45) and the 45th data value is nine.

c. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = 0.25(50) = 12.5 ≈ 13 the 13th data value. Thus, the 25th percentile is six.

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.24

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70^{th} percentile.

b. Find the 83^{rd} percentile.

Solution 2.24

a. $k = 70$
i = the index
$n = 29$
$i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21^{st} position in the ordered data set is 64. The 70^{th} percentile is 64 years.

b. $k = 83^{rd}$ percentile
i = the index
$n = 29$
$i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

Try It

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

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

Calculate the 20^{th} percentile and the 55^{th} percentile.

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

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

Example 2.25

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.

b. Find the percentile for 25.

Solution 2.25

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.
$x = 18$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{18 + 0.5(1)}{29}(100) = 63.80$. 58 is the 64^{th} percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.
$x = 3$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{3 + 0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12^{th} percentile.

Interpreting Percentiles, Quartiles, and Median

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

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

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

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

GUIDELINE

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

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

Example 2.26

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 2.26

- 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.)

Example 2.27

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

Solution 2.27

- 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

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.

Example 2.28

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

Solution 2.28

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

Example 2.29

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

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

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

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

Determine the following five values.

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

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

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

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

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

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

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

2.4 | Measures of the Center of the Data -- MtRoyal - Version2016RevA

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean**, **median** and **mode**. 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

Though the words “mean” and “average” are sometimes used interchangeably, they do not necessarily mean the same thing. In general, “average” is any centre of location and “mean” is a specific type of centre. Many people use average and mean as the same, but not always. For example, when people talk about average housing price, they are usually referring to the median house price.

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

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

Formula for Population Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Formula for Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

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

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

In the second example, the frequencies are 3(1) + 2(2) + 1(3) + 5(4).

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 $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49^{th} value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50^{th} and 51^{st} values. The location of the median and the value of the median are **not** the same. The upper case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.30

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 2.30

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 20^{th} and 21^{st} values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

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

Example 2.31

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 2.31

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

M = 30,000

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

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

Another measure of the center is the mode. The **mode** is the data value that occurs most frequently and at least twice.

A data set can have either

- no mode.
- one mode (unimodal)
- two modes (bimodal)
- or many modes (multimodal).

Example 2.32

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 2.32

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

Example 2.33

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

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

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, blue, the mode is red.

Comparing Measures of Centre

Measure	How Common
Mean	most familiar
Median	commonly used
Mode	sometimes used

Table 2.31

Measure	Every Score Used
Mean	yes
Median	no
Mode	no

Table 2.32

Measure	Affected by Outliers
Mean	yes
Median	no
Mode	no

Table 2.33

2.5 | Distribution -- MtRoyal - Version2016RevA

The **distribution** of a data set indicates the shape, centre and variation of the data set. The distribution gives a fairly complete idea of how the data behaves.

The **shape** of the data set is determined by looking at a visual representation of the data. Common ways of describing the shape include whether it is symmetrical or not, how many distinct peaks it has (the number of modes), and whether the data has a tail only on one side (skew).

Not all data sets have a clearly defined shape. It may only be non-symmetric.

Key idea: The distribution of sample data usually mimics the distribution of the population. But the smaller the sample size the greater the potential for there to be difference between the two distributions. This means that, for a large enough sample size, the distribution of the sample can give a good idea of distribution of the population.

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

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

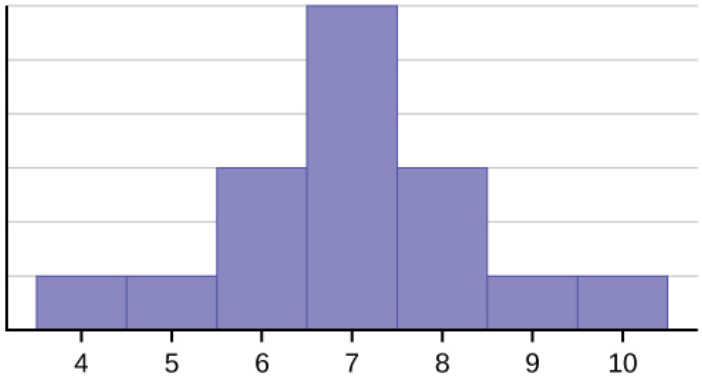


Figure 2.23

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

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

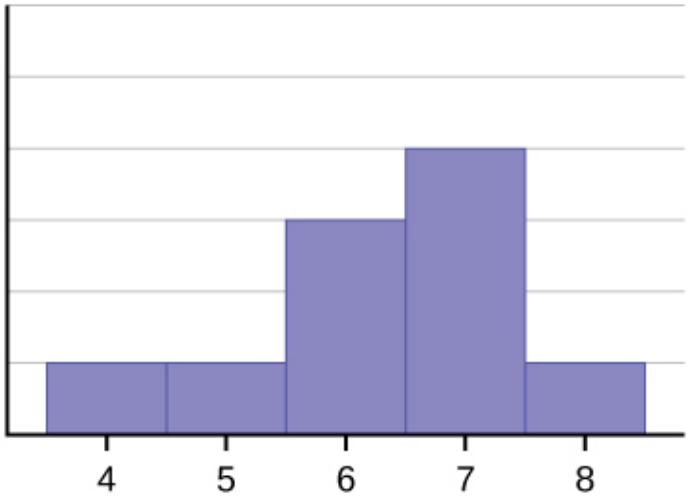


Figure 2.24

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

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

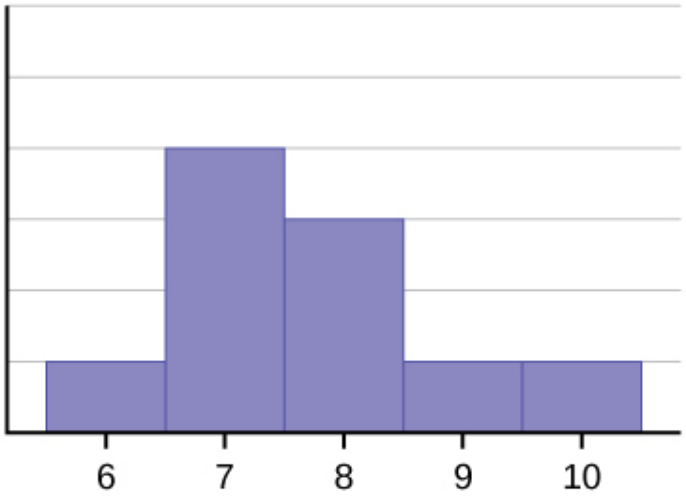


Figure 2.25

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

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

2.6 | Measures of Variaton -- MtRoyal - Version2016RevA

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. There are five measures of centre: range, standard deviation, variance, interquartile range and coefficient of variation. The range is the easiest to calculate. It is found by subtracting the maximum value in the data set from the minimum value in the data set. Though the

range is easy to calculate, it is very much affected by outliers. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

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

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

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

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

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

Calculating the Standard Deviation

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

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

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

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

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\left(\sum_{i=1}^{n} x^2\right) - n\bar{x}^2}{n - 1}}$
- For the sample standard deviation, the denominator is $\boldsymbol{n - 1}$, that is the sample size $-$ 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$
- For the population standard deviation, the denominator is N, the number of items in the population.

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

Example 2.34

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

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

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

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

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

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

Table 2.34

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

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

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

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, $s = 0.72$.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For **Example 2.34**, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

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

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

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

Example 2.35

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

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

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.

b. Calculate the following to one decimal place:

   i. The sample mean

   ii. The sample standard deviation

   iii. The median

   iv. The first quartile

   v. The third quartile

   vi. *IQR*

Solution 2.35

a. See **Table 2.35**

b. i. The sample mean = 73.5

   ii. The sample standard deviation = 17.9

   iii. The median = 73

   iv. The first quartile = 61

   v. The third quartile = 90

   vi. *IQR* = 90 – 61 = 29

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484

Table 2.35

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1? ANSWER: Rounding)

Table 2.35

Coefficient of variation

- The **coefficient of variation** is the standard deviation expressed as a percentage of the mean: coefficient of variation = $\frac{s}{\bar{x}}(100\%)$
- The coefficient of variation is not affected by multiplicative changes of scale.
- **The coefficient of variation is used to compare variation between data sets.** This is very important to remember. For multiple data sets, if the means are the same, you can compare the standard deviations. BUT if the means are different, you MUST use the coefficient of variation of compare the variation in the data sets.
- If the standard deviation is larger than the mean, the coefficient of variation is bigger than 100%.

$$\text{Coefficient of Variation} = \frac{s}{\bar{x}}(100\%)$$

Measure	When to use
Range	The range is rarely the best measure of variation to use. But it is a good quick calculation of variation.
Standard deviation	This is the most common measure of variation. It is best used when finding the variation for one data set.
Variance	As it the square of the standard deviation, it is NEVER the best measure of variation to use. It is helpful in later topics in statistics though.
Interquartile range	This is a not very well known measure of variation, but it is helpful in describing the range for middle 50% of the data values.
Coefficient of variation	This is not well known, but is the best measure to use when comparing the variations of two or more data sets that have different measures of centre.

Table 2.36

Comparing Values from Different Data Sets

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

- For each data value, calculate how many standard deviations away from its mean the value is.

- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\ of STDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

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

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

Table 2.37

Example 2.36

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

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

Table 2.38

Solution 2.36

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

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

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

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

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

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

Try It

2.36 Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Table 2.39

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 the number of times a value of the data occurs

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

Histogram a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

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 (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

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

Midpoint the mean of an interval in a frequency table

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

Outlier an observation that does not fit 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 50^{th} percentile. The first and third quartiles are the 25^{th} and the 75^{th} percentiles, respectively.

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

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

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

CHAPTER REVIEW

2.1 Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA

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

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on y-axis with the frequency being graphed on the x-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

2.2 Box Plots -- MtRoyal - Version2016RevA

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

2.3 Measures of the Location of the Data -- MtRoyal - Version2016RevA

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

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

2.4 Measures of the Center of the Data -- MtRoyal - Version2016RevA

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

2.5 Distribution -- MtRoyal - Version2016RevA

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like **Figure 2.24**. A **left (or negative) skewed** distribution has a shape like **Figure 2.25**. A **symmetrical** distrubtion looks like **Figure 2.23**.

2.6 Measures of Variaton -- MtRoyal - Version2016RevA

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

FORMULA REVIEW

2.3 Measures of the Location of the Data -- MtRoyal - Version2016RevA

$$i = \left(\frac{k}{100}\right)(n+1)$$

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

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x + 0.5y}{n}\right)(100)$

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

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

n = total number of data

2.4 Measures of the Center of the Data -- MtRoyal - Version2016RevA

$\mu = \frac{\sum fm}{\sum f}$ Where f = interval frequencies and m = interval midpoints.

The mean for a sample (denoted by $\bar{x}$) is

$$\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

The mean for a population (denoted by μ) is

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$$

2.6 Measures of Variaton -- MtRoyal - Version2016RevA

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

$s_x =$ sample standard deviation

$\bar{x} =$ sample mean

Formulas for Sample Standard Deviation $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\left(\sum_{i=1}^{n} x^2\right) - n\bar{x}^2}{n-1}}$ For the sample standard deviation, the denominator is ***n*** **- 1**, that is the sample size - 1.

Formulas for Population Standard Deviation $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$ For the population standard deviation, the denominator is N, the number of items in the population.

PRACTICE

2.1 Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA

For the next three exercises, use the data to construct a line graph.

1. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in **Table 2.40**.

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

Table 2.40

2. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in **Table 2.41**.

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

Table 2.41

3. Several children were asked how many TV shows they watch each day. The results of the survey are shown in **Table 2.42**.

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

Table 2.42

4. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. **Table 2.43** 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	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Table 2.43

5. Using the data from Mrs. Ramirez's math class supplied in **Exercise 2.4**, construct a bar graph showing the percentages.

6. David County has six high schools. Each school sent students to participate in a county-wide science competition. **Table 2.44** 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 that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
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%

Table 2.44

7. Use the data from the David County science competition supplied in **Exercise 2.6**. Construct a bar graph that shows the county-wide population percentage of students at each school.

8. 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

Table 2.45

9. What does the frequency column in **Table 2.45** sum to? Why?

10. What does the relative frequency column in **Table 2.45** sum to? Why?

11. What is the difference between relative frequency and frequency for each data value in **Table 2.45**?

12. What is the difference between cumulative relative frequency and relative frequency for each data value?

13. To construct the histogram for the data in **Table 2.45**, determine appropriate minimum and maximum *x* and *y* values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Figure 2.26

14. Construct a frequency polygon for the following:

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

Table 2.46

b.

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

Table 2.47

c.

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

Table 2.48

15. Construct a frequency polygon 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

Table 2.49

16. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed 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 – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Table 2.50

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

Table 2.51

17. 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/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	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

Table 2.52

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Table 2.53

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

Table 2.54

18. 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 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Table 2.55

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

Table 2.56

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.

2.2 Box Plots -- MtRoyal - Version2016RevA

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.

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

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

2.3 Measures of the Location of the Data -- MtRoyal - Version2016RevA

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

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

a. Find the 40^{th} percentile.
b. Find the 78^{th} percentile.

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

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

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

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

24.

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

25.

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

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

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

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

29. 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 90^{th} percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90^{th} percentile in the context of this problem.

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

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

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

31. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34^{th} percentile. The 34^{th} 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?

Use **Exercise 2.23** to calculate the following values:

32. First quartile = _____

33. Second quartile = median = 50^{th} percentile = _____

34. Third quartile = _____

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

36. 10^{th} percentile = _____

37. 70^{th} percentile = _____

2.4 Measures of the Center of the Data -- MtRoyal - Version2016RevA

38. Find the mean for the following frequency tables.

a.

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

Table 2.57

b.

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

Table 2.58

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

Table 2.59

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

39. Calculate the mean.

40. Identify the median.

41. Identify the mode.

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

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

43. median = _____

44. mode = _____

2.5 Distribution -- MtRoyal - Version2016RevA

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

45. 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

46. 16; 17; 19; 22; 22; 22; 22; 22; 23

47. 87; 87; 87; 87; 87; 88; 89; 89; 90; 91

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

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

50. What word describes a distribution that has two modes?

51. Describe the shape of this distribution.

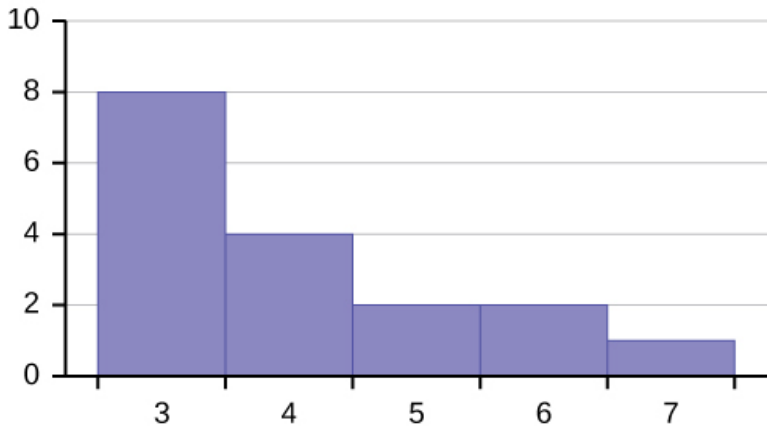


Figure 2.27

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

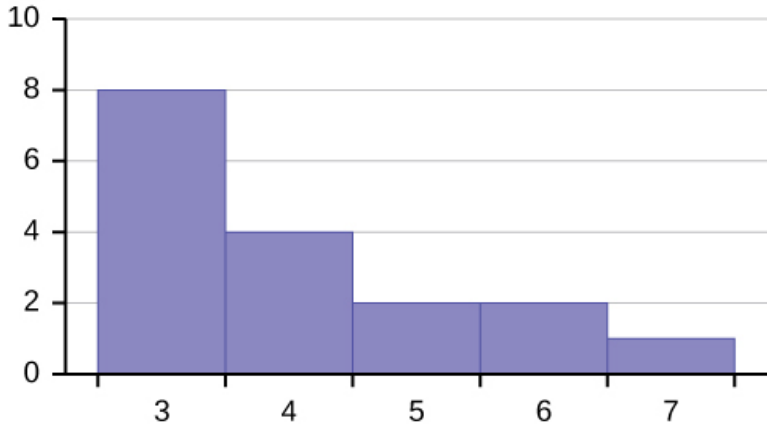


Figure 2.28

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

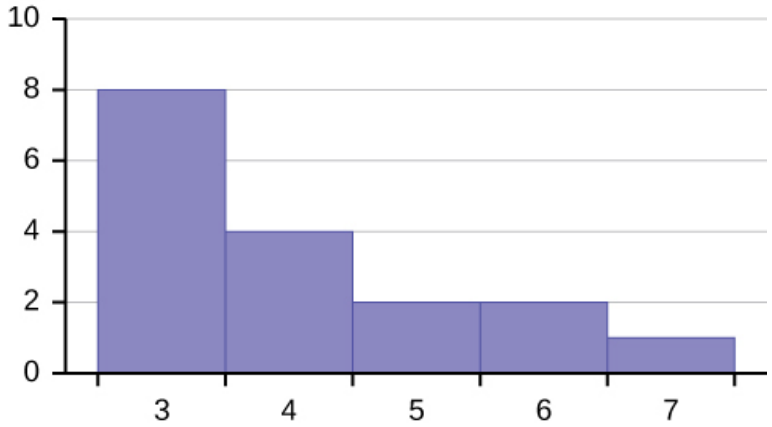


Figure 2.29

54. Describe the shape of this distribution.

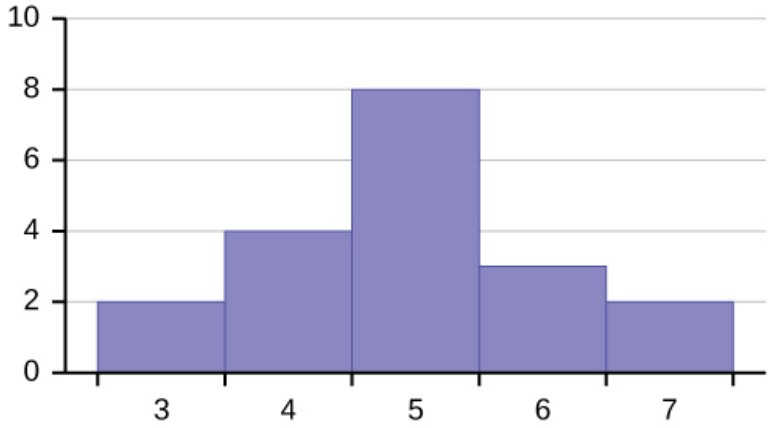


Figure 2.30

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

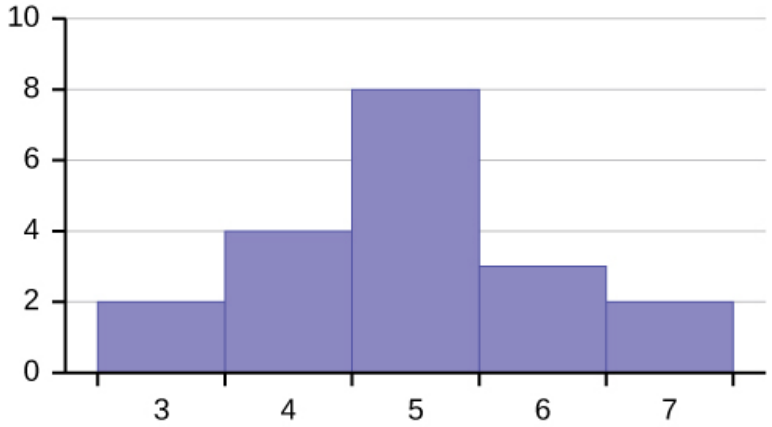


Figure 2.31

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

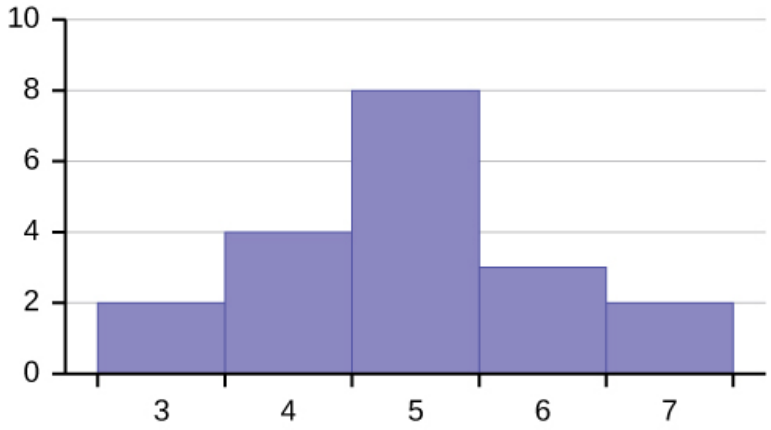


Figure 2.32

57. Describe the shape of this distribution.

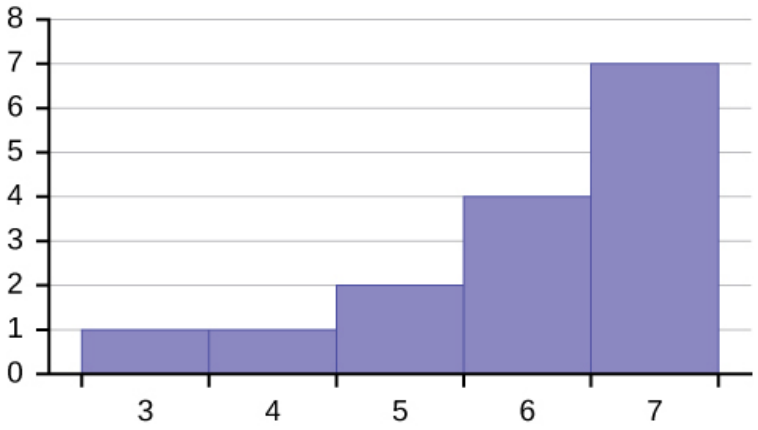


Figure 2.33

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

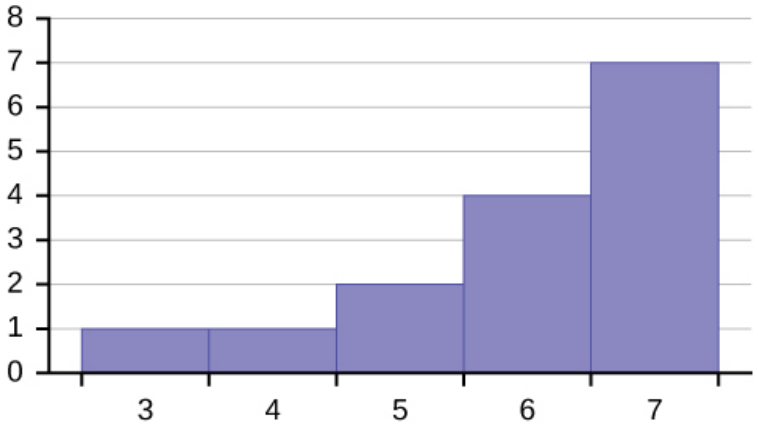


Figure 2.34

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

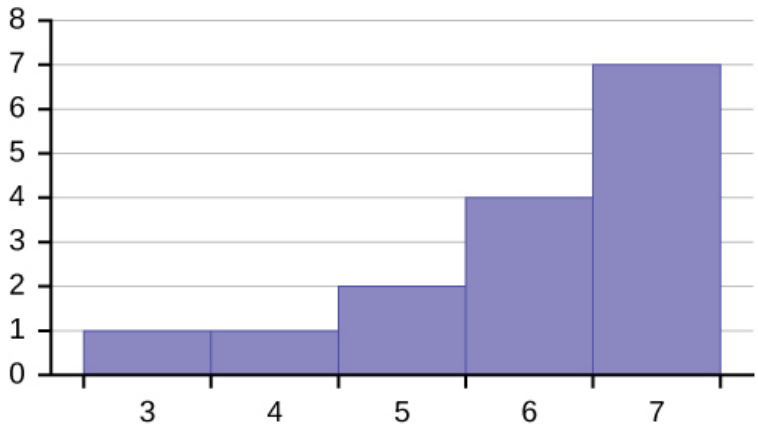


Figure 2.35

60. The mean and median for the data are the same.

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

Is the data perfectly symmetrical? Why or why not?

61. 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

62. 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

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

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

2.6 Measures of Variaton -- MtRoyal - Version2016RevA

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

65. Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

66. Find the value that is one standard deviation below the mean.

67. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

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

Table 2.60

68. Use **Table 2.60** to find the value that is three standard deviations:
a. above the mean
b. below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

69. Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

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

Table 2.61

b.

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

Table 2.62

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

Table 2.63

HOMEWORK

2.1 Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA

70. **Table 2.64** contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.64

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

71. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.65 Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.66 Publisher B

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Table 2.67 Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

72. Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Table 2.68 Singles

Amount($)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Table 2.69 Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

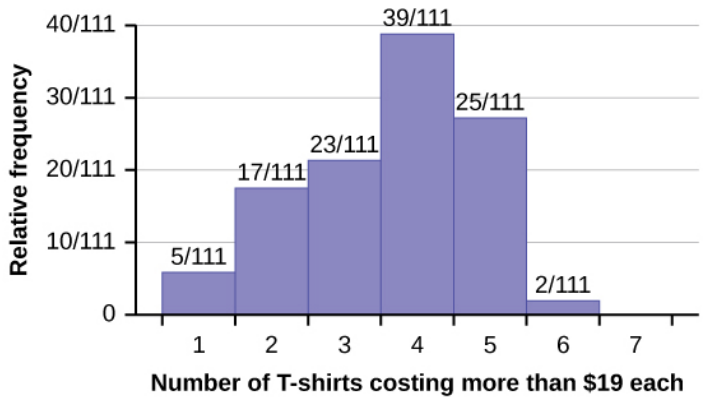
73. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

Table 2.70

a. Construct a histogram of the data.
b. Complete the columns of the chart.

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.



74. 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

75. 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

76. Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.71

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

2.2 Box Plots -- MtRoyal - Version2016RevA

77. 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.

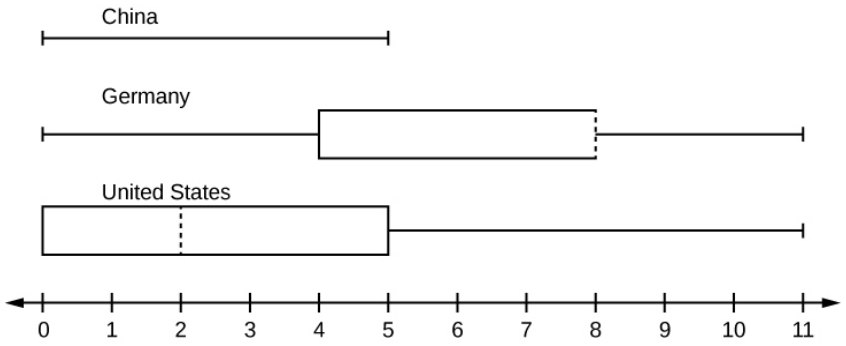


Figure 2.36

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over 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?

78. Given the following box plot, answer the questions.

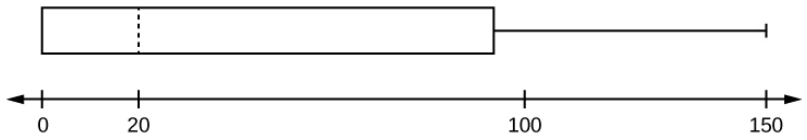


Figure 2.37

a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

79. Given the following box plots, answer the questions.

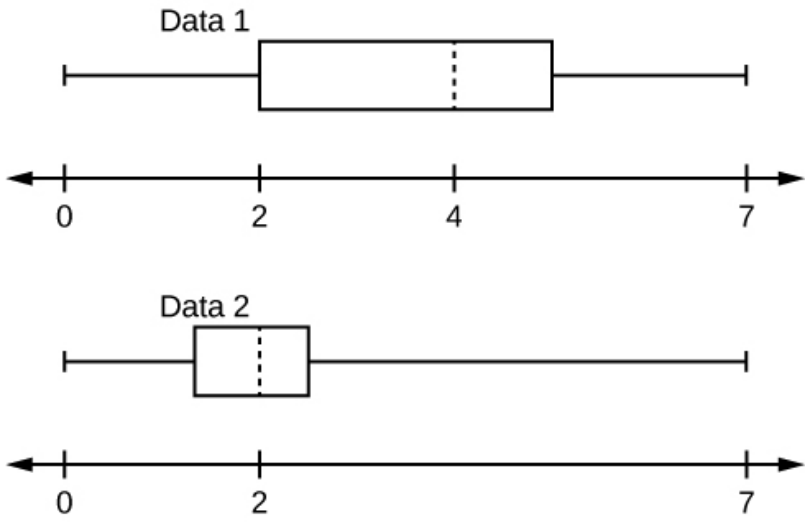


Figure 2.38

a. In complete sentences, explain why each statement is false.
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.

b. For which group, Data 1 or Data 2, is the value of "7" more likely to be an outlier? Explain why in complete sentences.

80. 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 it, people were asked the age they were when they purchased their car. The following box plots display the results.

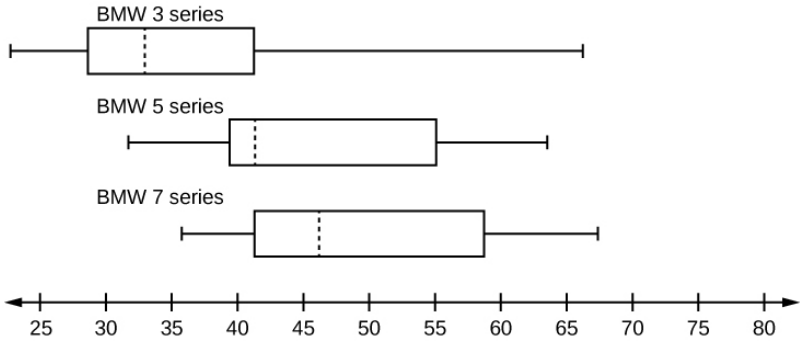


Figure 2.39

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. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 series. Estimate the interquartile range (IQR).
g. Look at 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. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
    i. 31–35
    ii. 38–41
    iii. 41–64

81. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Table 2.72

Construct a box plot of the data.

2.3 Measures of the Location of the Data -- MtRoyal - Version2016RevA

82. The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

83. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in **Table 2.73**. Also, include left endpoint, but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

Table 2.73

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
    i. Should all bars have the same width, based on the data? Why or why not?
    ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40^{th} and 80^{th} percentiles
e. Construct a bar graph of the data

2.4 Measures of the Center of the Data -- MtRoyal - Version2016RevA

84. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	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

Table 2.74

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. How does the United States compare to other countries?

85. **Table 2.75** gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

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

Table 2.75

2.5 Distribution -- MtRoyal - Version2016RevA

86. 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?

2.6 Measures of Variaton -- MtRoyal - Version2016RevA

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES

- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

87. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

88. 75% of all years have an FTES:
a. at or below: _____
b. at or above: _____

89. The population standard deviation = _____

90. What percent of the FTES were from 528.5 to 1447.5? How do you know?

91. What is the *IQR*? What does the *IQR* represent?

92. How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Table 2.76

93. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

94. Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

95. 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 Standard Deviation**
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Table 2.77

96. 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.

97. 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 why.

BRINGING IT TOGETHER: HOMEWORK

98. 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

Table 2.78

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

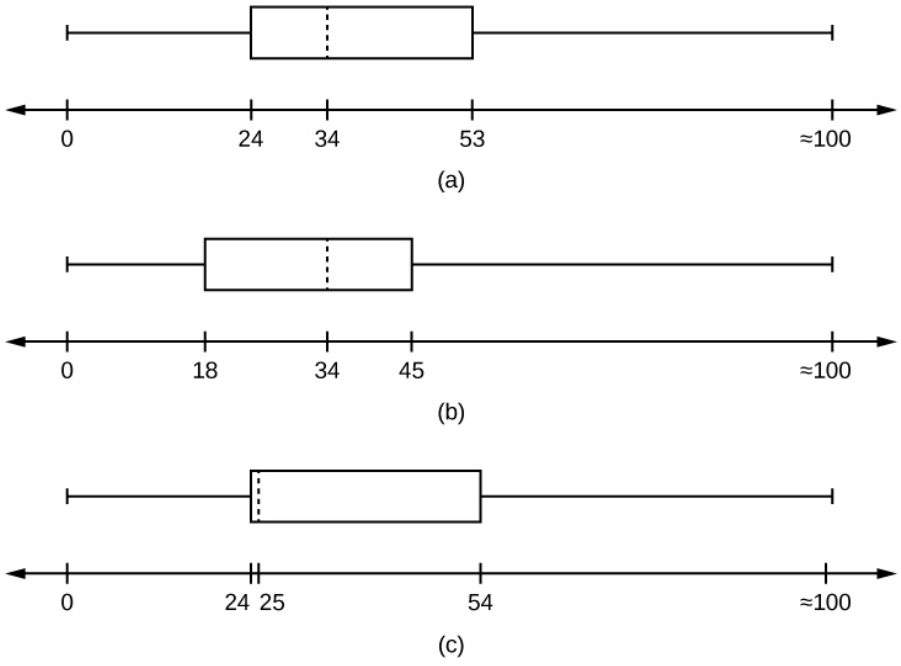


Figure 2.40

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

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

Table 2.79

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

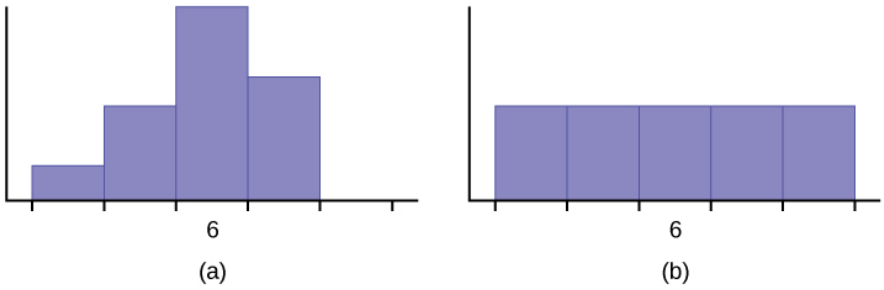


Figure 2.41

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

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

Table 2.80

100. What is the *IQR*?
a. 8
b. 11
c. 15
d. 35

101. What is the mode?

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

102. Is this a sample or the entire population?

a. sample
b. entire population
c. neither

103. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Table 2.81

a. Find the sample mean $\bar{x}$.
b. Find the approximate sample standard deviation, s.

104. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

Table 2.82

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. What percent of the students owned at least five pairs?
i. Find the 40^{th} percentile.
j. Find the 90^{th} percentile.
k. Construct a line graph of the data
l. Construct a stemplot of the data

105. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. The middle 50% of the weights are from _____ to _____.
f. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
g. 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?
h. Assume the population was the San Francisco 49ers. Find:
    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
    iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
i. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

106. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

107. Refer to **Figure 2.42** determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

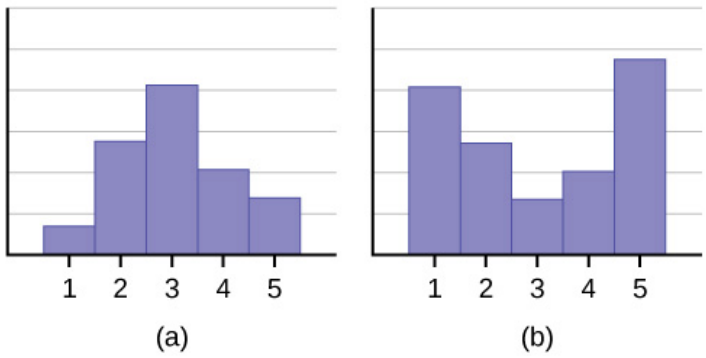


Figure 2.42

a. The medians for both graphs are the same.
b. We cannot determine if any of the means for both 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 both graphs is different.

108. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. The middle 50% of the conferences last from _____ days to _____ days.
f. Calculate the sample mean of days of engineering conferences.
g. Calculate the sample standard deviation of days of engineering conferences.
h. Find the mode.
i. 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.
j. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

109. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

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

Table 2.83

110. The 80^{th} percentile is _____

a. 5
b. 80
c. 3
d. 4

111. The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

112. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the **Table 2.84**.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.84

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

REFERENCES

2.1 Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA

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).

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).

2.2 Box Plots -- MtRoyal - Version2016RevA

Data from *West Magazine*.

2.3 Measures of the Location of the Data -- MtRoyal - Version2016RevA

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.

2.4 Measures of the Center of the Data -- MtRoyal - Version2016RevA

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).

2.6 Measures of Variaton -- MtRoyal - Version2016RevA

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).

SOLUTIONS

1

Figure 2.43

3

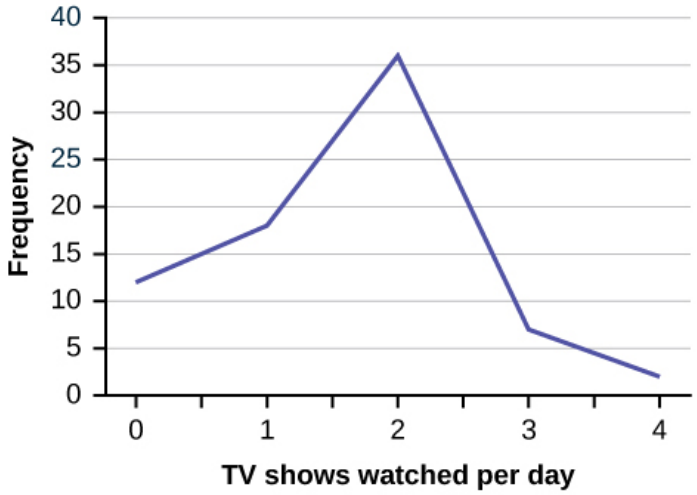


Figure 2.44

5

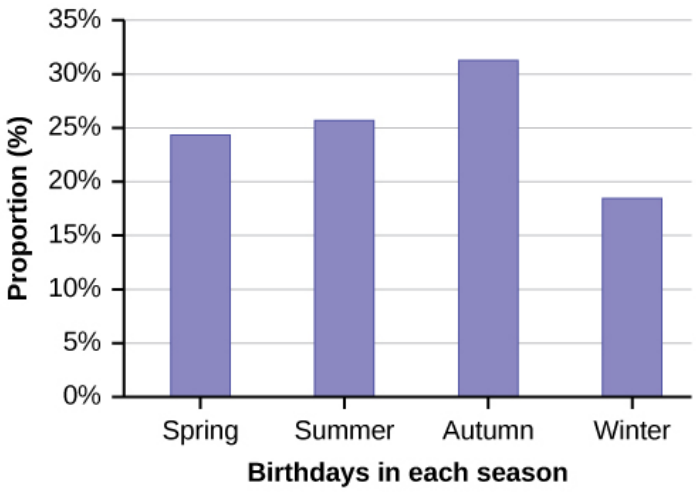


Figure 2.45

7

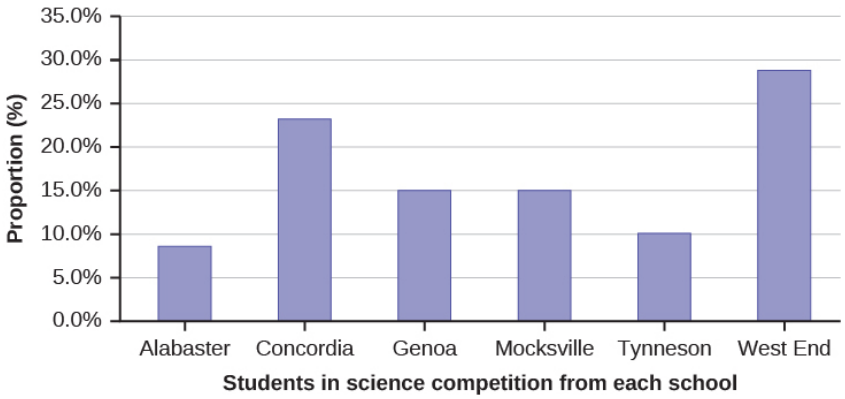


Figure 2.46

9 65

11 The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

13 Answers will vary. One possible histogram is shown:

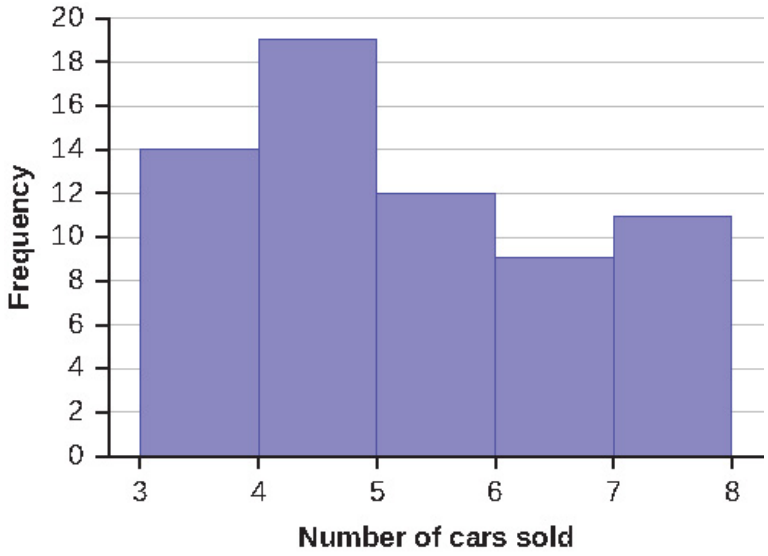


Figure 2.47

15 Find the midpoint for each class. These will be graped on the *x*-axis. The frequency values will be graphed on the *y*-axis values.

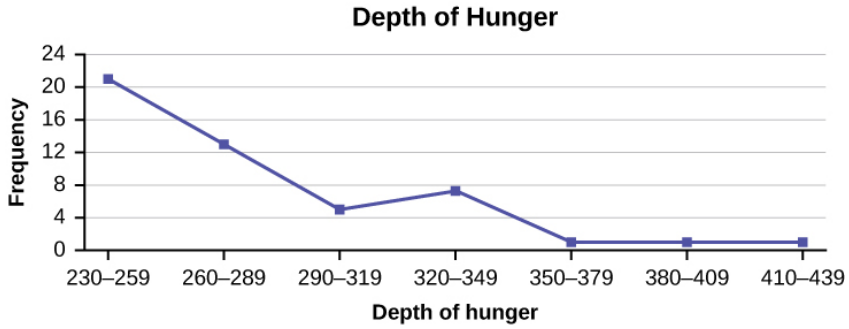


Figure 2.48

17

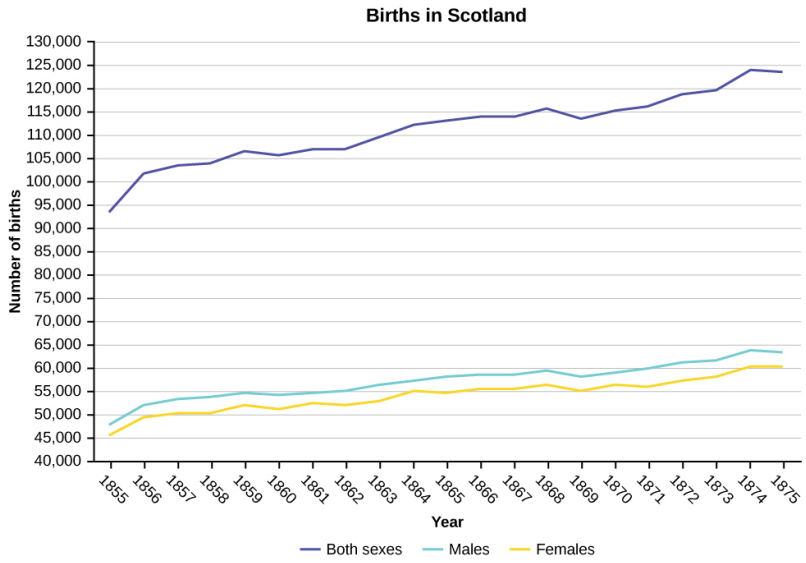


Figure 2.49

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

21

a. The 40^{th} percentile is 37 years.

b. The 78^{th} percentile is 70 years.

23 Jesse graduated 37^{th} out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37. $x = 143$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{143 + 0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80^{th} percentile.

25

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

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

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

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

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

33 4

35 6 – 4 = 2

37 6

39 Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738; $\frac{738}{27}$ = 27.33

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

43 4

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

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

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

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

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

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

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

59 The mean and the median are both six.

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

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

65 s = 34.5

67 For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$ For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$ Fredo's z-score of –0.67 is higher than Karl's z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

69

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

70

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.

Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)

Press MATH

Arrow over to PRB

Press 5:randInt(

Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

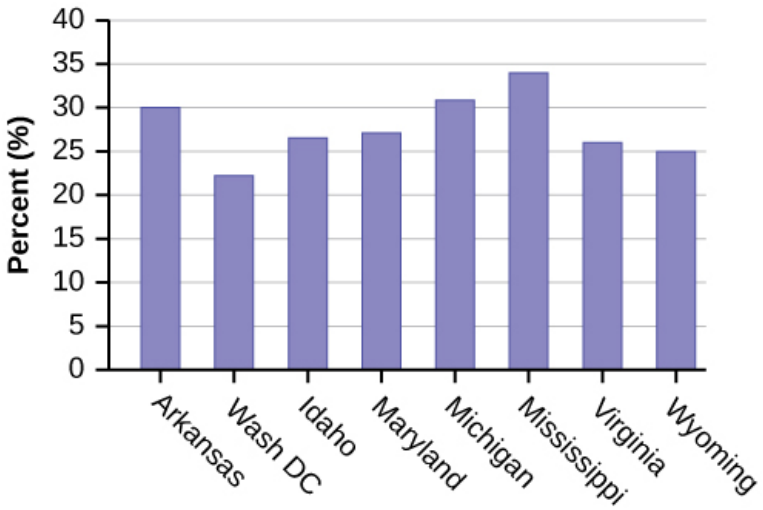


Figure 2.50

b.

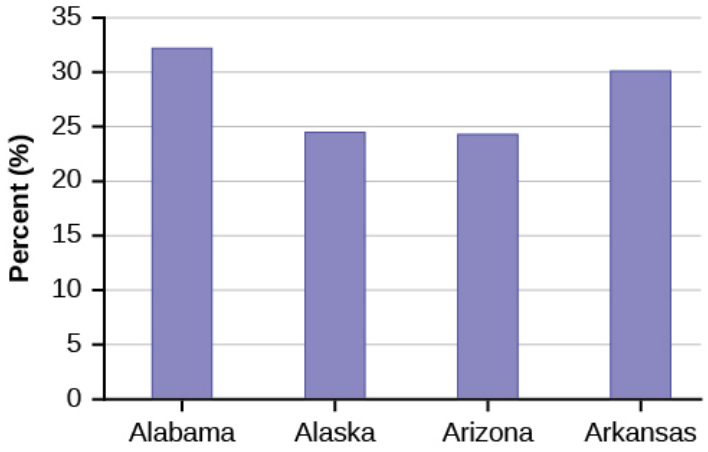
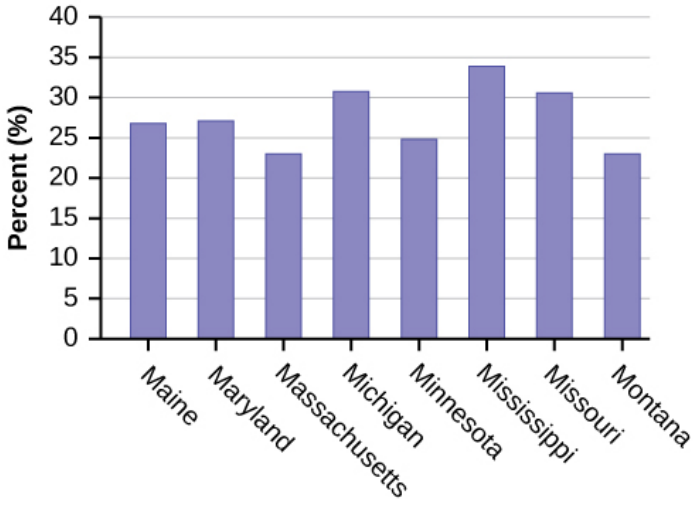


Figure 2.51



c.

Figure 2.52

72

Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Table 2.85 Singles

Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Table 2.86 Couples

a. See **Table 2.85** and **Table 2.86**.

b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

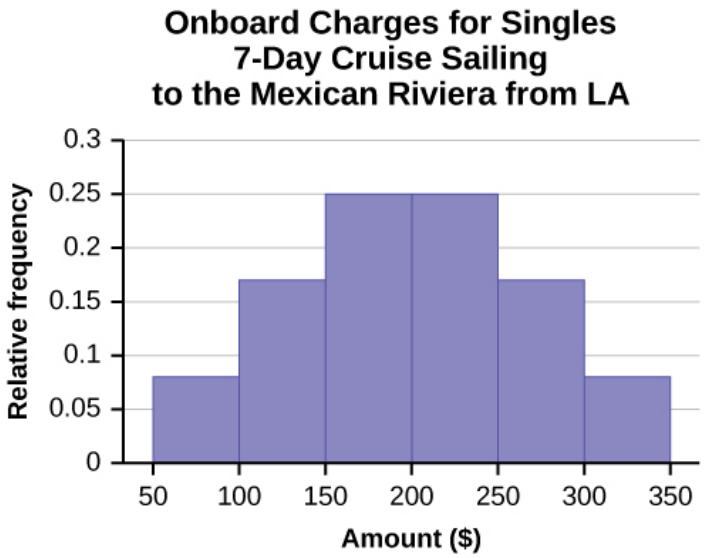


Figure 2.53

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

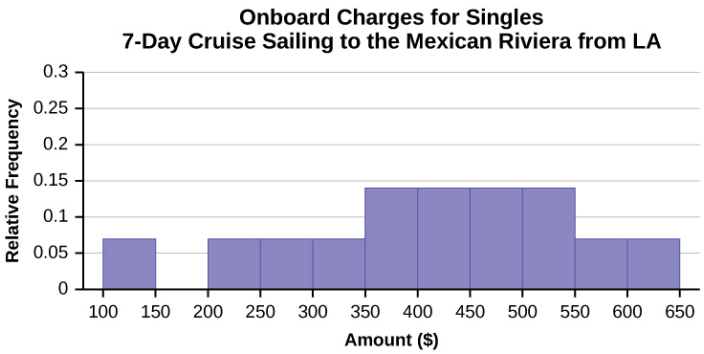


Figure 2.54

d. Compare the two graphs:

i. Answers may vary. Possible answers include:

- Both graphs have a single peak.
- Both graphs use class intervals with width equal to $50.

ii. Answers may vary. Possible answers include:

- The couples graph has a class interval with no values.
- It takes almost twice as many class intervals to display the data for couples.

iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the Singles with the new graph for the Couples:

i.
- Both graphs have a single peak.
- Both graphs display 6 class intervals.
- Both graphs show the same general pattern.

ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

74 c

76 Answers will vary.

78

a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.

b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

80

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. $IQR \sim 17$ 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.

83

a. $1 - (0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01) = 0.06$

b. $0.19+0.26+0.18 = 0.63$

c. Check student's solution.

d. 40^{th} percentile will fall between 30,000 and 40,000

80^{th} percentile will fall between 50,000 and 75,000

e. Check student's solution.

85 The mean percentage, $\bar{x} = \frac{1328.65}{50} = 26.75$

87 The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

89 474 FTES

91 919

93

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

94 Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

96 For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

98

a. For graph, check student's solution.

b. 49.7% of the community is under the age of 35.

c. Based on the information in the table, graph (a) most closely represents the data.

100 a

102 b

103

a. 1.48

b. 1.12

105

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302

b. 241

c. 205.5

d. 272.5

e. 205.5, 272.5

f. sample

g. population

h. i. 236.34

ii. 37.50

iii. 161.34

iv. 0.84 std. dev. below the mean

i. Young

107

a. True

b. True

c. True

d. False

109

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

Table 2.87

b. Check student’s solution.

c. mode

d. 8628.74

e. 6943.88

f. –0.09

111 a

3 | PROBABILITY TOPICS

Figure 3.1 Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Introduction

CHAPTER OBJECTIVE

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 and multiplication rules.
- Construct and interpret contingency tables and tree diagrams.
- Understand the difference between likely and unlikely events.

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

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

3.1 | Terminology -- Probability Topics -- MtRoyal - Version2016RevA

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

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

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

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means that event A has a 50% chance of happening. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

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

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A) = \frac{2}{4} = 0.5$.

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

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

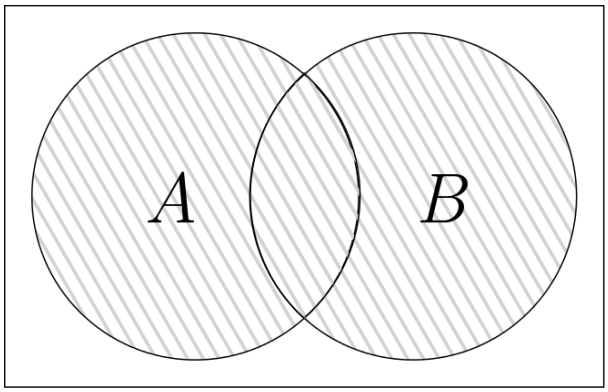
It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and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

A key concept in probability is whether an event is **likely** or **unlikely**. A likely event is an event that has a good chance of happening, while an unlikely event is rare. For example, it is likely to snow in Calgary in the winter, but it is unlikely to snow in Calgary in the summer (it can happen, but it would be a rare or strange event). In general, in statistics, unlikely

events usually have a probability of less than 1% of happening. Likely events usually have a probability of greater than 10% of happening. If the probability of the event is between 1% and 10%, it is up to the statistician or researcher to make a call to determine whether it is likely or unlikely.

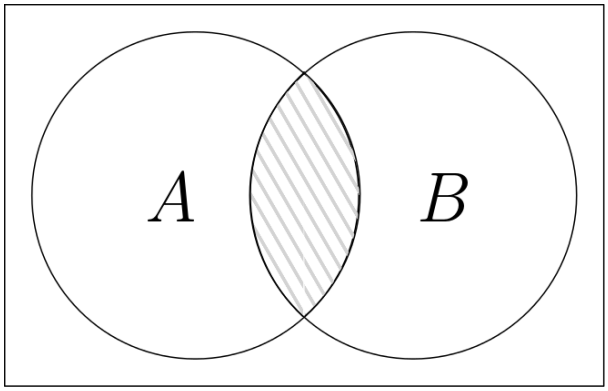
"OR" Event:

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



"AND" Event:

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



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

The **conditional probability** of *A* given *B* is written *P*(*A*|*B*). *P*(*A*|*B*) is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate *P*(*A*|*B*) is *P*(*A*|*B*) = $\frac{P(A \text{ AND } B)}{P(B)}$ where *P*(*B*) is greater than zero.

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

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

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

Odds

The odds of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:

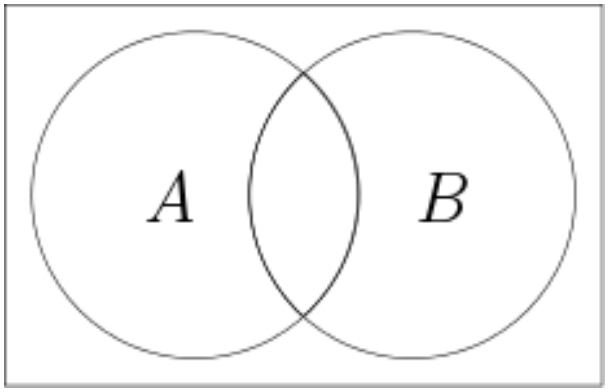
$$\frac{P(A)}{1 - P(A)}$$

where $P(A)$ is the probability of success and of course $1 - P(A)$ is the probability of failure. Odds are always quoted as "numerator to denominator," e.g. 2 to 1. Here the probability of winning is twice that of losing; thus, the probability of winning is 0.66. A probability of winning of 0.60 would generate odds in favor of winning of 3 to 2. While the calculation of odds can be useful in gambling venues in determining payoff amounts, it is not helpful for understanding probability or statistical theory.

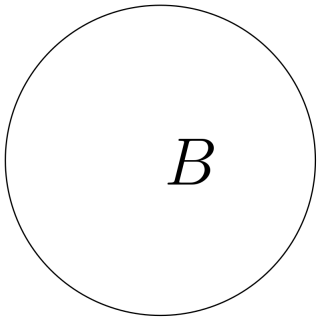
Understanding Terminology and Symbols

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

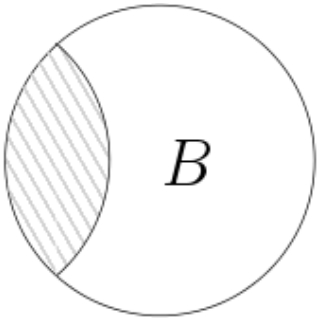
If the sample space is



then $P(A|B)$ is found by looking only at events that involved B:



and within B looking at the portion that involve A:



That portion is clearly the intersection of A and B.

Example 3.1

The sample space S is the whole numbers starting at one and less than 20.

a. S = _____
Let event A = the even numbers and event B = numbers greater than 13.

b. A = _____, B = _____

c. $P(A)$ = _____, $P(B)$ = _____

d. A AND B = _____, A OR B = _____

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

f. A' = _____, $P(A')$ = _____

g. $P(A) + P(A')$ = _____

h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

Solution 3.1

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$

b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$

c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$

d. A AND B = {14,16,18}, A OR $B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$

e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$

f. $A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(A') = \frac{10}{19}$

g. $P(A) + P(A') = 1$ ($\frac{9}{19} + \frac{10}{19} = 1$)

h. $P(A|B) = \frac{P(A \text{AND} B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{AND} B)}{P(A)} = \frac{3}{9}$, No

Try It

3.1 The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. S = _____

Let event A = the sum is even and event B = the first number is prime.

b. A = _____, B = _____

c. $P(A)$ = _____, $P(B)$ = _____

d. A AND B = _____, A OR B = _____

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

f. B' = _____, $P(B')$ = _____

g. $P(A) + P(A')$ = _____

h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

Example 3.2

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

a. Event T = the outcome is two.

b. Event A = the outcome is an even number.

c. Event B = the outcome is less than four.

d. The complement of A.

e. A GIVEN B

f. B GIVEN A

g. A AND B

h. A OR B

i. A OR B'

j. Event N = the outcome is a prime number.

k. Event I = the outcome is seven.

Solution 3.2

a. $T = \{2\}$, $P(T) = \frac{1}{6}$

b. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$

c. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$

d. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$

e. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$

f. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$

g. A AND $B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$

h. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$

i. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$

j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$

k. A six-sided die does not have seven dots. $P(7) = 0$.

Example 3.3

Table 3.1 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

Table 3.1

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$
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. $P(R|M)$
l. $P(F|L)$
m. $P(L|F)$

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

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

3.2 | Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

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

Independent Events

Two events are independent if the following are true:

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

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

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

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

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

Example 3.4

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit.

a. Sampling with replacement:
Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:
Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice. The probability of picking the three of diamonds

is called a conditional probability because it is conditioned on what was picked first. This is true also of the probability of picking the J of spades. The probability of picking the J of spades is actually conditioned on *both* the previous picks.

Try It

3.4 You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?

b. Suppose you know that the picked cards are *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

Example 3.5

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are *QS*, 1*D*, 1*C*, *QD*.

b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution 3.5

a. Without replacement; b. With replacement

Try It

3.5 You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. *QS*, 1*D*, 1*C*, *QD*

b. *KH*, 7*D*, 6*D*, *KH*

c. *QS*, 7*D*, 6*D*, *KS*

Mutually Exclusive Events

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

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

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example 3.6

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

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

Solution 3.6

Look at the sample space in **???**.

a. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(F) = \frac{3}{4}$
b. Two faces are the same if HH or TT show up. $P(G) = \frac{2}{4}$
c. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(H) = \frac{2}{4}$
d. F and G share HH so $P(F \text{ AND } G)$ is not equal to zero (0). 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 nothing in common so $P(J \text{ AND } H) = 0$. J and H are mutually exclusive.

Try It

3.6 A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Example 3.7

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

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.

- Let event C = odd faces larger than two. Then $C = \{3, 5\}$. Let event D = all even faces smaller than five. Then $D = \{2, 4\}$. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. $E = \{1, 2, 3, 4\}$.

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

Solution 3.7

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

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

Try It

3.7 Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

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

Example 3.8

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

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

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

NOTE

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

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

Solution 3.8

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

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution 3.8

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

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

Try It

3.8 In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Example 3.9

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

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

Justify your answers to the following questions numerically.

a. Are C and D independent?

b. Are C and D mutually exclusive?

c. What is $P(D|C)$?

Solution 3.9

a. Yes, because $P(C|D) = P(C)$.

b. No, because $P(C \text{ AND } D)$ is not equal to zero.

c. $P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{0.75} = 0.3$

Try It

3.9 A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

a. Find $P(B|D)$.

b. Find $P(D|B)$.

c. Are B and D independent?

d. Are B and D mutually exclusive?

Example 3.10

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

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

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

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

Try It

3.10 In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.
Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Example 3.11

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

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$

- $P(L|F) = 0.75$

NOTE

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

Solution 1

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

Solution 2

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

Interpretation of Results

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

Try It

3.11 Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.55$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Example 3.12

a. Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are ____ outcomes.

b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ___ outcomes.

c. Multiply the two numbers of outcomes. The answer is _____.

d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are $H1$ and $T6$.)

e. Event A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.
A = {_____}. Find $P(A)$.

f. Event B = heads on the coin followed by a three on the die. B = {_____}. Find $P(B)$.

g. Are A and B mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.)

h. Are A and B independent? (Hint: Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then A and B are independent. If not, then they are dependent).

Solution 3.12

a. H and T; 2

b. 1, 2, 3, 4, 5, 6; 6

c. 2(6) = 12

d. $T1$, $T2$, $T3$, $T4$, $T5$, $T6$, $H1$, $H2$, $H3$, $H4$, $H5$, $H6$

e. $A = \{H2, H4, H6\}$; $P(A) = \frac{3}{12}$

f. $B = \{H3\}$; $P(B) = \frac{1}{12}$

g. Yes, because $P(A \text{ AND } B) = 0$

h. $P(A \text{ AND } B) = 0.P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so A and B are dependent.

Try It

3.12 A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(T)$.

b. Compute $P(T|F)$.

c. Are T and F independent?.

d. Are F and S mutually exclusive?

e. Are F and S independent?

3.3 | Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

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

This rule may also be written as: $P\left(A\middle|B\right) = \frac{P(A \cap B)}{P(B)}$

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

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

One easy way to remember the multiplication rule is that the word "and" means that the event has to satisfy two conditions. For example the name drawn from the class roster is to be both a female and a sophomore. It is harder to satisfy two conditions than only one and of course when we multiply fractions the result is always smaller. This reflects the increasing difficulty satisfying two conditions.

The Addition Rule

If A and B are defined on a sample space, then: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If A and B are **mutually exclusive**, then $P(A \cap B) = 0$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes $P(A \cup B) = P(A) + P(B)$.

Example 3.13

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

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

Example 3.14

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal | that he made the first goal is 0.90.

a. What is the probability that he makes both goals?

Solution 3.14

a. The problem is asking you to find $P(A \cap B) = P(B \cap A)$. Since $P(B|A) = 0.90$: $P(B \cap A) = P(B|A)P(A) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

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

Solution 3.14

b. The problem is asking you to find $P(A \cup B)$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.65 - 0.585 = 0.715$

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

c. Are A and B independent?

Solution 3.14

c. No, they are not, because $P(B \cap A) = 0.585$.

$P(B)P(A) = (0.65)(0.65) = 0.423$

$0.423 \neq 0.585 = P(B \cap A)$

So, $P(B \cap A)$ is **not** equal to $P(B)P(A)$.

d. Are A and B mutually exclusive?

Solution 3.14

d. No, they are not because $P(A \cap B) = 0.585$.

To be mutually exclusive, $P(A \cap B)$ must equal zero.

Try It

3.14 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 3.15

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

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

Solution 3.15

a. $\frac{28}{150}$

b. What is the probability that the member practices four times a week?

Solution 3.15

b. $\frac{80}{150}$

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution 3.15

c. $\frac{40}{150}$

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?

Solution 3.15

d. $P(\text{advanced} \cap \text{intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

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

Solution 3.15

e. No, these are not independent events.
$P(\text{novice} \cap \text{practices four times per week}) = 0.0667$
$P(\text{novice})P(\text{practices four times per week}) = 0.0996$
$0.0667 \neq 0.0996$

Try It

3.15 A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Example 3.16

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 | that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, $M \mid S$ = math given speech

a. What is the probability that Felicity enrolls in math and speech?
Find $P(M \cap S) = P(M \mid S)P(S)$.

b. What is the probability that Felicity enrolls in math or speech classes?
Find $P(M \cup S) = P(M) + P(S) - P(M \cap S)$.

c. Are M and S independent? Is $P(M \mid S) = P(M)$?

d. Are M and S mutually exclusive? Is $P(M \cap S) = 0$?

Solution 3.16
a. 0.1625, b. 0.6875, c. No, d. No

Try It

3.16 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 $P(D \mid B) = 0.5$.

a. Find $P(B \cap D)$.

b. Find $P(B \cup D)$.

Example 3.17

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?

Solution 3.17
a. $P(B) = 0.143$; $P(N) = 0.85$

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

Solution 3.17
b. $P(N \mid B) = 0.02$

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

Solution 3.17
c. $P(B \cap N) = P(B)P(N \mid B) = (0.143)(0.02) = 0.0029$

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

Solution 3.17
d. $P(B \cup N) = P(B) + P(N) - P(B \cap N) = 0.143 + 0.85 - 0.0029 = 0.9901$

e. Are having breast cancer and testing negative independent events?

Solution 3.17
e. No. $P(N) = 0.85$; $P(N \mid B) = 0.02$. So, $P(N \mid B)$ does not equal $P(N)$.

f. Are having breast cancer and testing negative mutually exclusive?

Solution 3.17
f. No. $P(B \cap N) = 0.0029$. For B and N to be mutually exclusive, $P(B \cap N)$ must be zero.

Try It

3.17 A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Example 3.18

Refer to the information in **Example 3.17**. P = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P \mid B) = 1 - P(N \mid B)$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \cap P) = P(P \mid B)P(B)$.
c. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

Solution 3.18
a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Try It

3.18 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 $P(D \mid B) = 0.5$.

a. Find $P(B')$.
b. Find $P(D \cap B)$.
c. Find $P(B \mid D)$.
d. Find $P(D \cap B')$.

e. Find $P(D \mid B')$.

3.4 | Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA

Contingency Tables

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

Example 3.19

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

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

Table 3.2

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

Calculate the following probabilities using the table.

a. Find P(Person is a car phone user).

Solution 3.19

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

b. Find P(person had no violation in the last year).

Solution 3.19

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

c. Find P(Person had no violation in the last year AND was a car phone user).

Solution 3.19

c. $\frac{280}{755}$

d. Find P(Person is a car phone user OR person had no violation in the last year).

Solution 3.19

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

e. Find *P*(Person is a car phone user GIVEN person had a violation in the last year).

Solution 3.19

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

f. Find *P*(Person had no violation last year GIVEN person was not a car phone user)

Solution 3.19

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

Try It

3.19 **Table 3.3** shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

Table 3.3

a. What is *P*(athlete stretches before exercising)?

b. What is *P*(athlete stretches before exercising|no injury in the last year)?

Example 3.20

Table 3.4 shows a random sample of 100 hikers and the areas of hiking they prefer.

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

Table 3.4 Hiking Area Preference

a. Complete the table.

Solution 3.20

a.

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

Table 3.5 Hiking Area Preference

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

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

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

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

Solution 3.20

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$

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

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

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

Solution 3.20

c.

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

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

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solution 3.20

d.

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

Try It

3.20 **Table 3.6** shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

Table 3.6

a. Out of the males, what is the probability that the cyclist prefers a hilly path?

b. Are the events “being male” and “preferring the hilly path” independent events?

Example 3.21

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

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

Table 3.7 Door Choice

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

Verify the remaining entries.

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

Solution 3.21

a.

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

Table 3.8 Door Choice

b. What is the probability that Alissa does not catch Muddy?

Solution 3.21

b. $\frac{41}{60}$

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

Solution 3.21

c. $\frac{9}{19}$

Example 3.22

Table 3.9 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					

Table 3.9 United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

TOTAL each column and each row. Total data = 4,520.7

a. Find *P*(2009 AND Robbery).
b. Find *P*(2010 AND Burglary).
c. Find *P*(2010 OR Burglary).
d. Find *P*(2011|Rape).
e. Find *P*(Vehicle|2008).

Solution 3.22

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Try It

3.22 **Table 3.10** relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

Table 3.10

a. Find the total for each row and column

b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Tree Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams can be used to visualize and solve conditional probabilities.

Tree Diagrams

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

Example 3.23

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.

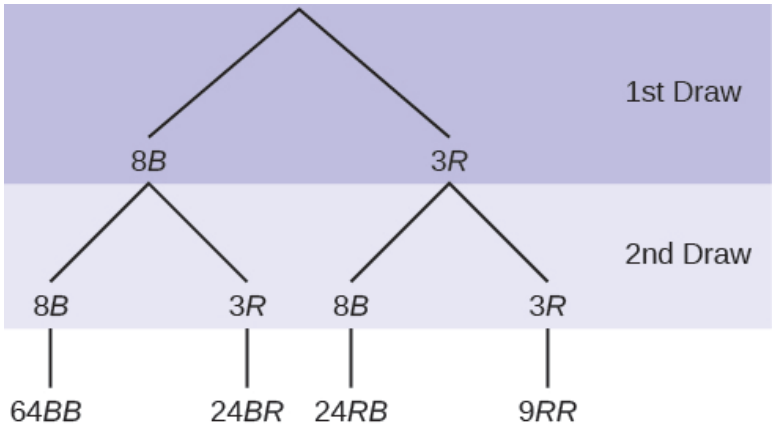


Figure 3.2 Total = 64 + 24 + 24 + 9 = 121

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

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

The other outcomes are similar.

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

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

Solution 3.23

a. *B*1*R*1; *B*1*R*2; *B*1*R*3; *B*2*R*1; *B*2*R*2; *B*2*R*3; *B*3*R*1; *B*3*R*2; *B*3*R*3; *B*4*R*1; *B*4*R*2; *B*4*R*3; *B*5*R*1; *B*5*R*2; *B*5*R*3; *B*6*R*1; *B*6*R*2; *B*6*R*3; *B*7*R*1; *B*7*R*2; *B*7*R*3; *B*8*R*1; *B*8*R*2; *B*8*R*3

b. Using the tree diagram, calculate $P(RR)$.

Solution 3.23

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

Solution 3.23

c. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

Solution 3.23

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

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Solution 3.23

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

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

f. Using the tree diagram, calculate $P(BB)$.

Solution 3.23

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

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.

Solution 3.23

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

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

Try It

3.23 In a standard deck, there are 52 cards. 12 cards are face cards (event *F*) and 40 cards are not face cards (event *N*). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate *P*(*FF*).

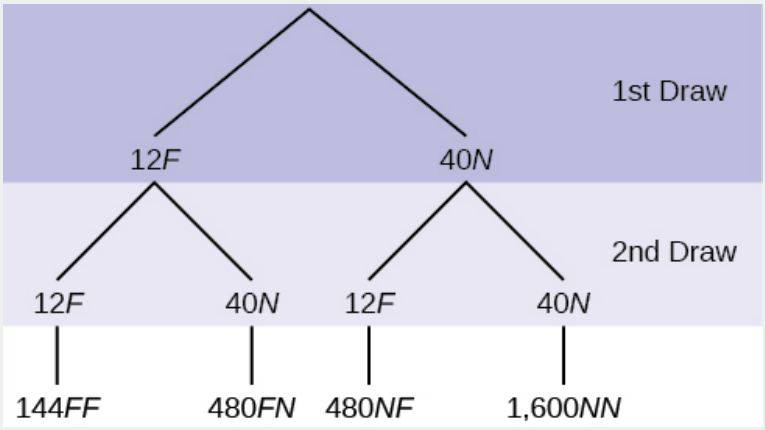


Figure 3.3

Example 3.24

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

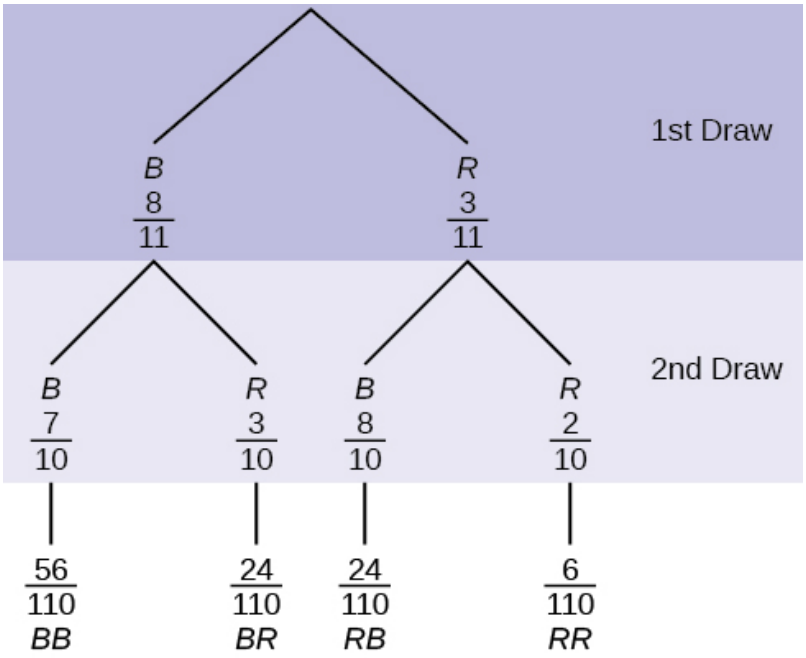


Figure 3.4 Total = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$

NOTE

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

Calculate the following probabilities using the tree diagram.

a. $P(RR)$ = _____

Solution 3.24

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

b. Fill in the blanks:

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

Solution 3.24

b. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

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

Solution 3.24

c. $P(R \text{ on 2nd}|B \text{ on 1st}) = \frac{3}{10}$

d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\_\_\_)(\_\_\_) = \frac{24}{100}$

Solution 3.24

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

e. Find $P(BB)$.

Solution 3.24

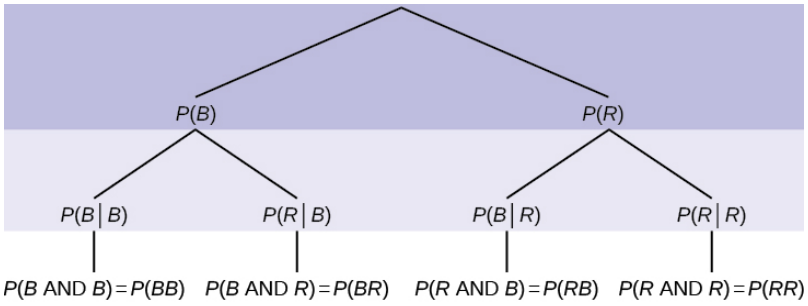
e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

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

Solution 3.24

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

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



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

Try It

3.24 In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

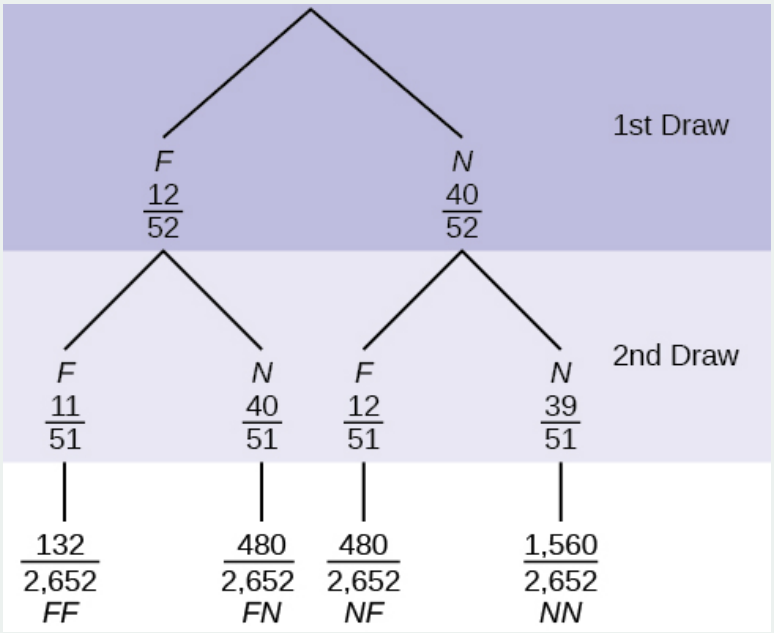


Figure 3.5

a. Find $P(FN \text{ OR } NF)$.

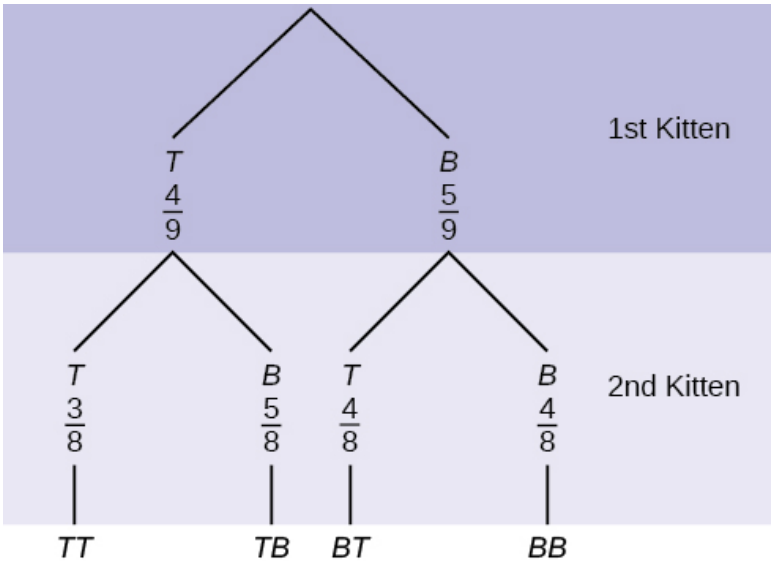
b. Find $P(N|F)$.

c. Find P(at most one face card).
Hint: "At most one face card" means zero or one face card.

d. Find P(at least on face card).
Hint: "At least one face card" means one or two face cards.

Example 3.25

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



a. What is the probability that both kittens are tabby?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. What is the probability that one kitten of each coloring is selected?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?

d. What is the probability of choosing two kittens of the same color?

Solution 3.25

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Try It

3.25 Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

KEY TERMS

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

Contingency Table the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

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; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment a planned activity carried out under controlled conditions

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

Mutually Exclusive Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \cap B) = 0$.

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; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

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

Sample Space the set of all possible outcomes of an experiment

Sampling with Replacement If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement When sampling is done without replacement, each member of a population may be chosen only once.

The Complement Event The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of *A* GIVEN *B* $P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Intersection: the AND Event An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Union: the OR Event An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Tree Diagram the useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

CHAPTER REVIEW

3.1 Terminology -- Probability Topics -- MtRoyal - Version2016RevA

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

3.2 Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

3.3 Two Basic Rules of Probability

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

3.4 Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

FORMULA REVIEW

3.1 Terminology -- Probability Topics -- MtRoyal - Version2016RevA

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

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

3.2 Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

If A and B are independent, $P(A \cap B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

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

3.3 Two Basic Rules of Probability

The multiplication rule: $P(A \cap B) = P(A \mid B)P(B)$

The addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

PRACTICE

3.1 Terminology -- Probability Topics -- MtRoyal - Version2016RevA

1. In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events 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.

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.
Let H = the event of getting a hat.
Let N = the event of getting a noisemaker.
Let F = the event of getting a finger trap.
Let C = the event of getting a bag of confetti.

2. Find $P(H)$.

3. Find $P(N)$.

4. Find $P(F)$.

5. Find $P(C)$.

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

6. Find $P(B)$.

7. Find $P(G)$.

8. Find $P(P)$.

9. Find $P(R)$.

10. Find $P(Y)$.

11. Find $P(O)$.

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).
Let A = the event that a country is in Asia.
Let E = the event that a country is in Europe.
Let F = the event that a country is in Africa.
Let N = the event that a country is in North America.

Let O = the event that a country is in Oceania.
Let S = the event that a country is in South America.

12. Find $P(A)$.

13. Find $P(E)$.

14. Find $P(F)$.

15. Find $P(N)$.

16. Find $P(O)$.

17. Find $P(S)$.

18. What is the probability of drawing a red card in a standard deck of 52 cards?

19. What is the probability of drawing a club in a standard deck of 52 cards?

20. What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

21. What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

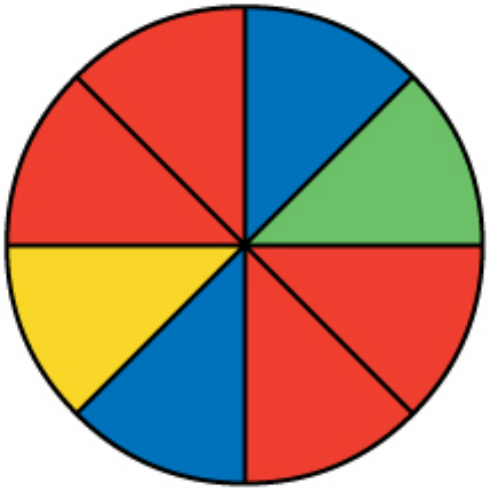


Figure 3.6

Let B = the event of landing on blue.
Let R = the event of landing on red.
Let G = the event of landing on green.
Let Y = the event of landing on yellow.

22. If you land on Y, you get the biggest prize. Find $P(Y)$.

23. If you land on red, you don't get a prize. What is $P(R)$?

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let 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.

24. Write the symbols for the probability that a player is not an outfielder.

25. Write the symbols for the probability that a player is an outfielder or is a great hitter.

26. Write the symbols for the probability that a player is an infielder and is not a great hitter.

27. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

28. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

29. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

30. Write the symbols for the probability that of all the great hitters, a player is an outfielder.

31. Write the symbols for the probability that a player is an infielder or is not a great hitter.

32. Write the symbols for the probability that a player is an outfielder and is a great hitter.

33. Write the symbols for the probability that a player is an infielder.

34. What is the word for the set of all possible outcomes?

35. What is conditional probability?

36. 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?

37. What is the sum of the probabilities of an event and its complement?

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.

38. What does $P(E|M)$ mean in words?

39. What does $P(E \text{ OR } M)$ mean in words?

3.2 Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

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

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

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

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

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

3.3 Two Basic Rules of Probability

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

44. Find $P(C)$.

45. Find $P(L)$.

46. Find $P(C \mid L)$.

47. In words, what is $C \mid L$?

48. Find $P(L \cap C)$.

49. In words, what is $L \cap C$?

50. Are L and C independent events? Show why or why not.

51. Find $P(L \cup C)$.

52. In words, what is $L \cup C$?

53. Are L and C mutually exclusive events? Show why or why not.

HOMEWORK

3.1 Terminology -- Probability Topics -- MtRoyal - Version2016RevA

54.

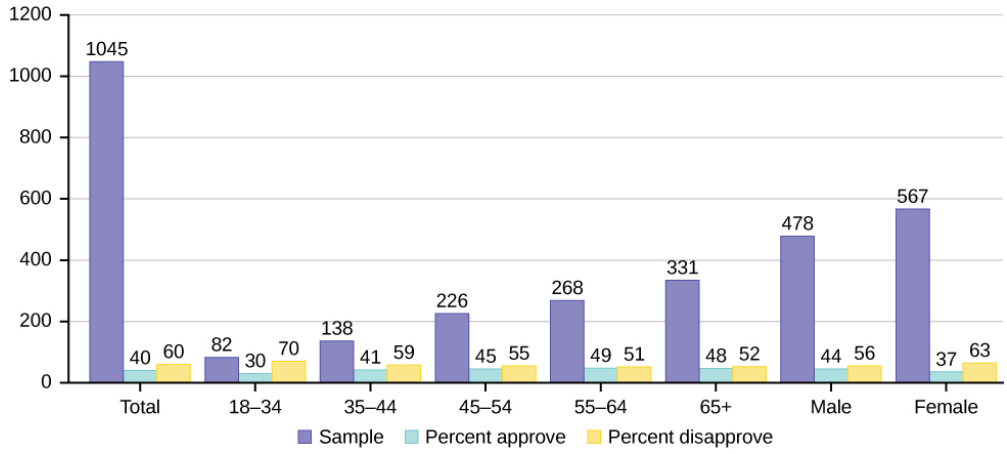


Figure 3.7 The graph in **Figure 3.7** displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

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 $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

55. 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.

3.2 Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA

Use the following information to answer the next 12 exercises. 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.

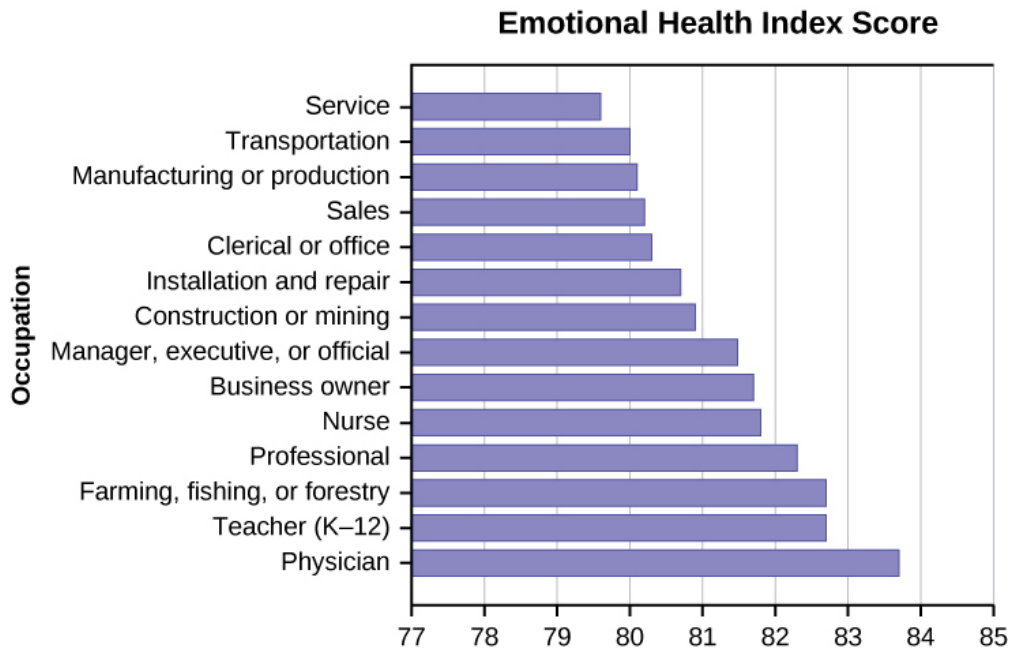


Figure 3.8

56. Find the probability that an Emotional Health Index Score is 82.7.

57. Find the probability that an Emotional Health Index Score is 81.0.

58. Find the probability that an Emotional Health Index Score is more than 81?

59. Find the probability that an Emotional Health Index Score is between 80.5 and 82?

60. If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

61. What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

62. What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

63. What occupation has the highest emotional index score?

64. What occupation has the lowest emotional index score?

65. What is the range of the data?

66. Compute the average EHIS.

67. 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?

3.3 Two Basic Rules of Probability

68. 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.

In this problem, let:

- 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 $P(C \mid A)$.
d. Find $P(B|C)$.
e. In words, what is $C \mid A$?
f. In words, what is $B \mid C$?
g. Find $P(C \cap B)$.
h. In words, what is $C \cap B$?
i. Find $P(C \cup B)$.
j. Are C and B mutually exclusive events? Show why or why not.

69. 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?

Use the following information to answer the next three exercises. 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.

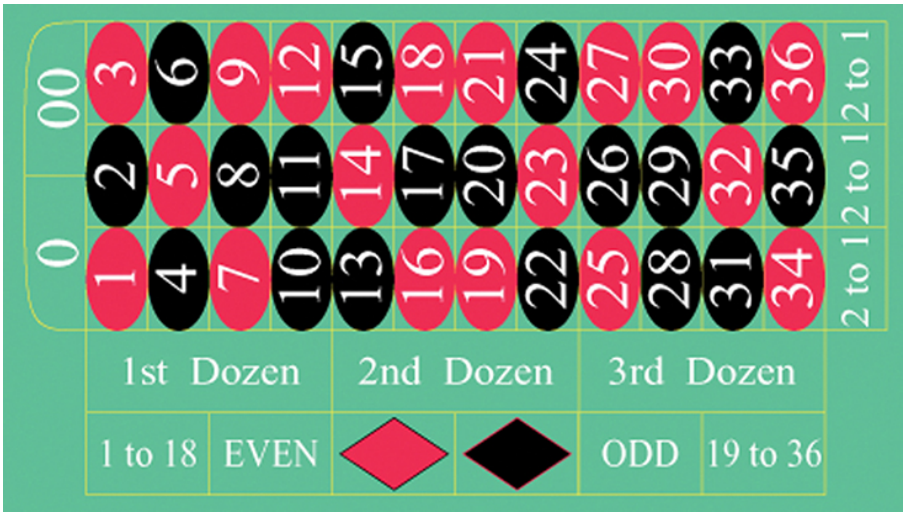


Figure 3.9 (credit: film8ker/wikibooks)

70.

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
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?

71. 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

72. 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)

73. 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.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(G) =$ _____
c. $P(G \mid E) =$ _____
d. $P(G \cap E) =$ _____
e. $P(G \cup E) =$ _____
f. Are G and E mutually exclusive? Justify your answer numerically.

74. 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 "$P(A \mid B)$" represents. Find $P(A \mid B)$.
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.

75. 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.

76. 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.

77. 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.

78. Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C \mid D) = 0.6$.

a. Find $P(C \cap D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(C \cup D)$.
e. Find $P(D \mid C)$.

79. Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \cup Z) = 0.71$ and $P(Y) = 0.42$.

80. G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

a. Explain why the following statement MUST be false: $P(H \mid G) = 0.4$.
b. Find $P(H \cup G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

81. 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. $P(S \cap E')$ =	iii. 0.1957
d. $P(S \mid E')$ =	iv. 0.1219

Table 3.11

82. 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 and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

83. 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 and explain it.
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.

84. 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	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Table 3.12

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit is a double})$

85. 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 bloodtype. 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.

86. 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.

87. 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).

88. 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 $P(D \cap E)$.
b. Find $P(E \mid D)$.
c. Find $P(D \cup E)$.
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.

BRINGING IT TOGETHER: HOMEWORK

89. A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into **Table 3.13**.

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

Table 3.13

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\# 1–33}|\leq 210 \text{ pounds})$?

90. 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.

91. Given events G and H: $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) = 0.14$

a. Find $P(H \text{ OR } G)$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

92. Given events J and K: $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

a. Find $P(J \text{ AND } K)$.
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J AND K).

REFERENCES

3.1 Terminology -- Probability Topics -- MtRoyal - Version2016RevA

"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 -- Probaility Topics -- MtRoyal - Version2016RevA

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 2,2 013).

3.4 Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA

"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).

"Human Blood Types." Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

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).

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).

SOLUTIONS

1

a. $P(L') = P(S)$

b. $P(M \text{ OR } S)$

c. $P(F \text{ AND } L)$

d. $P(M|L)$

e. $P(L|M)$

f. $P(S|F)$

g. $P(F|L)$

h. $P(F \text{ OR } L)$

i. $P(M \text{ AND } S)$

j. $P(F)$

3 $P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

5 $P(C) = \frac{5}{42} = 0.12$

7 $P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

9 $P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

11 $P(O) = \frac{150 - 22 - 38 - 20 - 28 - 26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

13 $P(E) = \frac{47}{194} = 0.24$

15 $P(N) = \frac{23}{194} = 0.12$

17 $P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

19 $\frac{13}{52} = \frac{1}{4} = 0.25$

21 $\frac{3}{6} = \frac{1}{2} = 0.5$

23 $P(R) = \frac{4}{8} = 0.5$

25 $P(O \text{ OR } H)$

27 $P(H|I)$

29 $P(N|O)$

31 $P(I \text{ OR } N)$

33 $P(I)$

35 The likelihood that an event will occur given that another event has already occurred.

37 1

39 the probability of landing on an even number or a multiple of three

41 $P(J) = 0.3$

43 $P(Q \text{ AND } R) = P(Q)P(R)$ $0.1 = (0.4)P(R)$ $P(R) = 0.25$

45 0.376

47 $C \mid L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

49 $L \cap C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

51 0.6492

53 No, because $P(L \cap C)$ does not equal 0.

55

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%

b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

57 0

59 0.3571

61 0.2142

63 Physician (83.7)

65 83.7 − 79.6 = 4.1

67 $P(\text{Occupation} < 81.3) = 0.5$

69

a. The Forum Research surveyed 1,046 Torontonians.

b. 58%

c. 42% of 1,046 = 439 (rounding to the nearest integer)

d. 0.57

e. 0.60.

71

a. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$

b. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$

c. $P(\text{Bettting on one number}) = \frac{1}{38}$

d. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$

e. $P(\text{Betting on two number that touch each other on the table }) = \frac{2}{38}$

f. $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$

g. $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

73

a. {*G*1, *G*2, *G*3, *G*4, *G*5, *Y*1, *Y*2, *Y*3}

b. $\frac{5}{8}$

c. $\frac{2}{3}$

d. $\frac{2}{8}$

e. $\frac{6}{8}$

f. No, because $P(G \cap E)$ does not equal 0.

75

NOTE

The coin toss is independent of the card picked first.

a. {(*G*,*H*) (*G*,*T*) (*B*,*H*) (*B*,*T*) (*R*,*H*) (*R*,*T*)}

b. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$

c. Yes, *A* and *B* are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \cap B) = 0$

d. No, *A* and *C* are not mutually exclusive because they can occur at the same time. In fact, *C* includes all of the outcomes of *A*; if the card chosen is blue it is also (red or blue). $P(A \cap C) = P(A) = \frac{3}{20}$

77

a. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$

b. $\frac{4}{8}$

c. Yes, because if *A* has occurred, it is impossible to obtain two tails. In other words, $P(A \cap B) = 0$.

79

a. If *Y* and *Z* are independent, then $P(Y \cap Z) = P(Y)P(Z)$, so $P(Y \cup Z) = P(Y) + P(Z) - P(Y)P(Z)$.

b. 0.5

81 iii; i; iv; ii

83

a. $P(R) = 0.44$

b. $P(R \mid E) = 0.56$

c. $P(R \mid O) = 0.31$

d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R \mid E) \neq P(R)$.

e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R \mid E) > P(R)$.

85

a. $P(\text{type O} \cup \text{Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O} \cap \text{Rh-})$

$0.52 = 0.43 + 0.15 - P(\text{type O} \cap \text{Rh-})$; solve to find $P(\text{type O} \cap \text{Rh-}) = 0.06$

6% of people have type O, Rh- blood

b. $P(\text{NOT}(\text{type O} \cap \text{Rh-})) = 1 - P(\text{type O} \cap \text{Rh-}) = 1 - 0.06 = 0.94$

94% of people do not have type O, Rh- blood

87

a. Let *C* = be the event that the cookie contains chocolate. Let *N* = the event that the cookie contains nuts.

b. $P(C \cup N) = P(C) + P(N) - P(C \cap N) = 0.36 + 0.12 - 0.08 = 0.40$

c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \cup N) = 1 - 0.40 = 0.60$

90

a. $P(C) = 0.4567$

b. not enough information

c. not enough information

d. No, because over half (0.51) of men have at least one false positive text

92

a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37 - P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$

b. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 0.10 = 0.90$

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

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 OBJECTIVE

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.

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 historical 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, that is, the random variable can only take on whole number values. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variable Notation

The upper case letter X denotes a random variable. Lower case letters like x or y denote the value of a random variable. If ***X* is a random variable, then *X* is written in words, and *x* is given as a number.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values as whole numbers that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Probability Density Functions (PDF) for a Random Variable

A **probability density function** or **probability distribution function** has two characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

A probability density function is a mathematical formula that calculates probabilities for specific types of events, what we have been calling experiments. There is a sort of magic to a probability density function (Pdf) partially because the same formula often describes very different types of events. For example, the binomial Pdf will calculate probabilities for flipping coins, yes/no questions on an exam, opinions of voters in an up or down opinion poll, indeed any binary event. Other probability density functions will provide probabilities for the time until a part will fail, when a customer will arrive at the turnpike booth, the number of telephone calls arriving at a central switchboard, the growth rate of a bacterium, and on and on. There are whole families of probability density functions that are used in a wide variety of applications, including medicine, business and finance, physics and engineering, among others.

For our needs here we will concentrate on only a few probability density functions as we develop the tools of inferential statistics.

Remember, a probability density function computes probabilities for us. We simply put the appropriate numbers in the formula and we get the probability of specific events. However, for these formulas to work they must be applied only to cases for which they were designed.

4.1 | Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

A binomial process, often called a Bernoulli process after the first person to fully develop its properties, is any case where there are only two possible outcomes in any one trial, called successes and failures. It gets its name from the binary number system where all numbers are reduced to either 1's or 0's, which is the basis for computer technology and CD music recordings.

Binomial Formula

$$b(x) = \binom{n}{x} p^x q^{n-x}$$

where b(x) is the probability of X successes in n trials when the probability of a success in ANY ONE TRIAL is p. And of course q=(1-p) and is the probability of a failure in any one trial.

We can see now why the combinatorial formula is also called the binomial coefficient because it reappears here again in the binomial probability function. For the binomial formula to work, the probability of a success in any one trial must be the same from trial to trial, or in other words, the outcomes of each trial must be independent. Flipping a coin is a binomial process because the probability of getting a head in one flip does not depend upon what has happened in PREVIOUS flips. (At this time it should be noted that using p for the parameter of the binomial distribution is a violation of the rule that population parameters are designated with Greek letters. In many textbooks θ (pronounced theta) is used instead of p and this is how it should be.

Just like a set of data, a probability density function has a mean and a standard deviation. For the binomial distribution these are given by the formulas:

$$\mu = np$$

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

Notice that p is the only parameter in these equations. The binomial distribution is thus seen as coming from the one-parameter family of probability distributions. In short, we know all there is to know about the binomial once we know p, the probability of a success in any one trial.

In probability theory, under certain circumstances, one probability distribution can be used to approximate another. We say that one is the limiting distribution of the other. If a small number is to be drawn from a large population, even if there is no replacement, we can still use the binomial even thought this is not a binomial process. If there is no replacement it violates the independence rule of the binomial. Nevertheless, we can use the binomial to approximate a probability that is really a hypergeometric distribution if we are drawing fewer than 10 percent of the population, i.e. X is less than 10 percent of N in the formula for the hypergeometric function. The rationale for this argument is that when drawing a small percentage of the population we do not alter the probability of a success from draw to draw in any meaningful way. Imagine drawing from not one deck of 52 cards but from 6 decks of cards. The probability of say drawing an ace does not change the conditional probability of what happens on a second draw in the same way it would if there were only 4 aces rather than the 24 aces now to draw from. This ability to use one probability distribution to estimate others will become very valuable to us later.

There are three characteristics of 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. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Think of this as drawing WITH replacement. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then $\sigma = \sqrt{npq}$.

Any experiment that has characteristics three and 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 4.1

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Try It

4.1 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 20 attempts, you want to find the probability that the dolphin succeeds 12 times. Find the P(X=12) using the binomial Pdf.

Example 4.2

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Solution 4.2
$P(x > 10)$

Example 4.3

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

a. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.

Solution 4.3
a. failure

b. If we are interested in the number of students who do their homework on time, then how do we define X?

Solution 4.3
b. X = the number of statistics students who do their homework on time

c. What values does x take on?

Solution 4.3
c. 0, 1, 2, …, 50

d. What is a "failure," in words?

Solution 4.3

d. Failure is defined as a student who does not complete his or her homework on time.

The probability of a success is $p = 0.70$. The number of trials is $n = 50$.

e. If $p + q = 1$, then what is q?

Solution 4.3

e. $q = 0.30$

f. The words "at least" translate as what kind of inequality for the probability question $P(x$ ____ $40)$.

Solution 4.3

f. greater than or equal to ($\geq$)
The probability question is $P(x \geq 40)$.

Try It

4.3 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.

Try It

4.3 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. 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.

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 the same for any trial).

Binomial Experiment a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

Binomial Probability Distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is

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

Probability Distribution Function (PDF) a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Random Variable (RV) a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

CHAPTER REVIEW

4.0 Introduction -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

The characteristics of a probability distribution or density function (PDF) are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

4.1 Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

The formula for the Binomial probability density function is

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x q^{(n-x)}$$

FORMULA REVIEW

4.1 Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

$X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x = 0, 1, 2, 3, ..., n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1$

$q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x q^{(n-x)}$$

where P(X) is the probability of X successes in n trials when the probability of a success in ANY ONE TRIAL is p.

PRACTICE

4.0 Introduction -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

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 **Table 4.1** 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

Table 4.1

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 six 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

Table 4.2

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 four 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. $P(x < 3) =$ _____

17. Find the probability that Javier volunteers for at least one event each month. $P(x > 0) =$ _____

4.1 Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

18. In words, define the random variable X.

19. $X \sim$ _____(_____,_____)

20. What values does the random variable X take on?

21. Construct the probability distribution function (PDF).

x	*P*(*x*)

Table 4.3

22. On average (μ), how many would you expect to answer yes?

23. What is the standard deviation (σ)?

24. What is the probability that at most five of the freshmen reply "yes"?

25. What is the probability that at least two of the freshmen reply "yes"?

HOMEWORK

4.1 Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

26. 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.

27. Define the random variable and list its possible values.

28. State the distribution of X.

29. Find the probability that at least four of the 25 patients actually have the flu.

30. On average, for every 25 patients calling in, how many do you expect to have the flu?

31. 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 **Table 4.4**. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	*P*(*x*)
0	0.03
1	0.50
2	0.24
3	
4	0.07
5	0.04

Table 4.4

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.

32. 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.

33. The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

Let X = the number of games won in that upcoming month.

34. 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

35. What is the probability that the San Jose Sharks win at least five games in that upcoming month

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

36. 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.

37. 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.

38. 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. On average, how many dice would you expect to show a one?
d. Find the probability that all six dice show a one.
e. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

39. 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. $X \sim$ _____(_____,_____)
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.

40. 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. $X \sim$ _____(_____,_____)
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.

41. 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. $X \sim$ _____(_____,_____)
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.

42. 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. $X \sim$ _____(_____,_____)
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. Based upon numerical values, 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.

43. 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. $X \sim$ _____(_____,_____)
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.

44. 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. $X \sim$ _____(_____,_____)
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?

45. 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. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
d. Calculate the average expected matches over the long run of playing this game for the player.
e. Calculate the average expected earnings over the long run of playing this game for the player.
f. Determine who has the advantage, the player or the house.

46. 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. 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.

47. 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 (i) mean and (ii) 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.

REFERENCES

4.1 Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA

"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).

SOLUTIONS

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

Table 4.5

3 0.10 + 0.05 = 0.15

5 1

7 0.35 + 0.40 + 0.10 = 0.85

9 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

Table 4.6

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

Table 4.7

17 1 – 0.05 = 0.95

18 X = the number that reply “yes”

20 0, 1, 2, 3, 4, 5, 6, 7, 8

22 5.7

24 0.4151

27 X = the number of patients calling in claiming to have the flu, who actually have the flu. X = 0, 1, 2, ...25

29 0.0165

31

a. X = the number of DVDs a Video to Go customer rents
b. 0.12
c. 0.11
d. 0.77

33 d. 4.43

35 c

37

- X = number of questions answered correctly
- $X \sim B\left(32, \frac{1}{3}\right)$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

39

a. X = the number of college and universities that offer online offerings.
b. 0, 1, 2, …, 13
c. $X \sim B(13, 0.96)$
d. 12.48
e. 0.0135
f. $P(x = 12) = 0.3186$ $P(x = 13) = 0.5882$ More likely to get 13.

41

a. X = the number of fencers who do **not** use the foil as their main weapon

b. 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.

43

a. X = the number of audits in a 20-year period

b. 0, 1, 2, …, 20

c. $X \sim B(20, 0.02)$

d. 0.4

e. 0.6676

f. 0.0071

45

1. X = the number of matches
2. 0, 1, 2, 3
3. In dollars: −1, 1, 2, 3
4. $\frac{1}{2}$
5. The answer is −0.0787. You lose about eight cents, on average, per game.
6. The house has the advantage.

47

a. $X \sim B(15, 0.281)$

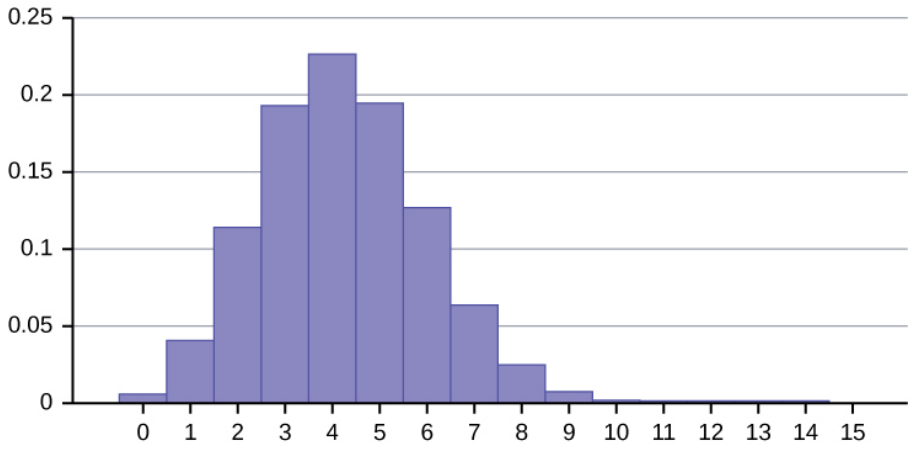


Figure 4.2

b. i. Mean = $\mu = np = 15(0.281) = 4.215$

ii. Standard Deviation = $\sigma = \sqrt{npq} = \sqrt{15(0.281)(0.719)} = 1.7409$

c. $P(x > 5)=1 – 0.7754 = 0.2246$
$P(x = 3) = 0.1927$

$P(x = 4) = 0.2259$
It is more likely that four people are literate that three people are.

5 | THE NORMAL DISTRIBUTION

Figure 5.1 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 OBJECTIVE

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

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal probability density function, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may

use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

The normal distribution is extremely important, but it cannot be applied to everything in the real world. Remember here that we are still talking about the distribution of population data. This is a discussion of probability and thus it is the population data that may be normally distributed, and if it is, then this is how we can find probabilities of specific events just as we did for population data that may be binomially distributed or Poisson distributed. This caution is here because in the next chapter we will see that the normal distribution describes something very different from raw data and forms the foundation of inferential statistics.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing the following formula of the normal probability density function:

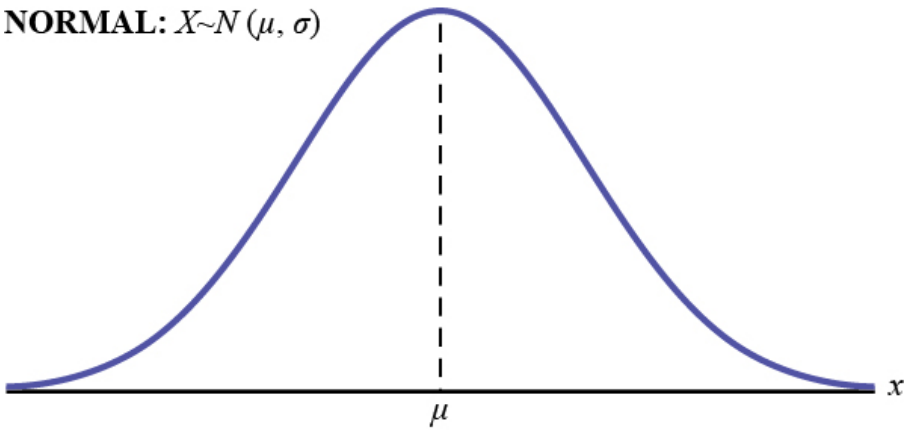


Figure 5.2

The curve is symmetrical about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Note that this is unlike several probability density functions we have already studied, such as the Poisson, where the mean is equal to μ and the standard deviation simply the square root of the mean, or the binomial, where p is used to determine both the mean and 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 and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

5.1 | The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value x = 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

The *z*-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. What this does is dramatically simplify the mathematical calculation of probabilities. Take a moment and substitute zero and one in the appropriate places in the above formula and you can see that the equation collapses into one that can be much more easily solved using integral calculus. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x comes from a known normal distribution with known mean μ and known standard deviation σ. The *z*-score tells how many standard deviations a particular x is away from the mean.

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, 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 5.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then:

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

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Now suppose $x = 1$. Then: $z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$ (rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$.

Example 5.2

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. $X \sim N(5, 2)$.

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 5.2

$$Z = \frac{x - \mu}{\sigma} = \frac{-3-5}{2} = -4$$

$z =$ **-4**. This z-score tells you that $x = -3$ is **four** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The *z*-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Try It

5.2 Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The *z*–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

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

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

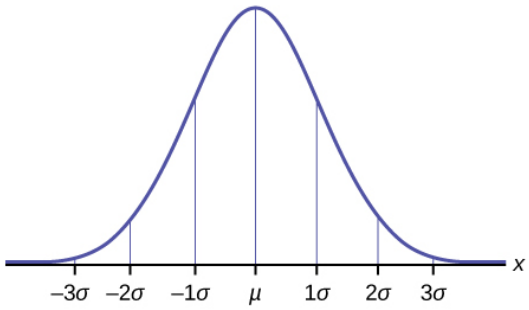


Figure 5.3

Example 5.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The *z*-score when $x = 168$ cm is $z =$ _____. This *z*-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Solution 5.3

$$Z = \frac{x - \mu}{\sigma} = \frac{168 - 170}{6.28} = -0.32$$

a. –0.32, 0.32, left, 170

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = 1.27$. What is the male's height? The *z*-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution 5.3

$$Z = \frac{x - \mu}{\sigma} = \frac{x - 170}{6.28} = 1.27 \rightarrow 1.27 * 6.28 + 170 = 177.98$$

b. 177.98, 1.27, right

Try It

5.3 In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the *z*-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each *z*-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Example 5.4

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between $–1\sigma = (–1)(6) = –6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 – 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50. The *z*-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie between $–2\sigma = (–2)(6) = –12$ and $2\sigma = (2)(6) = 12$. The values $50 – 12 = 38$ and $50 + 12 = 62$ are within two standard deviations of the mean 50. The *z*-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between $–3\sigma = (–3)(6) = –18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 – 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The *z*-scores are –3 and +3 for 32 and 68, respectively.

Try It

5.4 Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Try It

5.4 The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

b. About 95% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

c. About 99.7% of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

5.2 | Using the Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

The shaded area in the following graph indicates the area to the right of *x*. This area is represented by the probability $P(X > x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X > x)$.

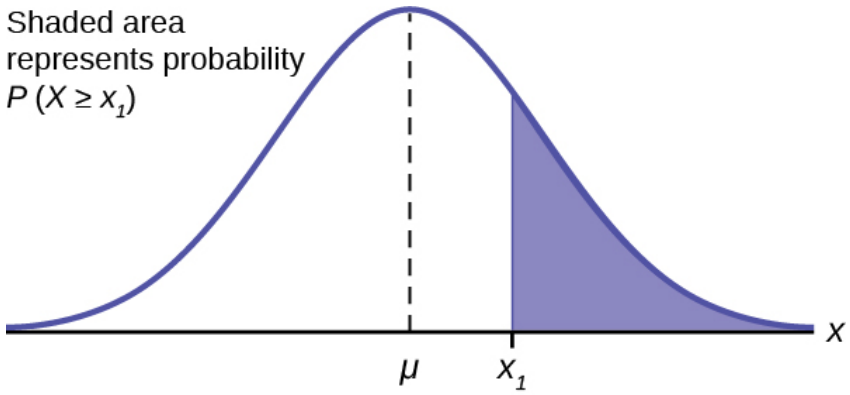


Figure 5.4

The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x)$ = **Area to the left** of the vertical line through *x*. $P(X < x) = 1 – P(X < x)$ = **Area to the right** of the vertical line through *x*. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

To find the probability for probability curves with a continuous random variable we need to calculate the area under the curve across the values of X we are interested in. For the normal distribution this seems a difficult task given the complexity of the formula. There is, however, a simply way to get what we want.

We start knowing that the area under a probability curve is the probability.

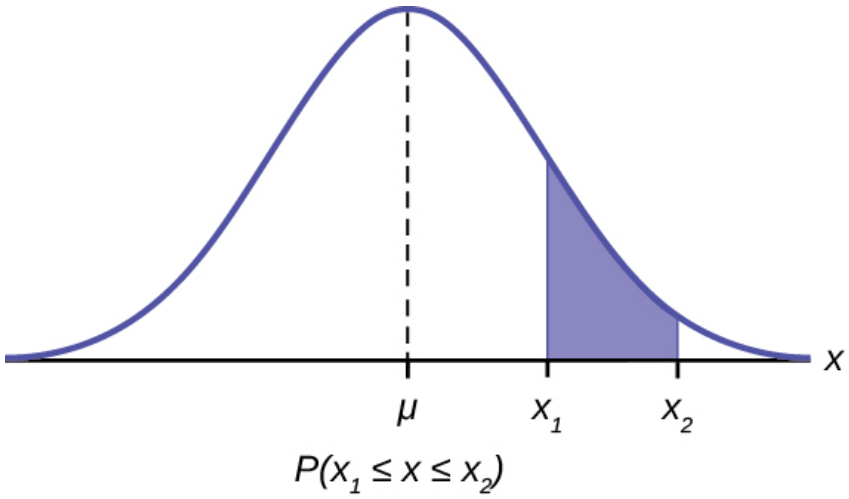


Figure 5.5

This shows that the area between X_1 and X_2 is the probability as stated in the formula: P ($X_1 \le x \le X_2$)

The mathematical tool needed to find the area under a curve is integral calculus. The integral of the normal probability density function between the two points x_1 and x_2 is the area under the curve between these two points and is the probability between these two points.

Doing these integrals is no fun and can be very time consuming. But now, remembering that there are an infinite number of normal distributions out there, we can consider the one with a mean of zero and a standard deviation of 1. This particular normal distribution is given the name Standard Normal Distribution. Putting these values into the formula it reduces to a very simple equation. We can now quite easily calculate all probabilities for any value of x, for this particular normal distribution, that has a mean of zero and a standard deviation of 1. These have been produced and are available here in the text or everywhere on the web. They are presented in various ways. The table in this text is the most common presentation and is set up with probabilities for one-half the distribution beginning with zero, the mean, and moving outward. The shaded area in the graph at the top of the table represents the probability from zero to the specific Z value noted on the horizontal axis, Z.

The only problem is that even with this table, it would be a ridiculous coincidence that our data had a mean of zero and a standard deviation of one. The solution is to convert the distribution we have with its mean and standard deviation to this new Standard Normal Distribution. The Standard Normal has a random variable called Z.

Using the standard normal table, typically called the normal table, to find the probability of one standard deviation, go to the Z column, reading down to 1.0 and then read at column 0. That number, 0.3413 is the probability from zero to 1 standard deviation. At the top of the table is the shaded area in the distribution which is the probability for one standard deviation. The table has solved our integral calculus problem. But only if our data has a mean of zero and a standard deviation of 1.

However, the essential point here is, the probability for one standard deviation on one normal distribution is the same on every normal distribution. If the population data set has a mean of 10 and a standard deviation of 5 then the probability from 10 to 15, one standard deviation, is the same as from zero to 1, one standard deviation on the standard normal distribution. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the **standardizing formula**:

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

where Z is the value on the standard normal distribution, X is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability so we always are dealing with the population, with **known** parameter values and a **known** distribution. It is also important to note that because the normal distribution is symmetrical it does not matter if the z-score is positive or negative when calculating a probability. One standard deviation to the left (negative Z-score) covers the same area as one standard deviation to the right (positive Z-score). This fact is why the Standard Normal tables do not provide areas for the left side of the distribution. Because of this symmetry, the Z-score formula is sometimes written as:

$$Z = \frac{|x - \mu|}{\sigma}$$

Where the vertical lines in the equation means the absolute value of the number.

What the standardizing formula is really doing is computing the number of standard deviations X is from the mean of its own distribution. The standardizing formula and the concept of counting standard deviations from the mean is the secret of all that we will do in this statistics class. The reason this is true is that **all** of statistics boils down to variation, and the counting of standard deviations is a measure of variation.

This formula, in many disguises, will reappear over and over throughout this course.

Example 5.5

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.

Solution 5.5

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$

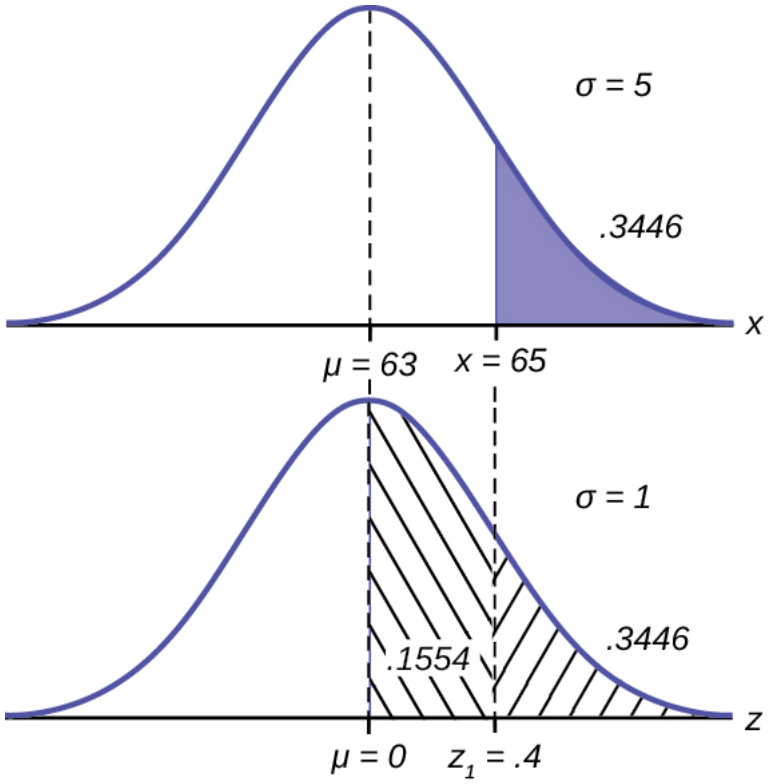


Figure 5.6

$$Z_1 = \frac{x_1 - \mu}{\sigma} = \frac{65 - 63}{5} = 0.4$$

$P(x \geq x_1) = P(Z \geq Z_1) = 0.3446$

The probability that any student selected at random scores more than 65 is 0.3446. Here is how we found this answer.

The normal table provides probabilities from zero to the value Z_1. For this problem the question can be written as: $P(X \geq 65) = P(Z \geq Z_1)$, which is the area in the tail. To find this area the formula would be $0.5 – P(X \leq 65)$. One half of the probability is above the mean value because this is a symmetrical distribution. The graph shows how to find the area in the tail by subtracting that portion from the mean, zero, to the Z_1 value. The final answer is: $P(X \geq 63) = P(Z \geq 0.4) = 0.3446$

$z = \frac{65 – 63}{5} = 0.4$

Area to the left of Z_1 to the mean of zero is 0.1554

$P(x > 65) = P(z > 0.4) = 0.5 – 0.1554 = 0.3446$

Solution 5.5

b.

$Z = \frac{x - \mu}{\sigma} = \frac{85 - 63}{5} = 4.4$ which is larger than the maximum value on the Standard Normal Table. Therefore, the probability that one student scores less than 85 is approximately one or 100%.

A score of 85 is 4.4 standard deviations from the mean of 63 which is beyond the range of the standard normal table. Therefore, the probability that one student scores less than 85 is approximately one (or 100%).

Try It

5.5 The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

Example 5.6

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution 5.6

a. Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

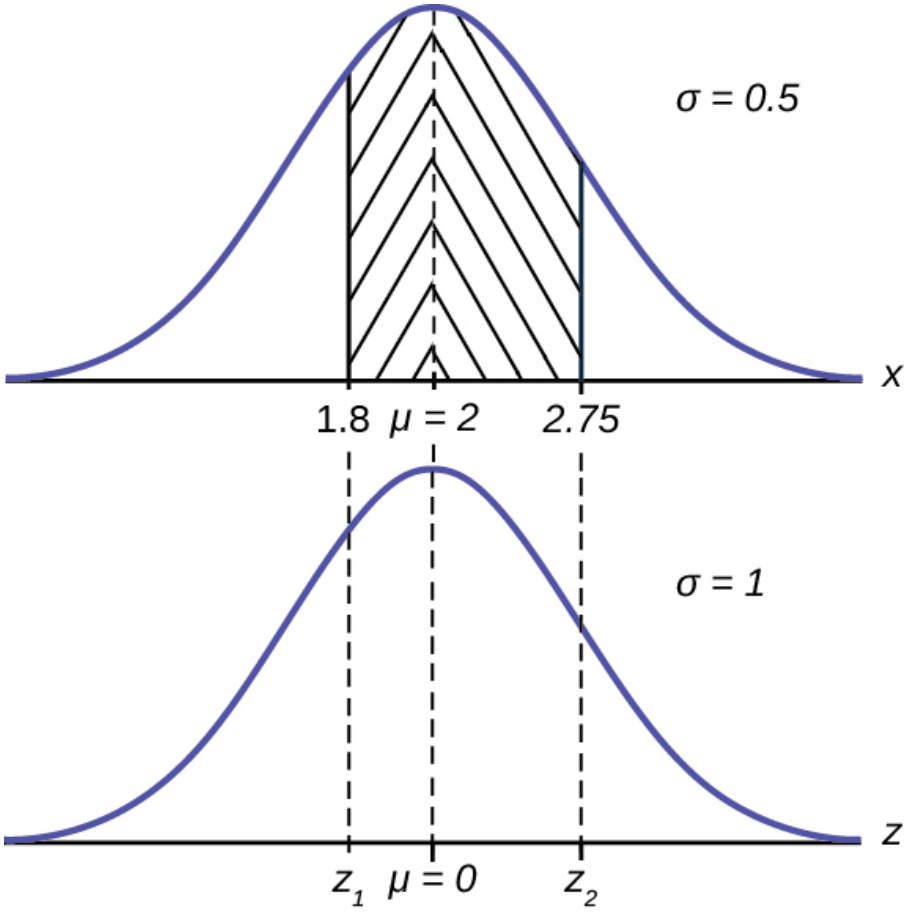


Figure 5.7

$P(1.8 \leq x \leq 2.75) = P(Z_i \leq Z \leq Z_2)$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution 5.6

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.

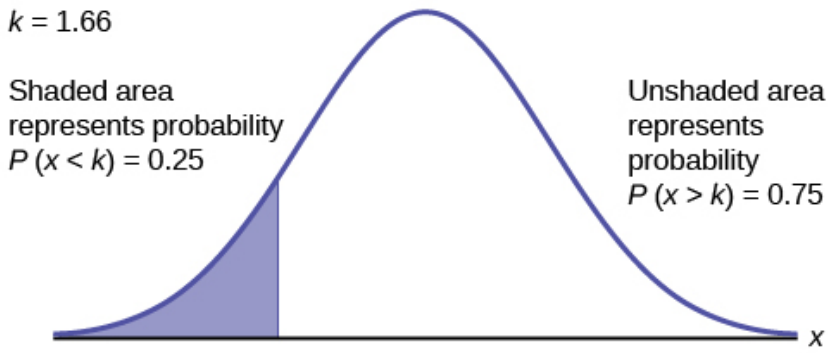


Figure 5.8

$f(Z) = 0.5 - 0.25 = 0.25$, *therefore* $Z \approx -0.675$ (*or just 0.67 using the table*) $Z = \frac{x - \mu}{\sigma} = \frac{x - 2}{0.5} = -0.675$, *therefore* $x = -0.675 * 0.5 + 2 = 1.66$ *hours.*

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

5.6 The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Example 5.7

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution 5.7
a. 0.8186

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution 5.7
b. 0.8413

Example 5.8

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.

Solution 5.8

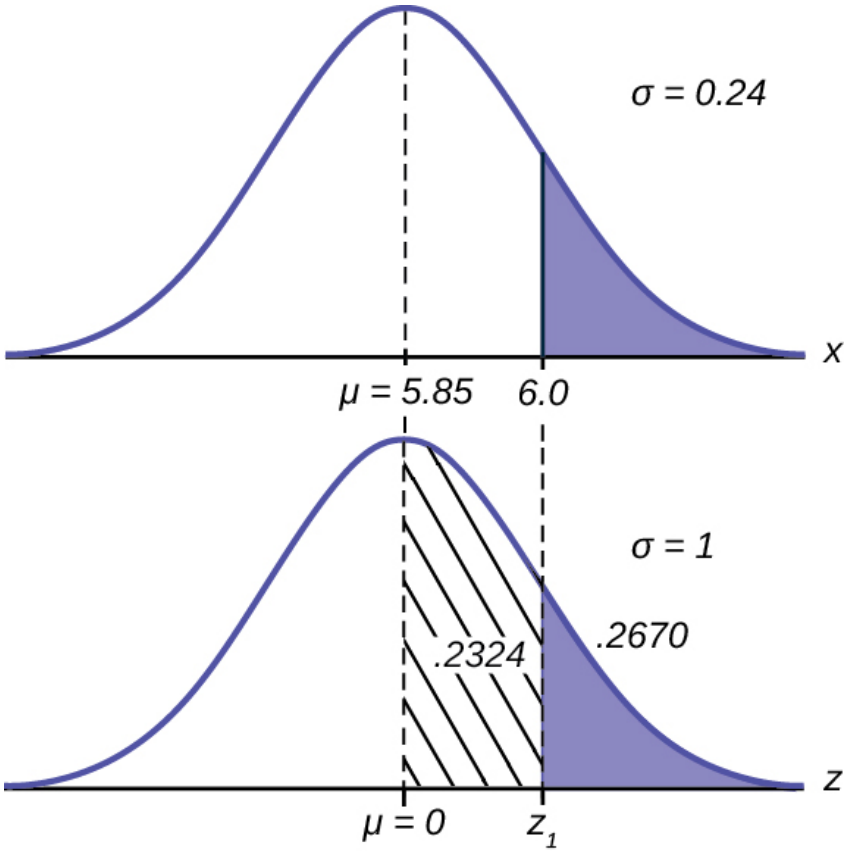


Figure 5.9

$$Z_1 = \frac{6 - 5.85}{.24} = .625$$

$P(x \geq 6) = P(z \geq 0.625) = 0.2670$

b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.

Solution 5.8

$f(Z) = \frac{0.20}{2} = 0.10$, *therefore* $Z \approx \pm 0.25$

$Z = \frac{x - \mu}{\sigma} = \frac{x - 5.85}{0.24} = \pm 0.25 \rightarrow \pm 0.25 * 0.24 + 5.85 = (5.79,\ 5.91)$

KEY TERMS

Normal Distribution a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the **standard normal distribution**.

Standard Normal Distribution a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score the linear transformation of the form $z = \frac{x - \mu}{\sigma}$ or written as $z = \frac{|x - \mu|}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

CHAPTER REVIEW

5.1 The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

FORMULA REVIEW

5.0 Introduction -- The Normal Distribution -- Mt Royal University -- Version 2016RevA

$X \sim N(\mu, \sigma)$

μ = the mean; σ = the standard deviation

5.1 The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

$Z \sim N(0, 1)$

z = a standardized value (z-score)

mean = 0; standard deviation = 1

To find the K^{th} percentile of X when the z-scores is known: $k = \mu + (z)\sigma$

z-score: $z = \frac{x - \mu}{\sigma}$ or $z = \frac{|x - \mu|}{\sigma}$

Z = the random variable for z-scores

$Z \sim N(0, 1)$

PRACTICE

5.1 The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

1. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. X = _____.

2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

3. $X \sim N(1, 2)$

σ = _____

4. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. X = _____.

5. $X \sim N(-4, 1)$

What is the median?

6. $X \sim N(3, 5)$

σ = _____

7. $X \sim N(-2, 1)$

μ = _____

8. What does a z-score measure?

9. What does standardizing a normal distribution do to the mean?

10. Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

11. What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

12. What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

13. What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

14. What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

15. Suppose $X \sim N(2, 6)$. What value of x has a z-score of three?

16. Suppose $X \sim N(8, 1)$. What value of x has a z-score of –2.25?

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 two 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 $X \sim N(-1, 2)$. What is the z-score of $x = 2$?

23. Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

24. Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

25. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

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 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

35. Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution(i.e., –3).

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 betwween the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

41. Define the random variable X in words. X = _____.

42. $X \sim$ _____(_____,_____)

HOMEWORK

5.1 The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

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.

43. What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

44. What is the z-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

45. 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

46. 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, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ 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.

47. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ = 125 and standard deviation σ = 14. Systolic blood pressure for males follows a normal distribution.

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?

48. 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 μ = 125 and standard deviation σ = 14. If X = a systolic blood pressure score then $X \sim N$ (125, 14).

a. Which answer(s) **is/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.

49. 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 μ = 10.2 kg and standard deviation σ = 0.8 kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the *z*-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

50. 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 μ = 520 and standard deviation σ = 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?

REFERENCES

5.1 The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

"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).

5.2 Using the Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA

"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).

"Facebook Statistics." Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

SOLUTIONS

1 ounces of water in a bottle

3 2

5 –4

7 –2

9 The mean becomes zero.

11 $z = 2$

13 $z = 2.78$

15 $x = 20$

17 $x = 6.5$

19 $x = 1$

21 $x = 1.97$

23 $z = –1.67$

25 $z \approx –0.33$

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 The lifetime of a Sunshine CD player measured in years.

44 c

46

a. Use the *z*-score formula. $z = –0.5141$. 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. $z = 1.5424$. 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.

48

a. iv

b. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

50 Let X = an SAT math score and Y = an ACT math score.

a. $X = 720$ $\frac{720 - 520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.

b. $z = 1.5$
The math SAT score is 520 + 1.5(115) ≈ 692.5. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.

c. $\frac{X - \mu}{\sigma} = \frac{700 - 514}{117} \approx 1.59$, the z-score for the SAT. $\frac{Y - \mu}{\sigma} = \frac{30 - 21}{5.3} \approx 1.70$, the z-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher z-score).

6 | THE CENTRAL LIMIT THEOREM

Figure 6.1 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 the normal probability density function. (credit: John Lodder)

Introduction

CHAPTER OBJECTIVE

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

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

6.1 | The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

The sampling distribution is a theoretical distribution. It is created by taking many many samples of size n from a population. Each sample mean is then treated like a single observation of this new distribution, the sampling distribution. The genius of thinking this way is that it recognizes that when we sample we are creating an observation and that observation must come from a particular distribution. The Central Limit Theorem answers the question: from what distribution did a sample mean come? If this is discovered, then we can treat a sample mean just like any other observation and calculate probabilities about what values it might take on. We have effectively moved from the world of statistics where we know only what we have from the sample, to the world of probability where we know the distribution from which the sample mean came and the parameters of that distribution.

The reasons that one samples a population are obvious. The time and expense of checking every invoice to determine its validity or every shipment to see if it contains all the items may well exceed the cost of errors in billing or shipping. For some products, sampling would require destroying them, called destructive sampling. One such example is measuring the ability of a metal to withstand saltwater corrosion for parts on ocean going vessels.

Sampling thus raises an important question; just which sample was drawn. Even if the sample were randomly drawn, there are theoretically an almost infinite number of samples. With just 100 items, there are more than 75 million unique samples of size five that can be drawn. If six are in the sample, the number of possible samples increases to just more than one billion. Of the 75 million, then, which one did you get? If there is variation in the items to be sampled, there will be variation in the samples. One could draw an "unlucky" sample and make very wrong conclusions concerning the population. This recognition that any sample we draw is really only one from a distribution of samples provides us with what is probably the single most important theorem is statistics: **the Central Limit Theorem**. Without the Central Limit Theorem it would be impossible to proceed to inferential statistics from simple probability theory. In its most basic form, the Central Limit Theorem states that **regardless** of the underlying probability density function of the population data, the theoretical distribution of the means of samples from the population will be normally distributed. In essence, this says that the mean of a sample should be treated like an observation drawn from a normal distribution. The Central Limit Theorem only holds if the sample size is "large enough" which has been shown to be only 30 observations or more.

Figure 6.2 graphically displays this very important proposition.

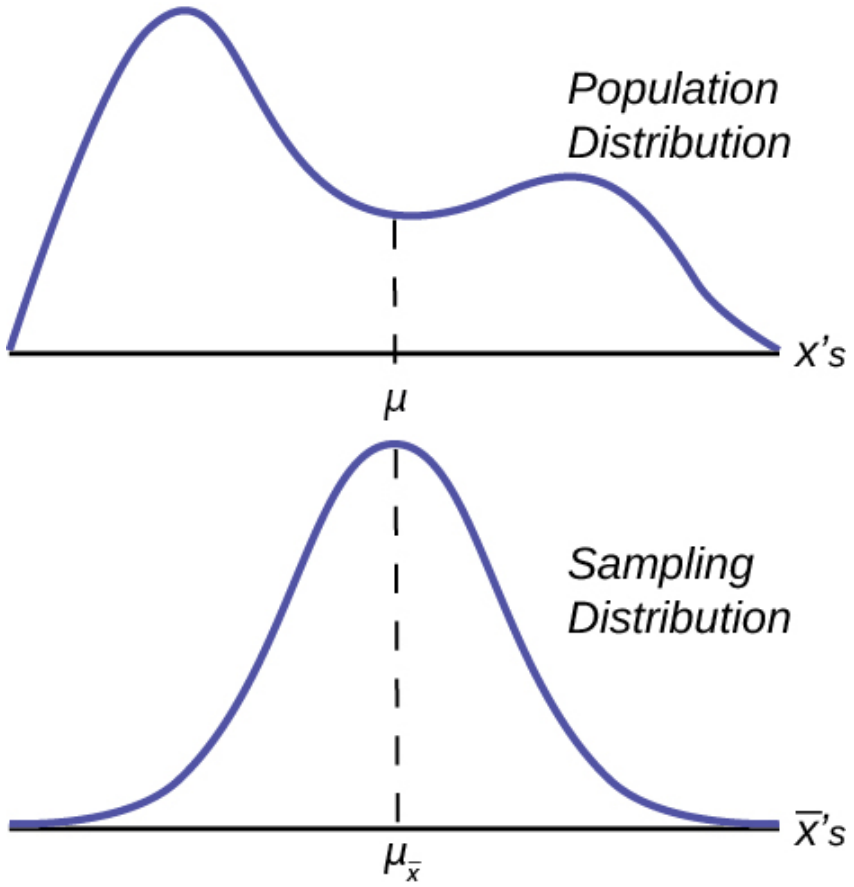


Figure 6.2

Notice that the horizontal axis in the top panel is labeled X. These are the individual observations of the population. This is the **unknown** distribution of the population values. The graph is purposefully drawn all squiggly to show that it does not matter just how odd ball it really is. Remember, we will never know what this distribution looks like, or its mean or standard deviation for that matter.

The horizontal axis in bottom panel is labeled $\bar{X}$'s. This is the theoretical distribution called sampling distribution of the means. Each observation on this distribution is a sample mean. All these sample means were calculated from individual samples with the same sample size. The theoretical sampling distribution contains all of the sample mean values from all the possible samples that could have been taken from the population. Of course, no one would ever actually take all of these samples, but if they did this is how they would look. And the Central Limit Theorem says that they will be normally distributed.

The Central Limit Theorem goes even further and tells us the mean and standard deviation of this theoretical distribution.

Parameter	Population Distribution	Sample	Sampling Distribution of $\bar{X}$'s
Mean	μ	$\bar{X}$	$\mu_{\bar{x}}$ and $E(\mu_{\bar{x}}) = \mu$
Standard Deviation	σ	s	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Table 6.1

The practical significance of The Central Limit Theorem is that now we can compute probabilities for drawing a sample mean, $\bar{X}$, in just the same way as we did for drawing specific observations, X's when we knew the population mean and standard deviation and that the population data were normally distributed.. The standardizing formula has to be amended to recognize that the mean and standard deviation of the sampling distribution, sometimes, called the standard error of the mean, are different from those of the population distribution, but otherwise nothing has changed. The new standardizing formula is

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Notice that $\mu_{\bar{X}}$ in the first formula has been changed to simply μ in the second version. The reason is that mathematically it can be shown that the expected value of $\mu_{\bar{X}}$ is equal to μ. This was stated in **Table 6.1** above. Mathematically, the E(x) symbol indicates the "expected value of x". This formula will be used in the next unit to provide estimates of the **unknown** population parameter μ.

6.2 | Using the Central Limit Theorem -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger size from any population, then the mean of the sampling distribution, $\mu_{\bar{x}}$ tends to get closer and closer to the true population mean, μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation of the sampling distribution gets. (Remember that the standard deviation for the sampling distribution of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ must be closer to the population mean μ as n increases. We can say that μ is the value that the sample means approach as n gets larger. The Central Limit Theorem illustrates the law of large numbers.

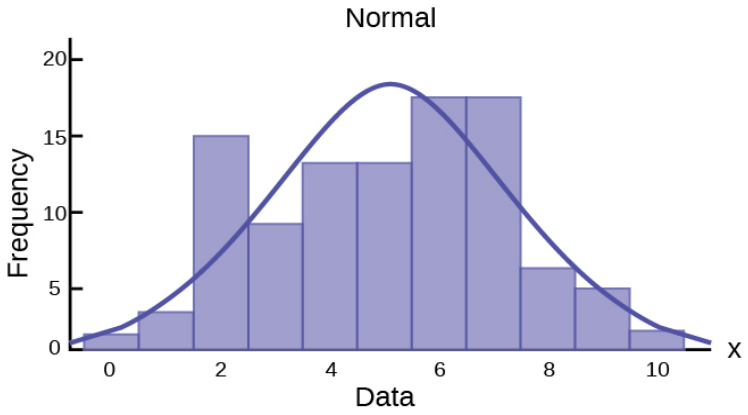
This concept is so important and plays such a critical role in what follows it deserves to be developed further. Indeed, there are two critical issues that flow from the Central Limit Theorem and the application of the Law of Large numbers to it. These are

1. The probability curve of the sampling distribution of means is normally distributed **regardless** of the underlying distribution of the population observations and
2. standard deviation of the sampling distribution decreases as the size of the samples that were used to calculate the means for the sampling distribution increases.

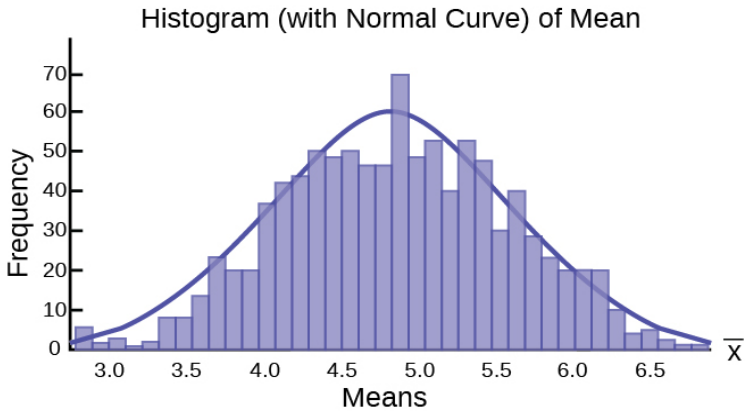
Taking these in order. It would seem counterintuitive that the population may have **any** distribution and the distribution of means coming from it would be normally distributed. With the use of computers, experiments can be simulated that show the process by which the sampling distribution changes as the sample size is increased. These simulations show visually the results of the mathematical proof of the Central Limit Theorem.

Here are three examples of very different population distributions and the evolution of the sampling distribution to a normal distribution as the sample size increases. The top panel in these cases represents the histogram for the original data. The three panels show the histograms for 1,000 randomly drawn samples for different sample sizes: n=10, n= 25 and n=50. As the sample size increases, and the number of samples taken remains constant, the distribution of the 1,000 sample means becomes closer to the smooth line that represents the normal distribution.

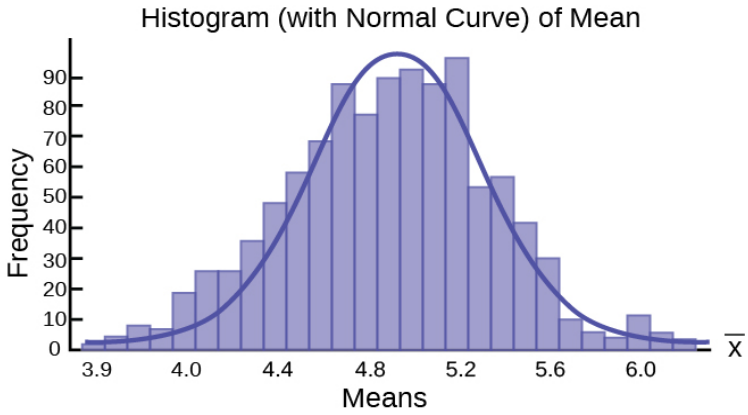
Figure 6.3 is for a normal distribution of individual observations and we would expect the sampling distribution to converge on the normal quickly. The results show this and show that even at a very small sample size the distribution is close to the normal distribution.



Sample Size $n = 10$



Sample Size $n = 25$



Sample Size $n = 50$

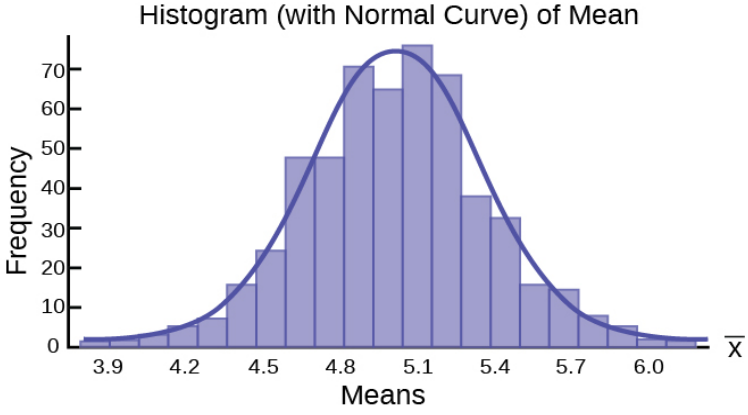
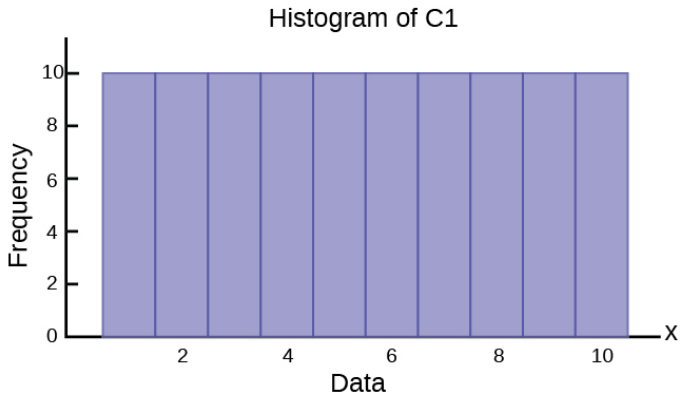


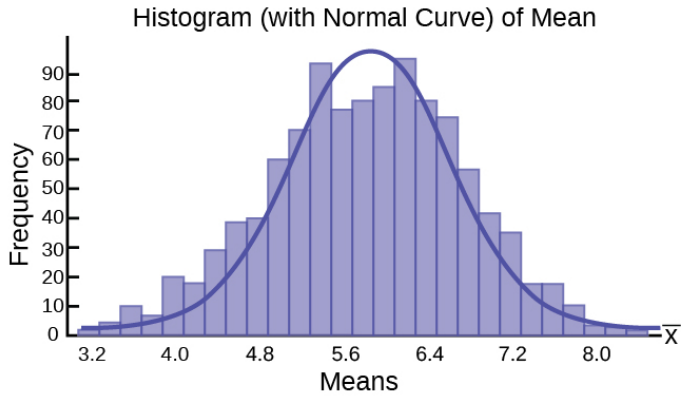
Figure 6.3

Figure 6.4 is a uniform distribution which, a bit amazingly, quickly approached the normal distribution even with only a sample of 10.

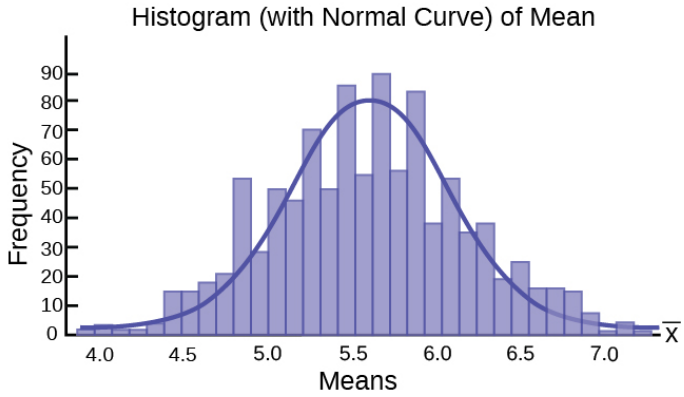
Distribution of Random Variable



Sample Size $n = 10$



Sample Size $n = 25$



Sample Size $n = 50$

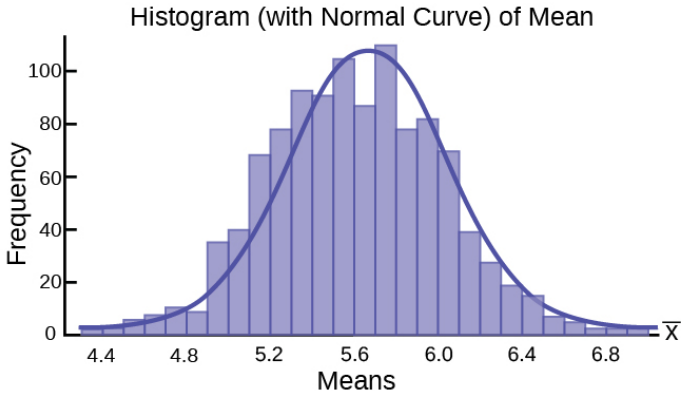
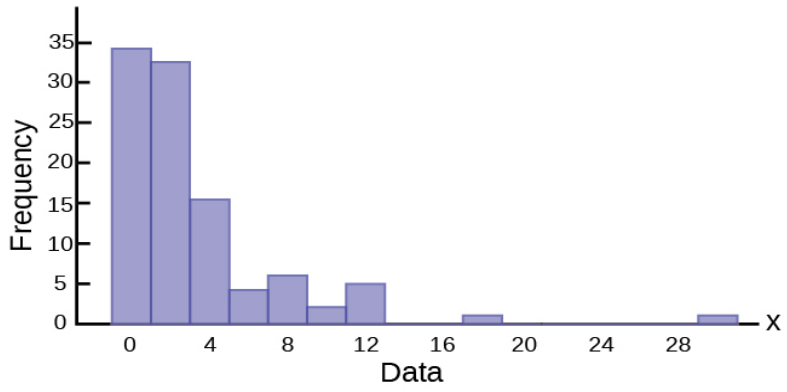
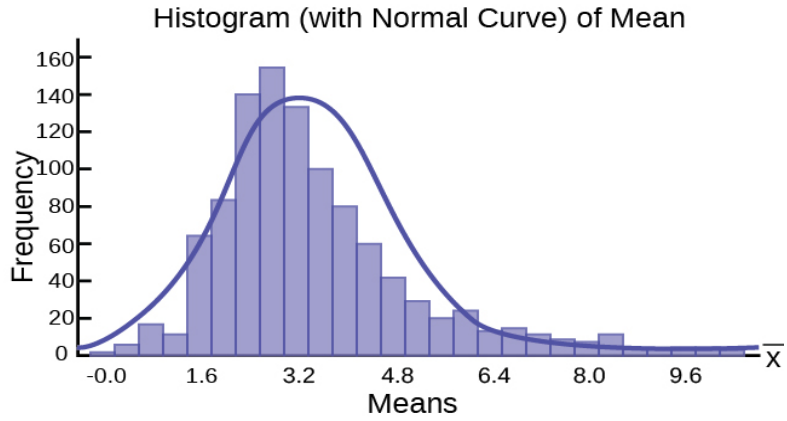


Figure 6.4

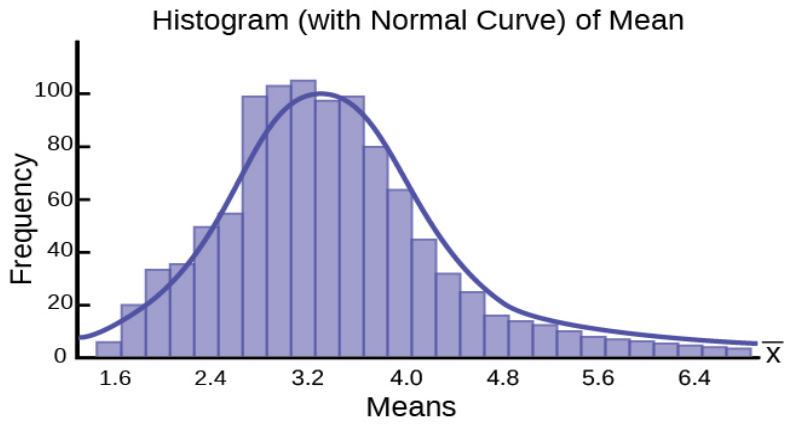
Figure 6.5 is a skewed distribution. This last one could be an exponential, geometric, or binomial with a small probability of success creating the skew in the distribution. For skewed distributions our intuition would say that this will take larger sample sizes to move to a normal distribution and indeed that is what we observe from the simulation. Nevertheless, at a sample size of 50, not considered a very large sample, the distribution of sample means has very decidedly gained the shape of the normal distribution.



Distribution of Sample means with $n = 10$



Distribution of Sample means with $n = 25$



Distribution of Sample means with $n = 50$

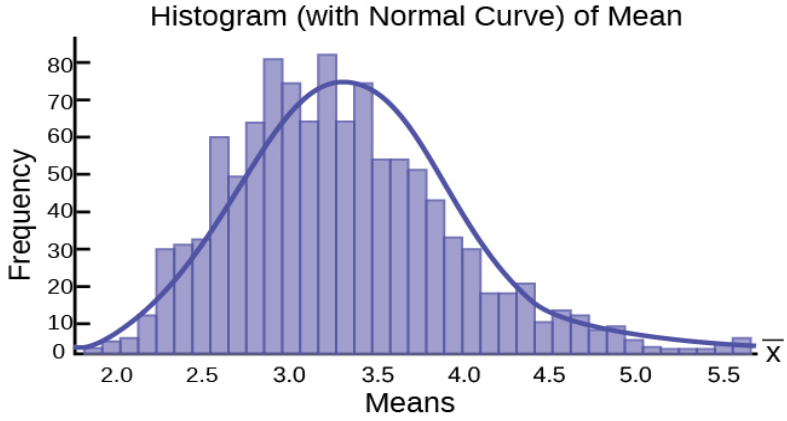


Figure 6.5

The Central Limit Theorem provides more than the proof that the sampling distribution of means is normally distributed. It also provides us with the mean and standard deviation of this distribution. Further, as discussed above, the expected value of the mean, $\mu_{\bar{x}}$, is equal to the mean of the population of the original data which is what we are interested in estimating from the sample we took. We have already inserted this conclusion of the Central Limit Theorem into the formula we use for standardizing from the sampling distribution to the standard normal distribution. And finally, the Central Limit Theorem has also provided the standard deviation of the sampling distribution, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and this is critical to have to calculate probabilities of values of the new random variable, $\bar{x}$.

Figure 6.6 shows a sampling distribution. The mean has been marked on the horizontal axis of the $\bar{x}$'s and the standard deviation has been written to the right above the distribution. Notice that the standard deviation of the sampling distribution is the original standard deviation of the population, divided by the sample size. We have already seen that as the sample size increases the sampling distribution becomes closer and closer to the normal distribution. As this happens, the standard deviation of the sampling distribution changes in another way; the standard deviation decreases as n increases. At very very large n, the standard deviation of the sampling distribution becomes very small and at infinity it collapses on top of the population mean. This is what it means that the expected value of $\mu_{\bar{x}}$ is the population mean, μ.

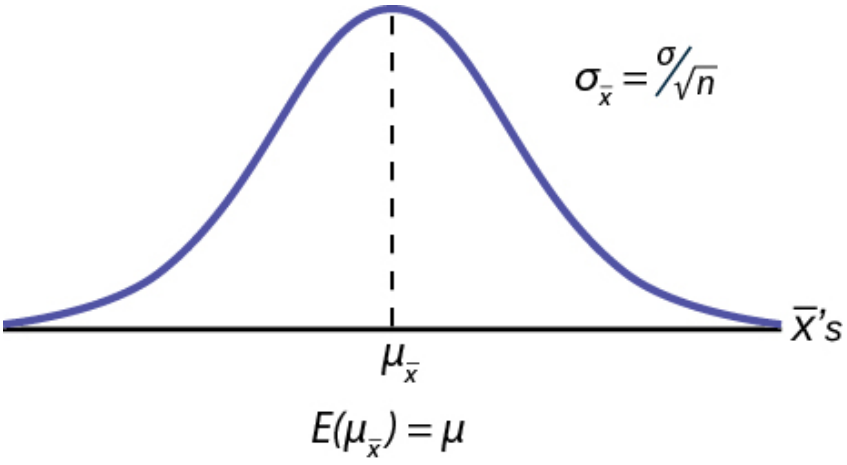


Figure 6.6

At non-extreme values of n,this relationship between the standard deviation of the sampling distribution and the sample size plays a very important part in our ability to estimate the parameters we are interested in.

Figure 6.7 shows three sampling distributions. The only change that was made is the sample size that was used to get the sample means for each distribution. As the sample size increases, n goes from 10 to 30 to 50, the standard deviations of the respective sampling distributions decrease because the sample size is in the denominator of the standard deviations of the sampling distributions.

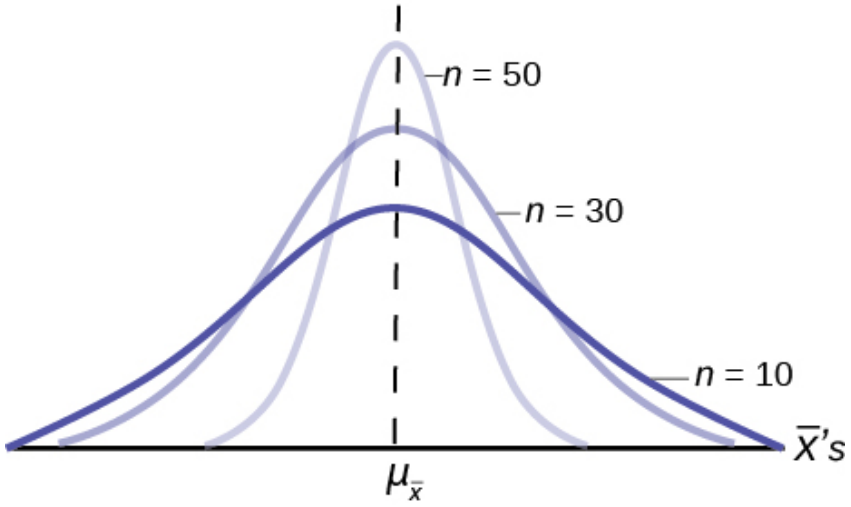


Figure 6.7

The implications for this are very important. **Figure 6.8** shows the effect of the sample size on the confidence we will have in our estimates. These are two sampling distributions from the same population. One sampling distribution was created with samples of size 10 and the other with samples of size 50. All other things constant, the sampling distribution with sample size 50 has a smaller standard deviation that causes the graph to be higher and narrower. The important effect of this is that for the same probability of one standard deviation from the mean, this distribution covers much less of a range of possible values than the other distribution. One standard deviation is marked on the $\bar{X}$ axis for each distribution. This is shown by the two arrows that are plus or minus one standard deviation for each distribution. If the probability that the true mean is one standard deviation away from the mean, then for the sampling distribution with the smaller sample size, the possible range of values is much greater. A simple question is, would you rather have a sample mean from the narrow, tight distribution, or the flat, wide distribution as the estimate of the population mean? Your answer tells us why people intuitively will always choose data from a large sample rather than a small sample. The sampling mean they are getting is coming from a more compact distribution. This concept will be the foundation for what will be called level of confidence in the next unit.

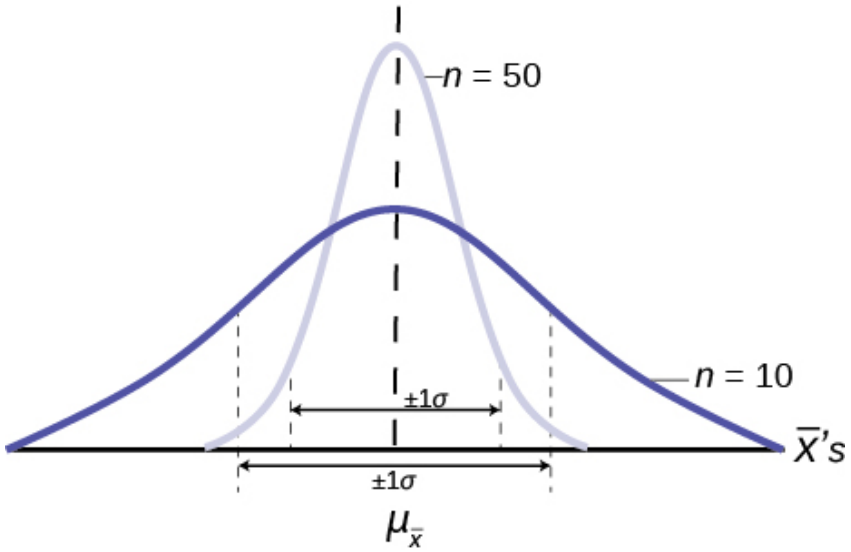


Figure 6.8

6.3 | Central Limit Theorem (Pocket Change) -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

Stats Lab

6.1 Central Limit Theorem (Pocket Change)

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate and compare properties of the central limit theorem.

NOTE

This lab works best when sampling from several classes and combining data.

Collect the Data

1. Count the change in your pocket. (Do not include bills.)
2. Randomly survey 30 classmates. Record the values of the change in **Table 6.2**.

Table 6.2

3. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

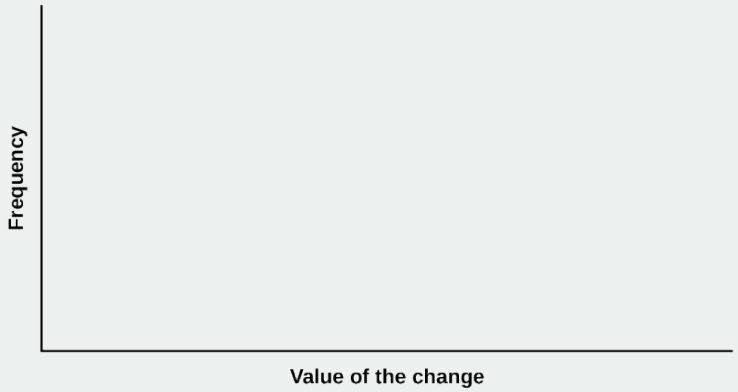


Figure 6.9

4. Calculate the following (n = 1; surveying one person at a time):
   a. $\bar{x}$ = _____
   b. s = _____

5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Collecting Averages of Pairs

Repeat steps one through five of the section **Collect the Data.** with one exception. Instead of recording the change of 30 classmates, record the average change of 30 pairs.

1. Randomly survey 30 **pairs** of classmates.
2. Record the values of the average of their change in **Table 6.3**.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 6.3

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled **Collect the Data**. Sketch the graph using a ruler and a pencil.

Figure 6.10

4. Calculate the following (n = 2; surveying two people at a time):
   a. $\bar{x}$ = _____
   b. s = _____
5. Draw a smooth curve through tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Collecting Averages of Groups of Five

Repeat steps one through five (of the section titled **Collect the Data**) with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of five.

1. Randomly survey 30 **groups of five** classmates.
2. Record the values of the average of their change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 6.4

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled **Collect the Data**. Sketch the graph using a ruler and a pencil.

Figure 6.11

4. Calculate the following (n = 5; surveying five people at a time):
   a. $\bar{x}$ = _____
   b. s = _____
5. Draw a smooth curve through tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Why did the shape of the distribution of the data change, as n changed? Use one to two complete sentences to explain what happened.
2. In the section titled **Collect the Data**, what was the approximate distribution of the data? X ~ _____(_____,_____)
3. In the section titled **Collecting Averages of Groups of Five**, what was the approximate distribution of the averages? $\bar{X}$ ~ _____(_____,_____)
4. In one to two complete sentences, explain any differences in your answers to the previous two questions.

KEY TERMS

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

Central Limit Theorem Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\bar{X}$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Finite Population Correction Factor adjusts the variance of the sampling distribution if the population is known and more than 5% of the population is being sampled.

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

Normal Distribution a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation.; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the random variable, Z, is called the **standard normal distribution**.

Sampling Distribution Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

Standard Error of the Mean the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

Standard Error of the Proportion the standard deviation of the sampling distribution of proportions

CHAPTER REVIEW

6.1 The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

6.2 Using the Central Limit Theorem -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

The Central Limit Theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean $\bar{x}$ gets to μ.

FORMULA REVIEW

6.1 The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

The Central Limit Theorem for Sample Means:

$$\bar{X} \sim N\left(\mu_{\bar{x}}, \frac{\sigma}{\sqrt{n}}\right)$$

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

The Mean $\bar{X}$: $\mu_{\bar{x}}$

Central Limit Theorem for Sample Means z-score $z = \frac{\bar{x} - \mu_{\bar{x}}}{\left(\frac{\sigma}{\sqrt{n}}\right)}$

Standard Error of the Mean (Standard Deviation ($\bar{X}$)): $\frac{\sigma}{\sqrt{n}}$

Finite Population Correction Factor for the sampling distribution of means: $Z = \frac{\bar{x} - \mu}{\frac{\theta}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}}$

Finite Population Correction Factor for the sampling distribution of proportions: $\theta_{p'} = \sqrt{\frac{p(1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}}$

PRACTICE

6.2 Using the Central Limit Theorem -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

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 taken.

1.

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?
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.

2. Draw the graph from **Exercise 6.1**

3. Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

4. Draw the graph from **Exercise 6.3**

5. Find the 90th percentile for the mean weight for the 100 weights.

6. Draw the graph from **Exercise 6.5**

7.

a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?
b. Find $P(\Sigma x < 2{,}450)$.

8. Draw the graph from **Exercise 6.7**

9. Find the 90th percentile for the total weight of the 100 weights.

10. Draw the graph from **Exercise 6.9**

Use the following information to answer the next five 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.

11.

a. What is the standard deviation?
b. What is the parameter *m*?

12. What is the distribution for the length of time one battery lasts?

13. What is the distribution for the mean length of time 64 batteries last?

14. What is the distribution for the total length of time 64 batteries last?

15. Find the probability that the sample mean is between seven and 11.

16. Find the 80^{th} percentile for the total length of time 64 batteries last.

17. Find the *IQR* for the mean amount of time 64 batteries last.

18. Find the middle 80% for the total amount of time 64 batteries last.

Use the following information to answer the next eight exercises: A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is taken.

19. Find $P(\Sigma x > 420)$.

20. Find the 90^{th} percentile for the sums.

21. Find the 15^{th} percentile for the sums.

22. Find the first quartile for the sums.

23. Find the third quartile for the sums.

24. Find the 80^{th} percentile for the sums.

HOMEWORK

6.1 The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

25. 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, $\bar{X} =$ _____
d. $\bar{X} \sim$ _____ (_____, _____)
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.

26. 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. If $\bar{X}$ = average distance in feet for 49 fly balls, then $\bar{X} \sim$ _____(_____,_____)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\bar{X}$. Shade the region corresponding to the probability. Find the probability.
c. Find the 80^{th} percentile of the distribution of the average of 49 fly balls.

27. 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, $\bar{X}$ = _____
c. $\bar{X}$ ~ _____(_____,_____)
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.

28. 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. Consider 49 of the races. Let $\bar{X}$ the average of the 49 races.

a. $\bar{X}$ ~ _____(_____,_____)
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80^{th} percentile for the average of these 49 marathons.
d. Find the median of the average running times.

29. 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, $\bar{X}$ = _____
d. $\bar{X}$ ~ _____(_____,_____)
e. Find the first quartile for the average song length.
f. The IQR(interquartile range) for the average song length is from _____–_____.

30. In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = _____
b. In words, $\bar{X}$ = _____
c. $\bar{X}$ ~ _____(_____,_____)
d. The IQR for $\bar{X}$ is from _____ acres to _____ acres.

31. 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, the mean of $\bar{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\bar{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\bar{X}$ is approximately the same as the standard deviation of X.

32. 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 about ten. Suppose that 16 individuals are randomly chosen. Let $\bar{X}$ = average percent of fat calories.

a. $\bar{X}$ ~ _____(_____, _____)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

33. 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, $\bar{X} =$ _____
c. $\bar{X} \sim$ _____(_____,_____)
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?

34. Which of the following is NOT TRUE about the distribution for averages?

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.

35. 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. $\bar{X} \sim N(4.59, 0.10)$
b. $\bar{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
c. $\bar{X} \sim N\left(4.59, \frac{16}{0.10}\right)$
d. $\bar{X} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

REFERENCES

6.1 The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA

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.

SOLUTIONS

1

a. $U(24, 26)$, 25, 0.5774
b. $N(25, 0.0577)$
c. 0.0416

3 0.0003

5 25.07

7

a. $N(2{,}500, 5.7735)$

b. 0

9 2,507.40

11

a. 10

b. $\frac{1}{10}$

13 $N\left(10, \frac{10}{8}\right)$

15 0.7799

17 1.69

19 0.0072

21 391.54

23 405.51

25

a. X = amount of change students carry

b. $X \sim E(0.88, 0.88)$

c. $\bar{X}$ = average amount of change carried by a sample of 25 students.

d. $\bar{X} \sim N(0.88, 0.176)$

e. 0.0819

f. 0.1882

g. The distributions are different. Part a is exponential and part b is normal.

27

a. length of time for an individual to complete IRS form 1040, in hours.

b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.

c. $N\left(10.53, \frac{1}{3}\right)$

d. Yes. I would be surprised, because the probability is almost 0.

e. No. I would not be totally surprised because the probability is 0.2312

29

a. the length of a song, in minutes, in the collection

b. $U(2, 3.5)$

c. the average length, in minutes, of the songs from a sample of five albums from the collection

d. $N(2.75, 0.0220)$

e. 2.74 minutes

f. 0.03 minutes

31

a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.

b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.

c. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.

33

a. X = the yearly income of someone in a third world country

b. the average salary from samples of 1,000 residents of a third world country

c. $\bar{X} \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$

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.
$P(2000 < \bar{X} < 2100) = 0.1537$
$P(2100 < \bar{X} < 2200) = 0.1317$

35 b

7 | CONFIDENCE INTERVALS

Figure 7.1 Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Introduction

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.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals. What statistics provides us beyond a simple average, or point estimate, is an estimate to which we can attach a probability of accuracy, what we will call a confidence level. We make inferences with a known level of probability.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

$\bar{x}$ and s are each called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include the unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1 .$$

The **empirical rule**, which applies to the normal distribution, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ with 95% probability. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then with 95% probability the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \text{ and } \bar{x} + 0.2 = 2 + 0.2 = 2.2$$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).** Please note that we talked in terms of 95% confidence using the empirical rule. The empirical rule for two standard deviations is only approximately 95% of the probability under the normal distribution. To be precise, two standard deviations under a normal distribution is actually 95.44% of the probability. To calculate the exact 95% confidence level we would use 1.96 standard deviations.

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ, or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form:

(Point estimate - margin of error, Point estimate + margin of error)

For the confidence interval for a mean the formula would be:

$$\mu = \bar{X} \pm Z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Or written another way as:

$$\bar{X} - Z_{\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Where $\bar{X}$ is the sample mean. Z_{α} is determined by the level of confidence and $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of the sampling distribution for means.

7.1 | A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, **where the population standard deviation is known**, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (*EBM*) is called the **error bound for a population mean** (abbreviated ***EBM***). The sample mean $\bar{x}$ is the **point estimate** of the unknown population mean μ.

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or, in symbols,($\bar{x} - EBM$, $\bar{x} + EBM$)

The margin of error (*EBM*) depends on the **confidence level** (abbreviated ***CL***). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, *CL*. α is the probability that the interval does not contain the unknown population parameter.
Mathematically, $\alpha + CL = 1$.

Example 7.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.

The sample mean is seven, and the error bound for the mean is 2.5.

$\bar{x}$ = 7 and EBM = 2.5

The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).

If the confidence level (CL) is 95%, then we say that, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Try It

7.1 Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

What is the confidence interval estimate for the population mean?

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x}$ = 10, and we have constructed the 90% confidence interval (5, 15) where EBM = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.

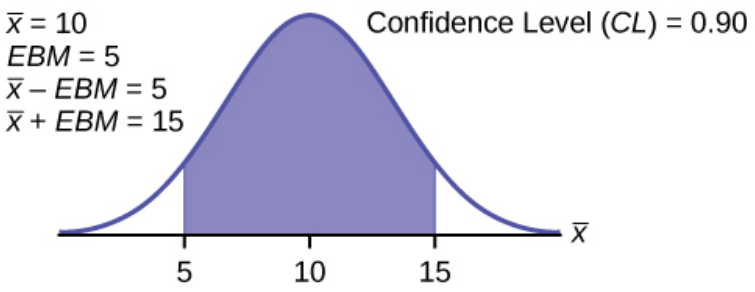


Figure 7.2

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the "standard error of the mean" in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the central limit theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the *z*-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a standard normal probability table.

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

`invNorm`(0.975, 0, 1) = 1.96

NOTE

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound (*EBM*)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

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

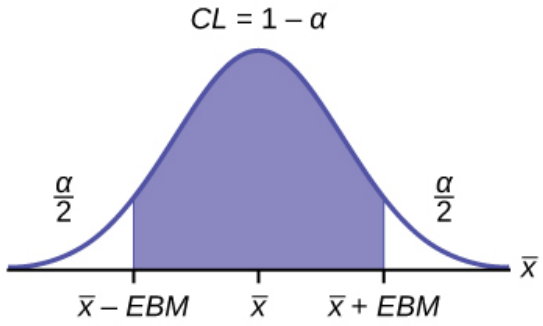


Figure 7.3

Writing the Interpretation

The interpretation should clearly state the confidence level (*CL*), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 7.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution 7.2

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

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

$\bar{x} = 68$

$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

$\sigma = 3$; $n = 36$; The confidence level is 90% ($CL = 0.90$)

$CL = 0.90$ so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

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

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

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

using invNorm(0.95, 0, 1) on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

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

$\bar{x}$ - *EBM* = 68 - 0.8225 = 67.1775

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

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

Solution 7.2

Solution B

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

Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.
Arrow to Stats and press ENTER.
Arrow down and enter three for σ, 68 for $\bar{x}$, 36 for n, and .90 for C-level.
Arrow down to Calculate and press ENTER.
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Try It

7.2 Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time.

Example 7.3

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 7.1** 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
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Table 7.1

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 7.3

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean.

$\bar{x} = 1.024$

Next, find the *EBM*. Because you are creating a 98% confidence interval, $CL = 0.98$.

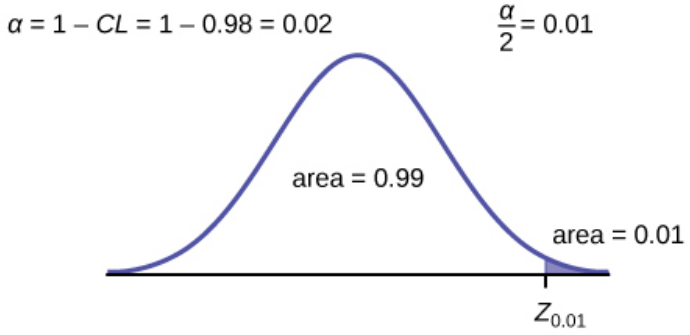


Figure 7.4

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$.

$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$

$\bar{x} - EBM = 1.024 - 0.1431 = 1.1671$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution 7.3

Solution B

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

Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.
Arrow to Stats and press ENTER.
Arrow down and enter the following values:
σ: 0.337
$\bar{x}$: 1.024
n: 30
C-level: 0.98
Arrow down to Calculate and press ENTER.
The confidence interval is (to three decimal places) (0.881, 1.167).

Try It

7.3 **Table 7.2** 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 $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Table 7.2

Notice the difference in the confidence intervals calculated in **Example 7.3** and the following **Try It** exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 7.4

Suppose we change the original problem in **Example 7.2** by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution 7.4

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

$\bar{x} = 68$

$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

$\sigma = 3$; $n = 36$; The confidence level is 95% ($CL = 0.95$).

$CL = 0.95$ so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \quad z_{\frac{\alpha}{2}} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$

$\bar{x} - EBM = 68 - 0.98 = 67.02$

$\bar{x} + EBM = 68 + 0.98 = 68.98$

Notice that the *EBM* is larger for a 95% confidence level in the original problem.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

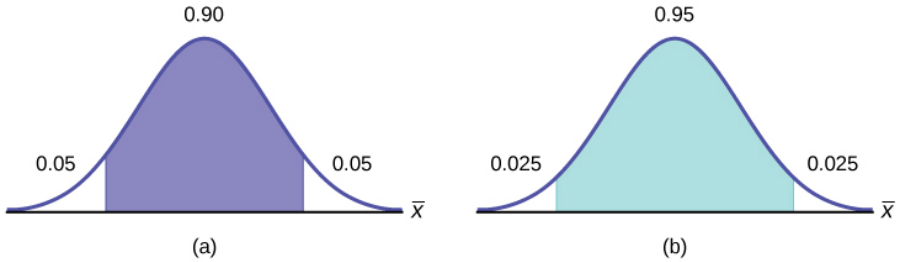


Figure 7.5

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Try It

7.4 Refer back to the pizza-delivery **Try It** exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Example 7.5

Suppose we change the original problem in **Example 7.2** to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution 7.5

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When $n = 100$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution 7.5

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Try It Σ

7.5 Refer back to the pizza-delivery **Try It** exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,

- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 7.6

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $EBM = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$.

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Try It

7.6 Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n.

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 7.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

From the problem, we know that $\sigma = 15$ and $EBM = 2$.
$z = z_{0.025} = 1.96$, because the confidence level is 95%.

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$ using the sample size equation.

Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Try It

7.7 The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

7.2 | A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and only used the Student's t-distribution only for sample sizes of at most 30.

If you draw a simple random sample of size n from a population with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student's t-distribution with $n - 1$ degrees of freedom**.

The t-score has the same interpretation as the ***z*-score**. It measures how far in standard deviation units $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The **degrees of freedom**, $n - 1$, come from the calculation of the sample standard deviation s. Remember when we first calculated a sample standard deviation we divided the sum of the squared deviations by $n - 1$, but we used n deviations $(x - \bar{x}\text{ values})$ to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n - 1$ deviations. The other $n - 1$ deviations can change or vary freely. **We call the number $n - 1$ the degrees of freedom (df)** in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the t-value increases and the confidence interval increases in width.

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution. You can confirm this by reading the bottom line at infinite degrees of freedom for a familiar level of confidence, e.g. at column 0.05, 95% level of confidence, we find the t-value of 1.96 at infinite degrees of freedom.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero, again like the standard normal distribution.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.

- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. This assumption comes from the Central Limit theorem because the individual observations in this case are the $\bar{x}$ s of the sampling distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

A probability table for the Student's t-distribution is used to calculate t-values at various commonly-used levels of confidence. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that at the bottom the table will show the t-value for infinite degrees of freedom. Mathematically, as the degrees of freedom increase, the t distribution approaches the standard normal distribution. You can find familiar Z-values by looking in the relevant alpha column and reading value in the last row.

A Student's t table (See **Appendix A**) gives *t*-scores given the degrees of freedom and the right-tailed probability.

The Student's t distribution has one of the most desirable properties of the normal: it is symmetrical. What the Student's t distribution does is spread out the horizontal axis so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's t distributions, one for each adjustment to the sample size. As the sample size increases, the Student's t distribution become more and more like the normal distribution. When the sample size reaches 30 the normal distribution is usually substituted for the Student's t because they are so much alike. This relationship between the Student's t distribution and the normal distribution is shown in **???**.

Figure (Figure_8_6.jpg)

Figure 7.6

This is another example of one distribution limiting another one, in this case the normal distribution is the limiting distribution of the Student's t when the degrees of freedom in the Student's t approaches infinity. This conclusion comes directly from the derivation of the Student's t distribution by Mr. Goset. He recognized the problem as having few observations and no estimate of the population standard deviation. He was substituting the sample standard deviation and getting volatile results. He therefore created the Student's t as a ratio of the normal distribution and Chi squared distribution. The Chi squared distribution is itself a ratio of two variances, in this case the sample variance and the unknown population variance. The Student's t thus is tied to the normal distribution, but has degrees of freedom that come from those from the Chi squared distribution. The algebraic solution demonstrates this result.

Development of Student's t-distribution:

1. $t = \frac{z}{\sqrt{\frac{\chi^2}{v}}}$

   Where Z is the standard normal distribution and χ^2 is the chi-squared distribution with v degrees of freedom.

2. $t = \frac{\frac{\left(\bar{x} - \mu\right)}{\sigma}}{\sqrt{\frac{\frac{s^2}{(n-1)}}{\frac{\sigma^2}{(n-1)}}}}$

   by substitution, and thus Student's t with $v = n - 1$ degrees of freedom is:

3. $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

Restating the formula for a confidence interval for the mean for cases when the sample size is smaller than 30 and we do not know the population standard deviation, σ:

$$\bar{x} - t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

Here the point estimate of the population standard deviation, s has been substituted for the population standard deviation, σ, and $t_{v,\alpha}$ has been substituted for Z_α. The Greek letter ν (pronounced nu) is placed in the general formula in recognition that there are many Student t_v distributions, one for each sample size. ν is the symbol for the degrees of freedom of the

distribution and depends on the size of the sample. **For this type of problem,** the degrees of freedom is ν = n-1, where n is the sample size. To look up a probability in the Student's t table we have to know the degrees of freedom in the problem.

Example 7.8

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 7.8

- The solution step-by-step.

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

$\bar{x}$ = 8.2267 s = 1.6722 n = 15

df = 15 – 1 = 14 CL so α = 1 – CL = 1 – 0.95 = 0.05

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

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

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

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

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Example 7.9

The average earnings per share (EPS) for 10 industrial stocks randomly selected from those listed on the Dow-Jones was found to be $\bar{X}$ = 1.85 with a standard deviation of s=0.395. Calculate a 99% confidence interval for the average EPS of all the industrials listed on the DJIA.

An alternative way to approach confidence intervals is shown below. In this approach instead of using the EMB as an independent calculation, this approach uses the broader equation as developed above.

$$\bar{x} - t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

Solution 7.9

To help visualize the process of calculating a confident interval we draw the appropriate distribution for the problem. In this case this is the Student’s -t because we do not know the population standard deviation and the sample is small, less than 30.

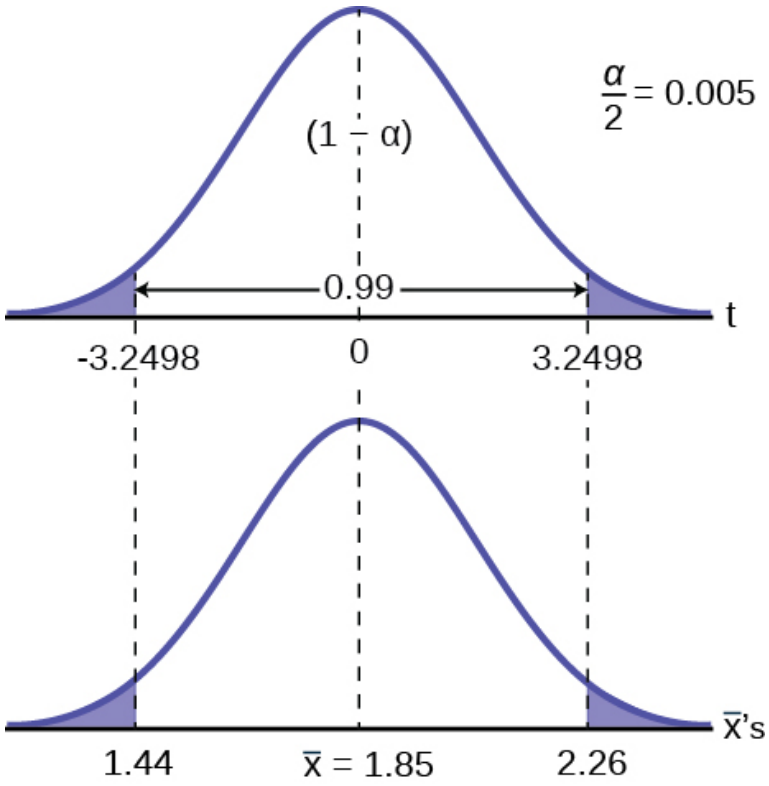


Figure 7.7

To find the appropriate t-value requires two pieces of information, the level of confidence desired and the degrees of freedom. The question asked for a 99% confidence level. On the graph this is shown where (1-α) , the level of confidence , is in the unshaded area. The tails, thus, have .005 probability each, α/2. The degrees of freedom for this type of problem is n-1= 9. From the Student's t table, at the row marked 9 and column marked .005, is the number of standard deviations to capture 99% of the probability, 3.2498. These are then placed on the graph remembering that the Student's t is symmetrical and so the t-value is both plus or minus on each side of the mean.

Inserting these values into the formula gives the result. These values can be placed on the graph to see the relationship between the distribution of the sample means, $\bar{X}$'s and the Student's t distribution.

$$\mu = \bar{X} \pm t_{\alpha/2,df=n-1}\frac{s}{\sqrt{n}} = 1.851 \pm 3.2498\frac{0.395}{\sqrt{10}} = 1.8551 \pm 0.406$$

$$1.445 \leq \mu \leq 2.257$$

We state the formal conclusion as :

With 99% confidence level, the average EPS of all the industries listed at DJIA is from $1.44 to $2.26.

Try It

7.9 You do a study of hypnotherapy to determine how effective it is in increasing the number of hourse of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Example 7.10

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. **Table 7.2** shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Table 7.3

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution 7.10

From the sample, you can calculate $\bar{x}$ = 127.45 and s = 25.965. There are 20 infants in the sample, so n = 20, and df = 20 – 1 = 19.

You are asked to calculate a 90% confidence interval: CL = 0.90, so α = 1 – CL = 1 – 0.90 = 0.10 $\frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is 1 – 0.05 = 0.95.

Use a table, calculator, or computer to find that $t_{0.05}$ = 1.729.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$\bar{x}$ – EBM = 127.45 – 10.038 = 117.412

$\bar{x}$ + EBM = 127.45 + 10.038 = 137.488

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Example 7.11

Determine the level of confidence that must have been used to solve the following interval for the population mean. The limits found were (40, 46), the sample size was 25, and the sample standard deviation was 7.2674.

Solution 7.11

We begin with the formula for a confidence interval:

$$\mu = \left[\bar{x} \pm \left[t\left(\frac{a}{2}\right)\frac{s}{\sqrt{n}}\right]\right]$$

We know that we need the test statistic for a Student's t distribution because the sample size is less than 30. Next we solve for $\bar{x}$, which is one half the distance between the upper bound and the lower bound of the confidence interval.

$$\bar{x} = \frac{40 + 46}{2} = 43$$

The ± value must be 3; 43 − 3 = 30 and 43 + 3 = 46. Therefore:

$\left[t \cdot \left(\frac{7.2674}{\sqrt{25}}\right)\right] = 3$, ***thus t = 2.064***

Solve for *t* and find this value on the Student's t table at the proper degrees of freedom.

row 24 on t-distribution with value 2.046 yields $\frac{a}{2} = 0.025$ **.**

therefore, $a = 0.05$ ***and*** $1 - a = 0.95$ ***or 95%***

Figure 7.8

Try It

7.11 A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in **Table 7.4**. Use this sample data to construct a 90% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Table 7.4

7.3 | A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

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.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

How do you know you are dealing with a proportion problem? First, the underlying **distribution has a binary random variable and therefore is a binomial distribution**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a proportion, take X,

the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

p' = the **estimated proportion** of successes or sample proportion of successes (p' is a **point estimate** for p, the true proportion, and thus q is the probability of a failure in any one trial.)

x = the **number** of successes in the sample

n = the size of the sample

The formula for the confidence interval for a proportion follows the same format as that for an estimate of a mean. Remembering the sampling distribution for the proportion from **Chapter 7 (http://legacy.cnx.org/content/m54530/latest/)** , the standard deviation was found to be:

$$E(p') = E\left(\frac{x}{n}\right) = \left(\frac{1}{n}\right)E(x) = \left(\frac{1}{n}\right)np = p$$

The confidence interval for a proportion, therefore, becomes the point estimate,

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)}\sqrt{\frac{p'(1-p')}{n}}\right]$$

$Z_{\left(\frac{a}{2}\right)}$ is set according to our desired degree of confidence and $\sqrt{\frac{p'(1-p')}{n}}$ is the standard deviation of the sampling distribution.

The **sample proportions p' and q' are estimates of the unknown population proportions p and q**. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

For the normal distribution of proportions, the z-score formula is as follows.

If $P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ then the z-score formula is $z = \frac{p' - p}{\sqrt{\frac{pq}{n}}}$ or $p' - Z_\alpha\sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha\sqrt{\frac{p'q'}{n}}$

This conclusion can be demonstrated through the following analysis. Proportions are based upon the binomial probability distribution. The possible outcomes are binary, either “success” or “failure”. This gives rise to a proportion, meaning the percentage of the outcomes that are “successes”. It was shown that the binomial distribution could be fully understood if we knew only the probability of a success in any one trial, called p. The mean and the standard deviation of the binomial were defined as:

$$\mu = \text{np}$$
$$\sigma = \sqrt{\text{npq}}$$

It was also shown that the binomial could be estimated by the normal distribution if BOTH np AND nq were greater than 5. From the discussion above, it was found that the standardizing formula for the binomial distribution is:

$$Z = \frac{\text{p'} - p}{\sqrt{\left(\frac{pq}{n}\right)}}$$

which is nothing more than a restatement of the general standardizing formula with appropriate substitutions for μ and σ from the binomial. We can use the standard normal distribution, the reason Z is in the equation, because the normal distribution is the limiting distribution of the binomial. This is another example of the Central Limit Theorem. We have already seen that the sampling distribution of means is normally distributed. Recall the extended discussion in **Chapter 7 (http://legacy.cnx.org/content/m54530/latest/)** concerning the sampling distribution of proportions and the conclusions of the Central Limit Theorem.

We can now manipulate this formula in just the same way we did for finding the confidence intervals for a mean, but to find the confidence intervals for the binomial population parameter, p.

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

Where p′ = x/n, the point estimate of p taken from the sample. Notice that p′ has replaced p in the formula. This is because we do not know p, indeed, this is just what we are trying to estimate.

Unfortunately, there is no correction factor for cases where the sample size is small so np′ and nq′ must always be greater than 5 to develop an interval estimate for p.

Example 7.12

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 sampled, 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 7.12

- The solution step-by-step.

Let X = the number of people in the sample who have cell phones. X is binomial. $X \sim B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find p', q'.

$n = 500$

x = the number of successes = 421

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

$p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$q' = 1 - p' = 1 - 0.842 = 0.158$

Since $CL = 0.95$, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $\left(\frac{\alpha}{2}\right) = 0.025$.

Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

This can be found using the Standard Normal probability table in Appendix A.

The confidence interval for the true binomial population proportion is

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

Thus the confidence interval is :$0.810 \leq p \leq 0.874$

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Try It

7.12 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 7.13

The Mixed Martial Arts gym known as R1 MMA has a larger than average proportion of clients who compete in amateur and professional MMA events. A confidence interval for the population proportion of fighters from 150

different gyms is constructed. The lower limit is determined to be 0.08 and the upper limit is determined to be 0.16. Determine the level of confidence used to construct the interval.

Solution 7.13

We begin with the formula for a confidence interval for a proportion because the random variable is binary; either the client competes in MMA events or they don't.

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

Next we find the sample proportion:

$$p' = \frac{0.08 + 0.16}{2} = 0.12$$

The ± that makes up the confidence interval is thus 0.04; 0.12 + 0.04 = 0.16 and 0.12 − 0.04 = 0.08, the boundaries of the confidence interval. Finally, we solve for *Z*.

$$\left[Z \cdot \sqrt{\frac{0.12(1-0.12)}{150}} \right] = 0.04 \text{ , } \textbf{\textit{therefore Z}} = \textbf{1.51}$$

And then look up the probability of 1.51 standard deviations on the standard normal table.

$p(Z = 1.51) = 0.4345$, $p(Z) \cdot 2 = 0.8690$ ***or*** $86.90\ \%$.

Example 7.14

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 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution 7.14

- The solution is step-by-step

$x = 300$ and $n = 500$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then $\alpha = 1 - CL = 1 - 0.90 = 0.10$ $\left(\frac{\alpha}{2}\right) = 0.05$

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

Remember that the area to the right of $Z_{0.05}$ is 0.05 and the area to the left of $Z_{0.05}$ is 0.95. This can also be found using a standard normal probability table. The student's t-table can also be used by entering he table at the 0.05 column and reading at the line for infinite degrees of freedom. The t-distribution is the normal distribution at infinite degrees of freedom. We use the same formula for a confidence interval for a proportion:

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

The confidence interval for the true binomial population proportion is $0.564 \leq p \leq 0.636$

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.

- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Try It

7.14 A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

7.4 | Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA

Means

Usually we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly in both time and product. Simple telephone surveys will cost approximately $30.00 each, for example.

If we go back to our standardizing formula for the sampling distribution for means, we can see that it is possible to solve it for n. If we do this we have $\left(\bar{X} - \mu\right)$ in the denominator.

$$n = \frac{Z_\alpha^2 \sigma^2}{\left(\bar{X} - \mu\right)^2} = \frac{Z_\alpha^2 \sigma^2}{e^2}$$

Because we have not taken a sample yet we do not know any of the variables in the formula except that we can set Z_α to the level of confidence we desire just as we did when determining confidence intervals. If we set a predetermined acceptable error, or tolerance, for the difference between $\bar{X}$ and μ, called e in the formula, we are much further in solving for the sample size n. We still do not know the population standard deviation, σ. In practice, a pre-survey is usually done which allows for fine tuning the questionnaire and will give a sample standard deviation that can be used. In other cases, previous information from other surveys may be used for σ in the formula. While crude, this method of determining the sample size may help in reducing cost significantly. It will be the actual data gathered that determines the inferences about the population, so caution in the sample size is appropriate calling for high levels of confidence and small sampling errors.

Proportions

What was done in cases when looking for the mean of a distribution can also be done when sampling to determine the population parameter p for proportions. Manipulation of the standardizing formula for proportions gives:

$$n = \frac{Z_\alpha^2 \text{pq}}{e^2}$$

where e = (p′-p), and is the acceptable sampling error, or tolerance, for the application.

In this case the very object of our search is in the formula, p, and of course q because q =1-p. This result occurs because the binomial distribution is a one parameter distribution. If we know p then we know the mean and the standard deviation. Therefore, p shows up in the standard deviation of the sampling distribution which is where we got this formula. If, in an abundance of caution, we substitute 0.5 for p we will draw the largest required sample size that will provide the level of confidence specified by Zα. This is true because of all combinations of two numbers that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p, this is the common practice. This may result in oversampling, but certainly not under sampling, thus, this is a cautious approach.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. **Table 7.5** shows the appropriate sample size at different levels of confidence and different level of the acceptable error, or tolerance.

Required Sample Size (90%)	Required Sample Size (95%)	Tolerance Level
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

Table 7.5

This table is designed to show the maximum sample size required at different levels of confidence given an assumed p= 0.5 and q=0.5 as discussed above.

The acceptable error, called tolerance in the table, is measured in plus or minus values from the actual proportion. For example, an acceptable error of 5% means that if the sample proportion was found to be 26 percent, the conclusion would be that the actual population proportion is between 21 and 31 percent with a 90 percent level of confidence if a sample of 271 had been taken. Likewise, if the acceptable error was set at 2%, then the population proportion would be between 24 and 28 percent with a 90 percent level of confidence, but would require that the sample size be increased from 271 to 1,691. If we wished a higher level of confidence, we would require a larger sample size. Moving from a 90 percent level of confidence of a 95 percent level at a plus or minus 5% tolerance requires changing the sample size from 271 to 384. A very common sample size often seen reported in political surveys is 384. With the survey results it is frequently stated that the results are good to a plus or minus 5% level of “accuracy”.

Example 7.15

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 7.15

From the problem, we know that the acceptable error, *e*, is **0.03** (3%=0.03) and $z_{\frac{\alpha}{2}}$ $z_{0.05}$ = 1.645 because the confidence level is 90%. The acceptable error, *e*, is the difference between the actual proportion *p*, and the sample proportion we expect to get from the sample.

However, in order to find *n*, we need to know the estimated (sample) proportion *p*′. Remember that *q*′ = 1 – *p*′. But, we do not know *p*′ yet. Since we multiply *p*′ and *q*′ together, we make them both equal to 0.5 because *p*′*q*′ = (0.5)(0.5) = 0.25 results in the largest possible product. (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16 and so on). The largest possible product gives us the largest *n*. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size *n*, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{e^2} \text{ gives } n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Try It

7.15 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?

7.5 | Confidence Interval (Home Costs) -- Confidence Intervals -- MtRoyal - Version2016RevA

Stats Lab

7.1 Confidence Interval (Home Costs)

Class Time:

Names:

Student Learning Outcomes

- The student will calculate the 90% confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

NOTE

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

1. Complete the table:

Table 7.6

Describe the Data

1. Compute the following:
   a. $\bar{x}$ = _____
   b. s_x = _____
   c. n = _____
2. In words, define the random variable $\bar{X}$.
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.
   a. Confidence Interval: _____

b. Error Bound: _____

2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.

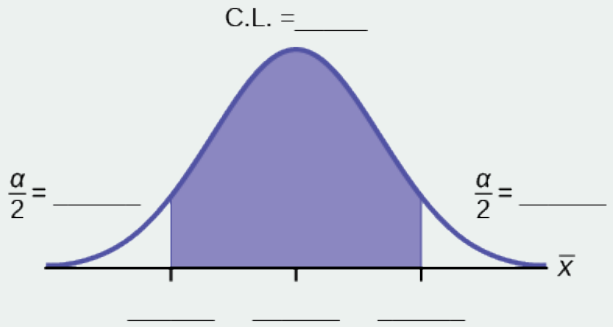


Figure 7.9

5. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the given information, construct a confidence interval for each confidence level given.

Confidence level	EBM/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

Table 7.7

2. What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

KEY TERMS

Binomial Distribution a discrete random variable (RV) which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(\mathbf{n},\mathbf{p})$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

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

Error Bound for a Population Mean (*EBM*) the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion (EBP) the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Inferential Statistics also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Normal Distribution a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation, notation: $X \sim N(\mu,\sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

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.
- 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 depends upon the application for which the t is being used.

CHAPTER REVIEW

7.1 A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

7.2 A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, *s*, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when *s* is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The *t*-score follows the Student's t-distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $\bar{x} \pm \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}}$ where $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$, *s* is the sample standard deviation, and *n* is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

7.3 A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$p' - Z_\alpha\sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha\sqrt{\frac{p'q'}{n}}$$

7.4 Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = Z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{Z_{\alpha}^2\sigma^2}{(\bar{x} - \mu)^2}$$

FORMULA REVIEW

7.1 A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

$\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by
(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$= (\bar{x} - EBM, \bar{x} + EBM)$

$= \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$

$EBM = z\frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z\frac{\alpha}{2}$ = the *z*-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the *z*-score used in the calculation of *"EBM* where $\alpha = 1 - CL$.

$n = \frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (*n*) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate−error bound, point estimate + error bound)

To find the error bound when you know the confidence interval

error bound = upper value−point estimate OR error bound

$= \frac{\text{upper value} - \text{lower value}}{2}$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z\frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

7.2 A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

s = the standard deviation of sample values.

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

$df = n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T, has a Student's t-distribution with df degrees of freedom

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by: $\bar{x} - t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right)$

7.3 A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

$p' = \frac{x}{n}$ where x represents the number of successes in a sample and n represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$q' = 1 - p'$

$p' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ The variable p' has a binomial distribution that can be approximated with the normal distribution shown here. The confidence interval for the true population proportion is given by the formula:

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

$n = \frac{Z_{\frac{\alpha}{2}}^2 p' q'}{e^2}$ provides the number of observations needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error e. Where e = the acceptable difference between the actual population proportion and the sample proportion.

7.4 Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by $\bar{X} - Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$

This formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of *"EBM* where $\alpha = 1 - CL$.

$n = \frac{Z^2 \sigma^2}{(\bar{x} - \mu)^2}$ = the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence for a continuous random variable

PRACTICE

7.1 A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

1. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

2. In words, define the random variables X and $\bar{X}$.

3. Which distribution should you use for this problem?

4. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

5. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

6. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

7. In words, define the random variables X and $\bar{X}$.

8. Which distribution should you use for this problem?

9. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

10. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

11. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

12. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

13. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

14. In words, define the random variable X.

15. In words, define the random variable $\bar{X}$.

16. Which distribution should you use for this problem?

17. Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

18. Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

19. In complete sentences, explain why the confidence interval in **Exercise 7.17** is larger than in **Exercise 7.18**.

20. In complete sentences, give an interpretation of what the interval in **Exercise 7.18** means.

21. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

22. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

23. $\bar{x}$ = _____

24. n = _____

25. _____ = 15

26. In words, define the random variable $\bar{X}$.

27. What is $\bar{x}$ estimating?

28. Is σ_x known?

29. As a result of your answer to **Exercise 7.26**, state the exact distribution to use when calculating the confidence interval.

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

30. How much area is in both tails (combined)? α =_____

31. How much area is in each tail? $\frac{\alpha}{2}$ =_____

32. Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

33. The 95% confidence interval is:_____.

34. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

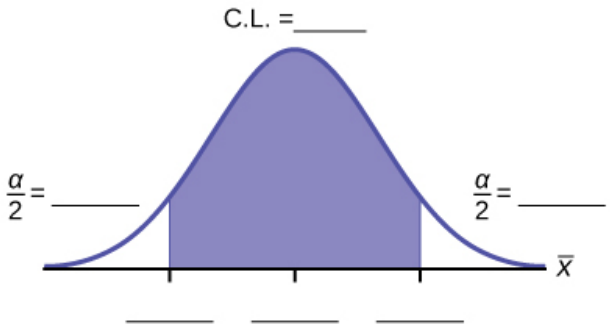


Figure 7.10

35. In one complete sentence, explain what the interval means.

36. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

37. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

7.2 A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

38. Identify the following:

a. $\bar{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

39. Define the random variables X and $\bar{X}$ in words.

40. Which distribution should you use for this problem?

41. Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

42. Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

43. Identify the following:

a. $\bar{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

44. Define the random variable X in words.

45. Define the random variable $\bar{X}$ in words.

46. Which distribution should you use for this problem?

47. Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval, (b) sketch the graph, and (c) calculate the error bound.

48. Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in **Table 7.8** 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

Table 7.8

49. Calculate the following:

a. $\bar{x}$ =_____
b. s_x =_____
c. n =_____

50. Define the random variable $\bar{X}$ in words.

51. What is $\bar{x}$ estimating?

52. Is σ_x known?

53. As a result of your answer to **Exercise 7.52**, state the exact distribution to use when calculating the confidence interval.

Construct a 95% confidence interval for the true mean number of colors on national flags.

54. How much area is in both tails (combined)?

55. How much area is in each tail?

56. Calculate the following:

a. lower limit
b. upper limit
c. error bound

57. The 95% confidence interval is_____.

58. Fill in the blanks on the graph with the areas, the upper and lower limits of the Confidence Interval and the sample mean.

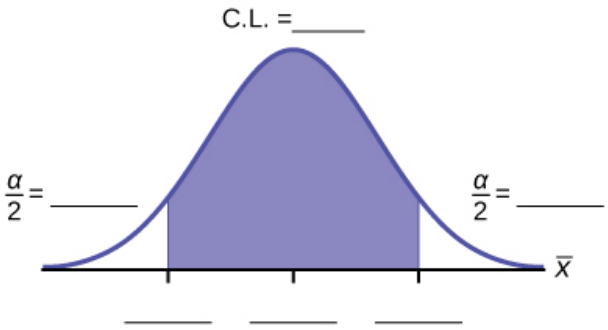


Figure 7.11

59. In one complete sentence, explain what the interval means.

60. Using the same $\bar{x}$, s_x, and level of confidence, suppose that *n* were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

61. Using the same $\bar{x}$, s_x, and *n* = 39, how would the error bound change if the confidence level were reduced to 90%? Why?

7.3 A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

62. 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?

63. 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?

Use the following information to answer the next five exercises: Suppose the marketing company did do a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

64. Identify the following:

a. x = _____
b. n = _____
c. p' = _____

65. Define the random variables X and P' in words.

66. Which distribution should you use for this problem?

67. Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

68. List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

69. We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and P' in words.

70. Which distribution should you use for this problem?

71. Construct a 95% confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

72. Suppose we want to lower the sampling error. What is one way to accomplish that?

73. The sampling error given in the survey is ±2%. Explain what the ±2% means.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the population proportion of voters who feel the economy is the most important.

74. Define the random variable X in words.

75. Define the random variable P' in words.

76. Which distribution should you use for this problem?

77. Construct a 90% confidence interval, and state the confidence interval and the error bound.

78. What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

79. What is being counted?

80. In words, define the random variable X.

81. Calculate the following:

a. x = _____
b. n = _____
c. p' = _____

82. State the estimated distribution of X. X~_____

83. Define a new random variable P'. What is p' estimating?

84. In words, define the random variable P'.

85. State the estimated distribution of P'. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

86. How much area is in both tails (combined)?

87. How much area is in each tail?

88. Calculate the following:

a. lower limit
b. upper limit
c. error bound

89. The 92% confidence interval is _____.

90. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

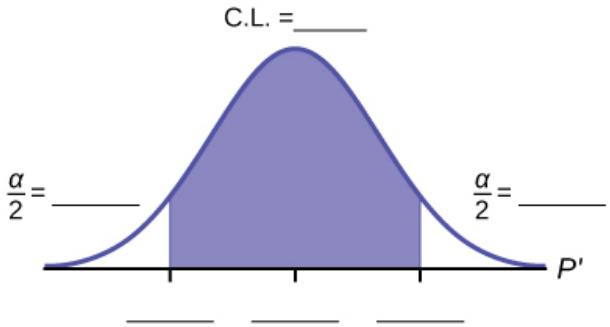


Figure 7.12

91. In one complete sentence, explain what the interval means.

92. Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

93. Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

94. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

7.4 Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

95. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

96. In words, define the random variables X and $\bar{X}$.

97. Which distribution should you use for this problem?

98. Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

99. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

100. Identify the following:

a. $\bar{x}$ = _____

b. σ = _____

c. n = _____

101. In words, define the random variables X and $\bar{X}$.

102. Which distribution should you use for this problem?

103. Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

104. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

105. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

106. Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

107. Identify the following:

a. $\bar{x}$ = _____

b. σ = _____

c. n = _____

108. In words, define the random variable X.

109. In words, define the random variable $\bar{X}$.

110. Which distribution should you use for this problem?

111. Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

112. Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

113. In complete sentences, explain why the confidence interval in **Exercise 7.111** is larger than in **Exercise 7.112**.

114. In complete sentences, give an interpretation of what the interval in **Exercise 7.112** means.

115. What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

116. What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

117. $\bar{x}$ = _____

118. n = _____

119. _____ = 15

120. In words, define the random variable $\bar{X}$.

121. What is $\bar{x}$ estimating?

122. Is σ_x known?

123. As a result of your answer to **Exercise 7.120**, state the exact distribution to use when calculating the confidence interval.

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

124. How much area is in both tails (combined)? α =_____

125. How much area is in each tail? $\frac{\alpha}{2}$ =_____

126. Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

127. The 95% confidence interval is:_____.

128. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

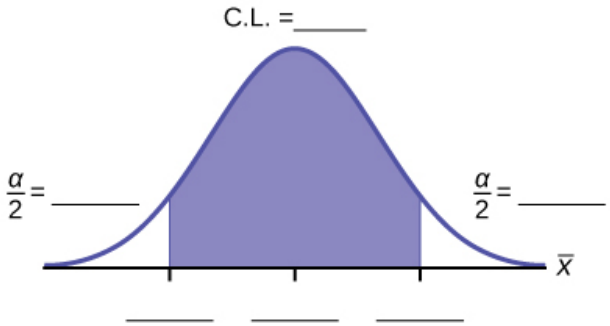


Figure 7.13

129. In one complete sentence, explain what the interval means.

130. Using the same mean, standard deviation, and level of confidence, suppose that *n* were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

131. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

HOMEWORK

7.1 A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

132. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean height of male Swedes.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

133. 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.

a. In words, define the random variables X and $\bar{X}$.

b. Which distribution should you use for this problem? Explain your choice.

c. Construct a 95% confidence interval for the population mean length of engineering conferences.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

134. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

f. If the firm did another survey, kept the error bound 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?

135. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. s_X =_____

b. In words, define the random variable X.

c. In words, define the random variable $\bar{X}$.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 90% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. Construct a 98% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.

h. In complete sentences, give an interpretation of what the interval in part f means.

136. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

137. 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.

138. 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. **Table 7.9** 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.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

Table 7.9

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the error bound.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

139. 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 error bound for mean U.S. household income.

140. 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?

7.2 A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

141. 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. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice
c. Calculate p'.
d. Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

142. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

143. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean time wasted.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

144. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variable X in words.

c. Define the random variable $\bar{X}$ in words.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 95% confidence interval for the population mean length of time.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. What does it mean to be "95% confident" in this problem?

145. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variable X in words.

c. Define the random variable $\bar{X}$ in words.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 99% confidence interval for the population mean length of time using training wheels.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. Why would the error bound change if the confidence level were lowered to 90%?

146. 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 20 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

Table 7.10

$\bar{x} = \$251,\ 854.23$

$s = \$521,\ 130.41$

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.

147. *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 7.11** 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

Table 7.11

Use this sample data to construct a 90% confidence interval for the mean age of CEO's for these top small firms. Use the Student's t-distribution.

148. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. n-1 = _____

b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

149. 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. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\bar{X}$ in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Explain what a "95% confidence interval" means for this study.

150. 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.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?

c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.

d. Calculate the mean.

e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

151. 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.
   i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. n-1 = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

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.

Use the following information to answer the next two exercises: 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.

152. Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

153. What is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

7.3 A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

154. 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?

155. 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.
   i. $x =$ _____
   ii. $n =$ _____
   iii. $p' =$ _____

b. Define the random variables X and P', in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

156. 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. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

157. 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.

a. 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. Define the random variables X and P', in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

158. Refer to the information in **Exercise 7.157**.

a. Construct three 95% confidence intervals.
   i. percent of all Asians who would welcome a white person into their families.
   ii. percent of all Asians who would welcome a Latino into their families.
   iii. percent 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?

159. 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 P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. Explain what a "97% confidence interval" means for this study.

160. 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 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.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. 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.

161. Refer to **Exercise 7.160**. 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. Define the random variables X and 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 are worried a lot about the quality of education in our schools.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. 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.

Use the following information to answer the next three exercises: 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.

162. A point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

163. A 90% confidence interval for the population proportion is _____.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

164. The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

Use the following information to answer the next two exercises: 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.

165. 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.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

166. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

a. 0.6614
b. 0.3386
c. 173
d. 338

167. 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 error bound based on the information provided.
e. Compare the error bound 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?

168. 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 error bound for this confidence interval.
b. Can we (with 95% confidence) conclude 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?

169. 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 error bound 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. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

170. 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?

7.4 Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA

171. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95% confidence interval for the population mean height of male Swedes.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

172. 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.

a. In words, define the random variables X and $\bar{X}$.

b. Which distribution should you use for this problem? Explain your choice.

c. Construct a 95% confidence interval for the population mean length of engineering conferences.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

173. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90% confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

f. If the firm did another survey, kept the error bound 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?

174. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. s_X =_____

b. In words, define the random variable X.
c. In words, define the random variable $\bar{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Construct a 98% confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
h. In complete sentences, give an interpretation of what the interval in part f means.

175. 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.
   i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

176. 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.

177. 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. **Table 7.12** 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.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

Table 7.12

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the error bound.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

178. 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 error bound for mean U.S. household income.

179. 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?

REFERENCES

7.1 A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

"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/2000growthchart-us.pdf (accessed July 2, 2013).

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).

7.2 A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA

"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).

7.3 A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA

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).

SOLUTIONS

1

a. 244

b. 15

c. 50

3 $N\left(244, \frac{15}{\sqrt{50}}\right)$

5 As the sample size increases, there will be less variability in the mean, so the interval size decreases.

7 X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

9 CI: (7.9441, 8.4559)

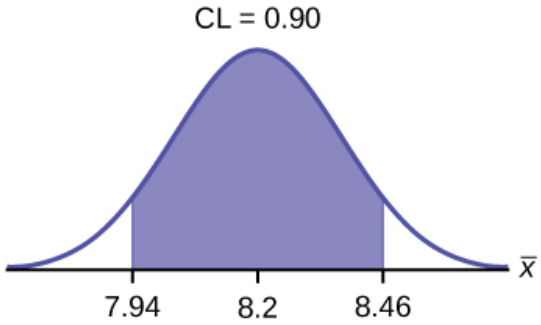


Figure 7.14

EBM = 0.26

11 The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

13

a. $\bar{x} = 2.2$

b. $\sigma = 0.2$

c. $n = 20$

15 $\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

17 *EBM* = 0.07
CI: (2.1264, 2.2736)

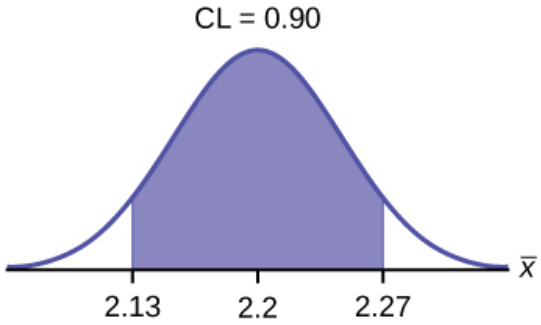


Figure 7.15

19 The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

21 The confidence level would increase.

23 30.4

25 σ

27 μ

29 normal

31 0.025

33 (24.52,36.28)

35 We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

37 The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

39 X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

41 CI: (1.3808, 1.6192)

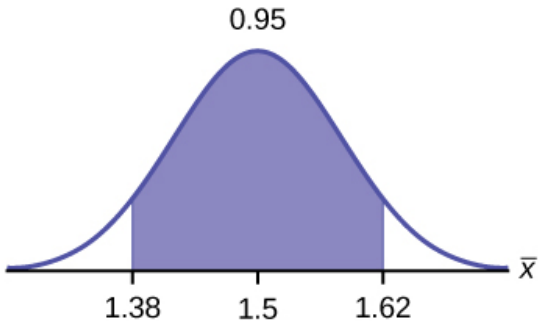


Figure 7.16

$EBM = 0.12$

43

a. $\bar{x} = 151$

b. $s_x = 32$

c. $n = 108$

d. $n - 1 = 107$

45 $\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

47 CI: (142.92, 159.08)

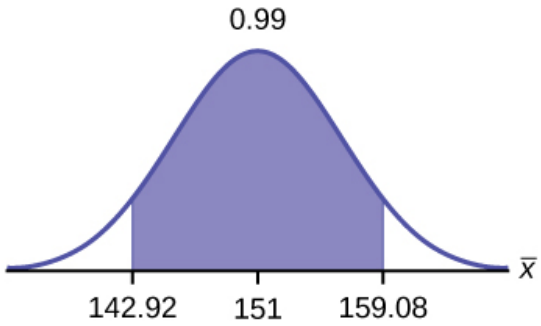


Figure 7.17

EBM = 8.08

49

a. 3.26

b. 1.02

c. 39

51 μ

53 t_{38}

55 0.025

57 (2.93, 3.59)

59 We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

60 The error bound would become EBM = 0.245. This error bound decreases because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

63 It would decrease, because the z-score would decrease, which reducing the numerator and lowering the number.

65 X is the number of "successes" where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

67 CI: (0.5321, 0.6679)

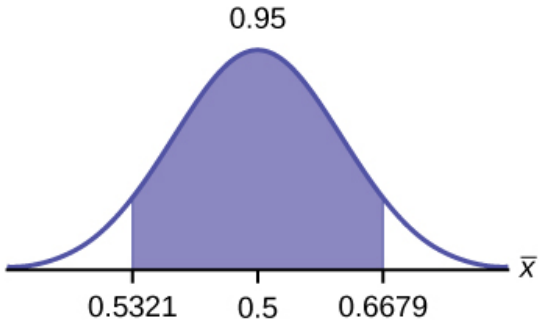


Figure 7.18

EBM: 0.0679

69 X is the number of "successes" where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

71 CI: (0.19432, 0.33068)

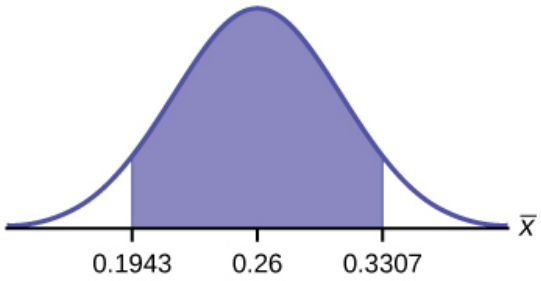


Figure 7.19

EBM: 0.0707

73 The sampling error means that the true mean can be 2% above or below the sample mean.

75 P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

77 CI: (0.62735, 0.67265) *EBM*: 0.02265

79 The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

81

a. $x = 64$

b. $n = 80$

c. $p' = 0.8$

83 p

85 $P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. (0.72171, 0.87829).

87 0.04

89 (0.72; 0.88)

91 With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

93 The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

95

a. 244

b. 15

c. 50

97 $N\left(244, \frac{15}{\sqrt{50}}\right)$

99 As the sample size increases, there will be less variability in the mean, so the interval size decreases.

101 X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

103 CI: (7.9441, 8.4559)

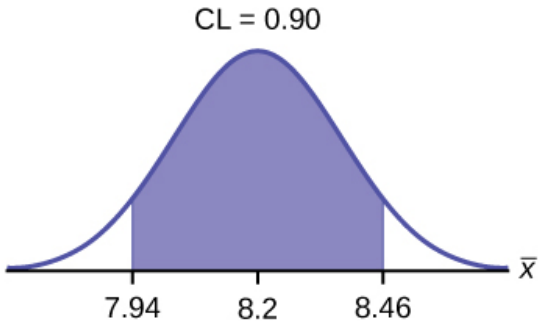


Figure 7.20

EBM = 0.26

105 The level of confidence would decrease because decreasing *n* makes the confidence interval wider, so at the same error bound, the confidence level decreases.

107

a. $\bar{x}$ = 2.2

b. σ = 0.2

c. n = 20

109 $\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

111 *EBM* = 0.07
CI: (2.1264, 2.2736)

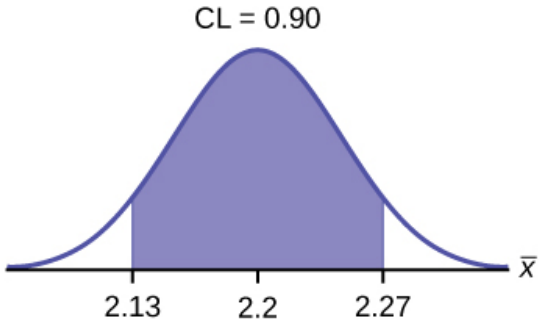


Figure 7.21

113 The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

115 The confidence level would increase.

117 30.4

119 σ

121 μ

123 normal

125 0.025

127 (24.52,36.28)

129 We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28.

131 The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

132

a. i. 71

ii. 3

iii. 48

b. X is the height of a Swiss male, and is the mean height from a sample of 48 Swiss males.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.49)

ii.

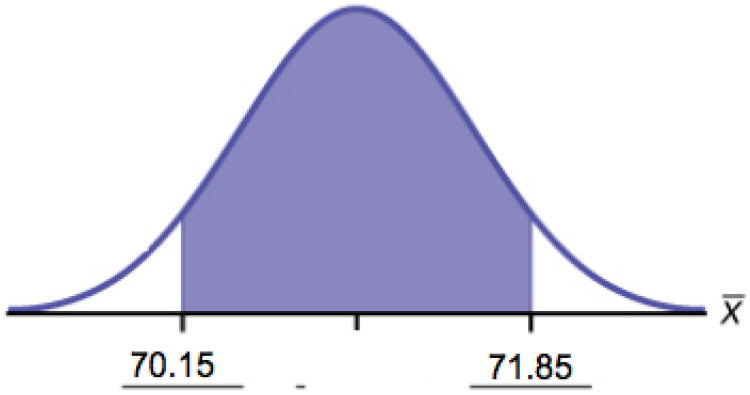


Figure 7.22

iii. $EBM = 0.849$

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

134

a. i. $\bar{x} = 23.6$

ii. $\sigma = 7$

iii. n = 100

b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

d. ii. (22.228, 24.972)

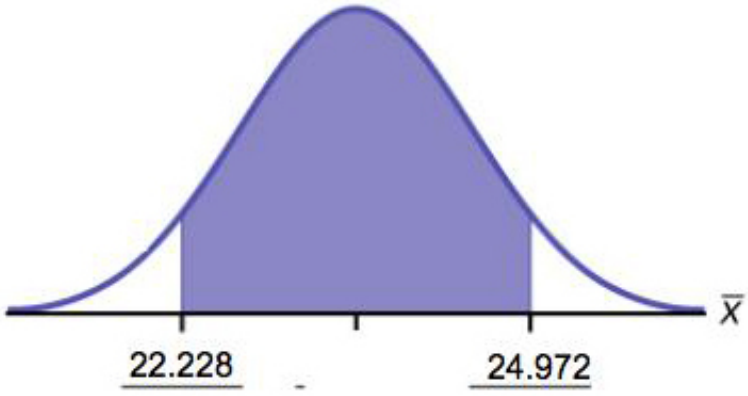


Figure 7.23

iii. *EBM* = 1.372

e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

136

a. i. 7.9

ii. 2.5

iii. 20

b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.

c. $N\ 7.9\left(\frac{2.5}{\sqrt{20}}\right)$

d. i. CI: (6.98, 8.82)

ii.

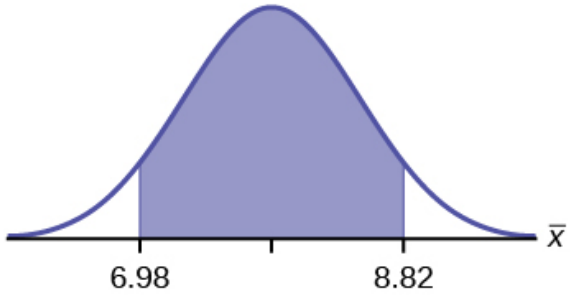


Figure 7.24

iii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

138

a. $\bar{x}$ = \$568,873

b. $CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $z_{\frac{\alpha}{2}} = 1.96$

$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}}$ = \$281,764

c. $\bar{x} - EBM = 568{,}873 - 281{,}764 = 287{,}109$

$\bar{x} + EBM = 568{,}873 + 281{,}764 = 850{,}637$

Alternate solution:

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

1. Press STAT and arrow over to TESTS.
2. Arrow down to 7:ZInterval.
3. Press ENTER.
4. Arrow to Stats and press ENTER.
5. Arrow down and enter the following values:

   σ: 909,200

   $\bar{x}$: 568,873

   n: 40

   CL: 0.95
6. Arrow down to Calculate and press ENTER.
7. The confidence interval is (\$287,114, \$850,632).
8. Notice the small difference between the two solutions–these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.

140 Use the formula for EBM, solved for n:

$n = \frac{z^2\sigma^2}{EBM^2}$ From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$. $z = z_{0.035} = 1.812$ (This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$ You need to measure at least 21 male students to achieve your goal.

142

a. i. 8629

ii. 6944

iii. 35

iv. 34

b. t_{34}

c. i. CI: (6244, 11,014)

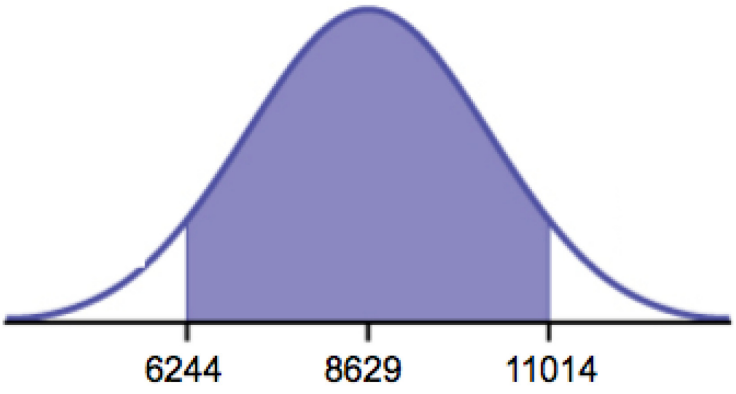
ii. 

Figure 7.25

iii. EB = 2385

d. It will become smaller

144

a. i. $\bar{x}$ = 2.51

ii. s_x = 0.318

iii. n = 9

iv. n - 1 = 8

b. the effective length of time for a tranquilizer

c. the mean effective length of time of tranquilizers from a sample of nine patients

d. We need to use a Student's-t distribution, because we do not know the population standard deviation.

e. i. CI: (2.27, 2.76)

ii. Check student's solution.

iii. *EBM*: 0.25

f. If we were to sample many groups of nine patients, 95% of the samples would contain the true population mean length of time.

146 $\bar{x} = \$251,854.23$ $s = \$521,130.41$ Note that we are not given the population standard deviation, only the standard deviation of the sample. There are 30 measures in the sample, so n = 30, and df = 30 - 1 = 29 CL = 0.96, so α = 1 - CL = 1 - 0.96 = 0.04 $\frac{\alpha}{2} = 0.02 t_{\frac{\alpha}{2}} = t_{0.02}$ = 2.150 $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.150\left(\frac{521,130.41}{\sqrt{30}}\right)$ ~ $\$204,561.66$ $\bar{x}$ - EBM = \$251,854.23 - \$204,561.66 = \$47,292.57 $\bar{x}$ + EBM = \$251,854.23+ \$204,561.66 = \$456,415.89 We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.

148

a. i. $\bar{x}$ =

ii. s_x =

iii. n =

iv. n - 1 =

b. X is the number of unoccupied seats on a single flight. $\bar{X}$ is the mean number of unoccupied seats from a sample of 225 flights.

c. We will use a Student's-t distribution, because we do not know the population standard deviation.

d. i. CI: (11.12 , 12.08)

ii. Check student's solution.

iii. *EBM*: 0.48

150

a. i. CI: (7.64 , 9.36)

ii.

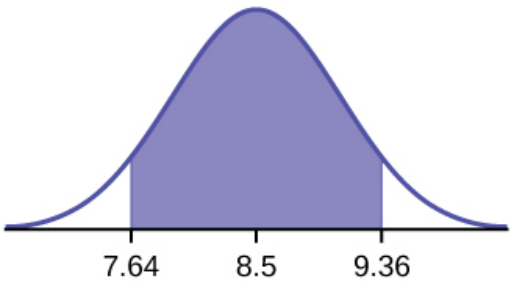


Figure 7.26

iii. *EBM*: 0.86

b. The sample should have been increased.

c. Answers will vary.

d. Answers will vary.

e. Answers will vary.

152 b

154

a. 1,068

b. The sample size would need to be increased since the critical value increases as the confidence level increases.

156

a. X = the number of people who feel that the president is doing an acceptable job;

P' = the proportion of people in a sample who feel that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

c. i. CI: (0.59, 0.63)

ii. Check student's solution

iii. *EBM*: 0.02

158

a. i. (0.72, 0.82)

ii. (0.65, 0.76)

iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.

c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.

d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

160

a. X = the number of adult Americans who feel that crime is the main problem; P' = the proportion of adult Americans who feel that crime is the main problem

b. Since we are estimating a proportion, given $P' = 0.2$ and $n = 1000$, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. CI: (0.18, 0.22)

ii. Check student's solution.

iii. *EBM*: 0.02

d. One way to lower the sampling error is to increase the sample size.

e. The stated "± 3%" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%. Thus, they estimate the percentage of adult Americans who feel that crime is the main problem to be between 18% and 22%.

162 c

164 d

166 a

168

a. $p' = \frac{(0.55 + 0.49)}{2} = 0.52$; $EBP = 0.55 - 0.52 = 0.03$

b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.

c. $CL = 0.75$, so $\alpha = 1 - 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125$ $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1 - 0.125 = 0.875$.)

$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1,\ 000}} \approx 0.018$

$(p' - EBP, p' + EBP) = (0.52 - 0.018, 0.52 + 0.018) = (0.502, 0.538)$

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

170 $CL = 0.95$ $\alpha = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$ $z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$n = \frac{z_{\frac{\alpha}{2}}^2 p' q'}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$ You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

171

a. i. 71

ii. 3

iii. 48

b. X is the height of a Swiss male, and is the mean height from a sample of 48 Swiss males.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.49)

ii.

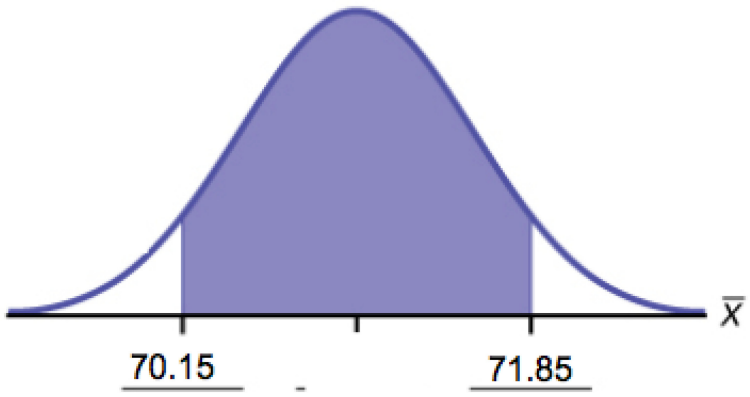


Figure 7.27

iii. $EBM = 0.849$

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

173

a. i. $\bar{x} = 23.6$

ii. $\sigma = 7$

iii. n = 100

b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

d. i. (22.228, 24.972)

ii.

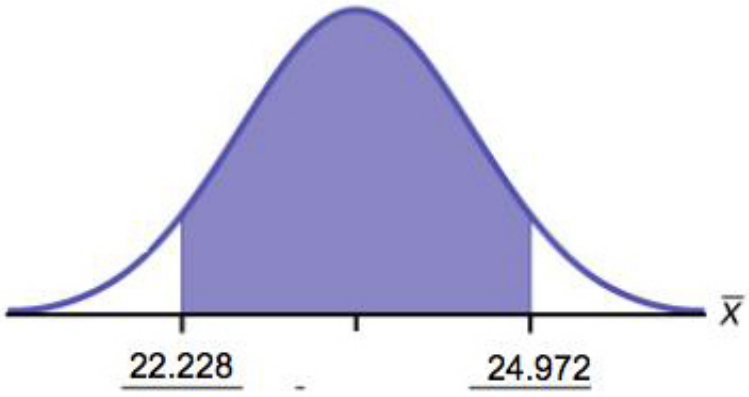


Figure 7.28

iii. *EBM* = 1.372

e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

175

a. i. 7.9

ii. 2.5

iii. 20

b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.

c. $N\ 7.9\left(\frac{2.5}{\sqrt{20}}\right)$

d. i. CI: (6.98, 8.82)

ii.

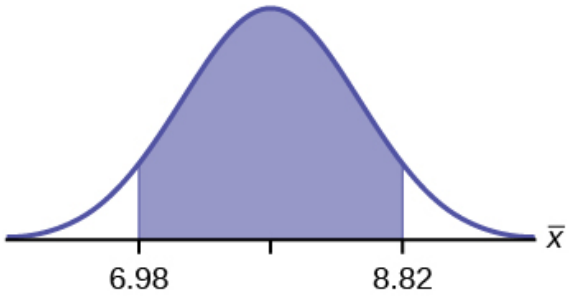


Figure 7.29

iii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

177

a. $\bar{x}$ = $568,873

b. $CL = 0.95\ \alpha = 1 - 0.95 = 0.05\ z_{\frac{\alpha}{2}} = 1.96$

$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}} = \$281,764$

c. $\bar{x} - EBM = 568,873 - 281,764 = 287,109$
$\bar{x} + EBM = 568,873 + 281,764 = 850,637$

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between $287,109 and $850,637.

179 Use the formula for *EBM*, solved for *n*:

$n = \frac{z^2\sigma^2}{EBM^2}$ From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$. $z = z_{0.035} =$

1.812 (This is the value of z for which the area under the density curve to the ***right*** of z is 0.035.)

$n = \frac{z^2 \sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$ You need to measure at least 21 male students to achieve your goal.

8 | HYPOTHESIS TESTING WITH ONE SAMPLE

Figure 8.1 You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

Introduction

CHAPTER OBJECTIVES

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

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.

- Conduct and interpret hypothesis tests for a single population proportion.

8.1 | Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis:** It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a: **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we cannot accept H_0. The alternative hypothesis is the contender and must win with significant evidence to overthrow the status quo. This concept is sometimes referred to the tyranny of the status quo.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "cannot accept H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Table 9.1 presents the various hypotheses in the relevant pairs. For example, if the null hypothesis is equal to some value, the alternative has to be not equal to that value.

H_0	H_a
equal (=)	not equal (≠)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Table 8.1

NOTE

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test.

Example 8.1

H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Try It

8.1 A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Example 8.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:
H_0: $\mu = 2.0$
H_a: $\mu \neq 2.0$

Try It

8.2 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 8.3

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:
H_0: $\mu \geq 5$
H_a: $\mu < 5$

Try It

8.3 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 8.4

In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.
H_0: $p \leq 0.066$
H_a: $p > 0.066$

Try It

8.4 On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: p __ 0.40

b. H_a: p __ 0.40

8.2 | Outcomes and the Type I and Type II Errors -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

STATISTICAL DECISION	***H₀* IS ACTUALLY...**	
	True	False
Cannot reject *H₀*	Correct Outcome	Type II error
Cannot accept *H₀*	Type I Error	Correct Outcome

Table 8.2

The four possible outcomes in the table are:

1. The decision is **cannot reject *H₀*** when ***H₀* is true (correct decision).**
2. The decision is **cannot accept *H₀*** when ***H₀* is true** (incorrect decision known as a **Type I error**). This case is described as "rejecting a good null". As we will see later, it is this type of error that we will guard against by setting the probability of making such an error. The goal is to NOT take an action that is an error.
3. The decision is **cannot reject *H₀*** when, in fact, ***H₀* is false** (incorrect decision known as a **Type II error**). This is called "accepting a false null". In this situation you have allowed the status quo to remain in force when it should be overturned. As we will see, the null hypothesis has the advantage in competition with the alternative.
4. The decision is **cannot accept *H₀*** when ***H₀* is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors.

Statistics allows us to set the probability that we are making a Type I error. The probability of making a Type I error is α. Recall that the confidence intervals in the last unit were set by choosing a value called Z_α (or t_α) and the alpha value determined the confidence level of the estimate because it was the probability of the interval capturing the true mean (or proportion parameter p). This alpha and that one are the same.

By way of example, the American judicial system begins with the concept that a defendant is "presumed innocent". This is the status quo and is the null hypothesis. The judge will tell the jury that they can not find the defendant guilty unless the evidence indicates guilt beyond a "reasonable doubt" which is usually defined in criminal cases as 95% certainty of guilt. If the jury cannot accept the null, innocent, then action will be taken, jail time. The burden of proof always lies with the alternative hypothesis. (In civil cases, the jury needs only to be more than 50% certain of wrongdoing to find culpability, called "a preponderance of the evidence").

The example above was for a test of a mean, but the same logic applies to tests of hypotheses for all statistical parameters one may wish to test.

The following are examples of Type I and Type II errors.

Example 8.5

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. **Type II 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.)

This is a situation described as "accepting a false null".

Try It

8.5 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 8.6

Suppose the null hypothesis, H_0, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Try It

8.6 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 8.7

It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome. The status quo is that the claim is false. The burden of proof always falls to the person making the claim, in this case the Genetics Lab.

Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.

Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Try It

8.7 "Red tide" is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Example 8.8

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is at least 75%.

Type II: A cancer patient believes the experimental drug has at least a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Try It

8.8 Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%.

Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

8.3 | Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

Earlier in the course, we discussed sampling distributions. Particular distributions are associated with hypothesis testing.We will perform hypotheses tests of a population mean using a normal distribution or a Student's *t*-distribution. (Remember, use a Student's *t*-distribution when the population standard deviation is unknown and the sample size is small, where small is considered to be less than 30 observations.) We perform tests of a population proportion using a normal distribution when we can assume that the distribution is normally distributed. We consider this to be true if the sample proportion, p', times the sample size is greater than 5 and 1- p' times the sample size is also greater then 5. This is the same rule of thumb we used when developing the formula for the confidence interval for a population proportion.

Assumptions

When you perform a **hypothesis test of a single population mean *μ*** using a **Student's *t*-distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean *μ*** using a normal distribution (often called a *z*-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion *p***, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are: there are a certain number *n* of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success *p*. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities *np* and *nq* must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Hypothesis Test for the Mean

Going back to the standardizing formula we can derive the **test statistic** for testing hypotheses concerning means.

$$Z_c = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

The standardizing formula can not be solved as it is because we do not have μ, the population mean. However, if we substitute in the hypothesized value of the mean, μ_0 in the formula, we can compute a Z value.

This calculated Z is nothing more than the number of standard deviations that the **hypothesized** mean is from the sample mean. If the sample mean falls "too many" standard deviations from the hypothesized mean we conclude that the **sample** mean could not have come from the distribution with the hypothesized mean, given our pre-set required level of confidence. It **could** have come from H_0, but it is deemed just too unlikely. If in fact this sample mean did come from H_0, but from in the tail, we have made a Type I error. Our only real comfort is that we know the probability of making such an error, α, and we can control the size of α.

This gives us the decision rule for testing a hypothesis:

Decision Rule: Two-tail Test
If $Z_c < \left\|Z_{\frac{\alpha}{2}}\right\|$: then cannot REJECT H_0
If $Z_c > \left\|Z_{\frac{\alpha}{2}}\right\|$: then cannot ACCEPT H_0

Table 8.3

This rule will always be the same no matter what hypothesis we are testing or what formulas we are using to make the test. The only change will be to change the Z_c to the appropriate symbol for the test statistic for the parameter being tested.

An alternative decision rule can be developed by calculating the probability that the sample mean came from the hypothesized distribution and comparing that probability with the probability that the analyst has decided as the level of confidence, α.This probability is called the *p*-value, p for probability. A large *p*-value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the *p*-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

8.4 | Rare Events, the Sample, Decision and Conclusion -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption**—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely,

Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the ***p*-value**. The *p*-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large *p*-value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the *p*-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the *p*-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 8.9

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. and the distribution of heights is normal.

The null hypothesis could be H_0: $\mu \leq 15$ The alternate hypothesis is H_a: $\mu > 15$

The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since ***σ* is known** ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The *p*-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The *p*-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

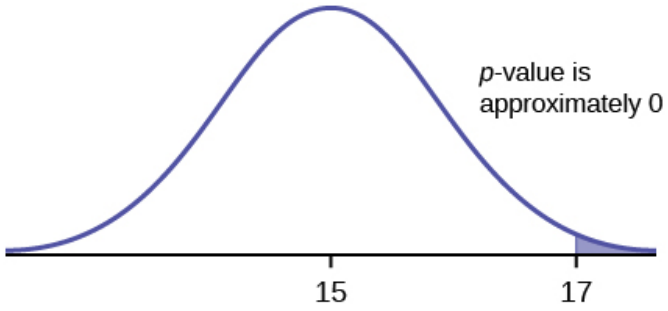


Figure 8.2

p-value= $P(\bar{x} > 17)$ which is approximately zero.

A *p*-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Try It

8.9 A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

H_0: $\mu \leq 12$
H_a: $\mu > 12$
The *p*-value is 0.0013
Draw a graph that shows the *p*-value.

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the *p*-value and a **preset or preconceived α (also called a "significance level")**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_0, do as follows:

- If $\alpha > p$-value, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If $\alpha \leq p$-value, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis,H_a, may be correct.

- When you "do not reject H_0", it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Example 8.10

When using the *p*-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

If the *p*-value is low, the null must go.

If the *p*-value is high, the null must fly.

This memory aid relates a *p*-value less than the established alpha (the *p* is low) as rejecting the null hypothesis and, likewise, relates a *p*-value higher than the established alpha (the *p* is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null when hypothesis when _____.

The results of the sample data _____.

Solution 8.10

Reject the null hypothesis when **the *p*-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the *p*-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Try It

8.10 It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

H_0: $p = 0.50$, H_a: $p > 0.50$

$\alpha = 0.01$

p-value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

8.5 | Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, a common standard to use is $\alpha = 0.05$.**
- When you calculate the *p*-value and draw the picture, the *p*-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.

- The **alternative hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the *p*-value**: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller *p*-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large *p*-value such as 0.4, as opposed to a *p*-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 8.11

H_0: $\mu = 5$, H_a: $\mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the *p*-value is as follows:

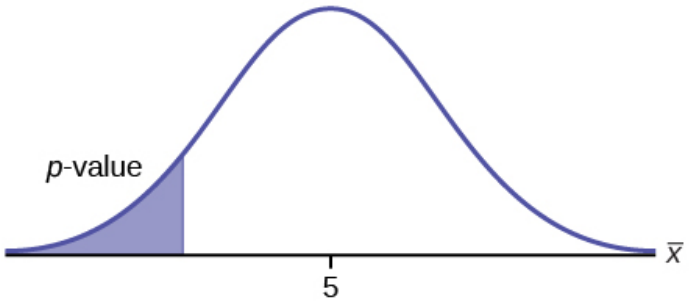


Figure 8.3

Try It

8.11 H_0: $\mu = 10$, H_a: $\mu < 10$

Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

Example 8.12

H_0: $p \leq 0.2$ H_a: $p > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:

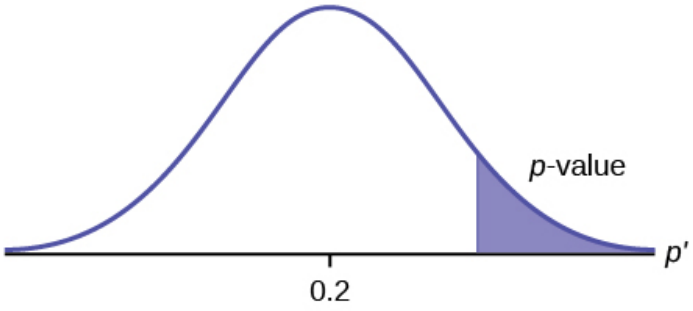


Figure 8.4

Try It

8.12 H_0: $\mu \leq 1$, H_a: $\mu > 1$

Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

Example 8.13

H_0: $p = 50$ H_a: $p \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the *p*-value is as follows.

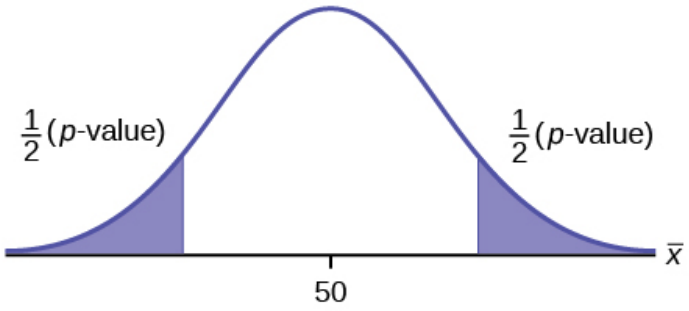


Figure 8.5

Try It

8.13 H_0: $p = 0.5$, H_a: $p \neq 0.5$

Assume the *p*-value is 0.2564. What type of test is this? Draw the picture of the *p*-value.

Full Hypothesis Test Examples

Example 8.14

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution 8.14

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

H_0: $\mu = 16.43$ H_a: $\mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population **standard deviation** is known: $\sigma = 0.8$)

$\bar{X} \sim N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right)$ Therefore, $\bar{X} \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the p-value using the normal distribution for a mean:

p-value = $P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187 (This is called the **actual level of significance**.) The p-value is the area to the left of the sample mean is given as 16.

Graph:

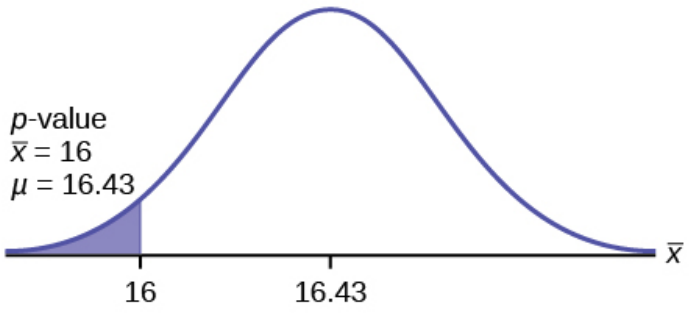


Figure 8.6

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the p-value: If H_0 is true, there is a 0.0187 probability (1.87%)that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the p-value:

$\alpha = 0.05$ p-value = 0.0187 $\alpha > p$-value

Make a decision: Since $\alpha > p$-value, reject H_0.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated.

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

Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value ($p = 0.0187$) but it also calculates the test statistic (z-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw(instead of Calculate). Press ENTER. A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the Z-Test function finds the p-value by doing a normal probability calculation using the **central limit theorem**:

$P(\bar{x} < 16) =$ 2nd DISTR normcdf $(-10\text{^}99, 16, 16.43, 0.8/\sqrt{15})$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Try It

8.14 The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

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

Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for $\mu 0$ (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value but it also calculates the test statistic (z-score) for the sample mean. Select <, ≠, or > for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and p-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

HISTORICAL NOTE (EXAMPLE 8.11)

The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{\bar{x} - \mu_X}{(\frac{\sigma_X}{\sqrt{n}})}$. For this problem, $\bar{x} = 16$, $\mu_X = 16.43$ from the null hypothes is, $\sigma_X = 0.8$, and $n = 15$.) You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p$-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, and the test statistics and the critical value.

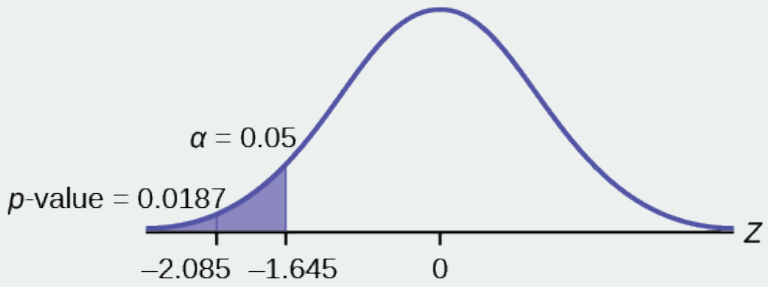


Figure 8.7

Example 8.15

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution 8.15

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a **test of a single population mean**.

H_0: $\mu = 275$
H_a: $\mu > 275$
This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X}$ = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$\bar{X} \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

$\bar{x} = 286.2$ pounds (from the data).

$\sigma = 55$ pounds **(Always use σ if you know it.)** We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see **m47896 (http://legacy.cnx.org/content/m47896/latest/)** for using the data as input):

$p\text{-value} = P(\bar{x} > 286.2) = 0.1323$.

Interpretation of the p-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

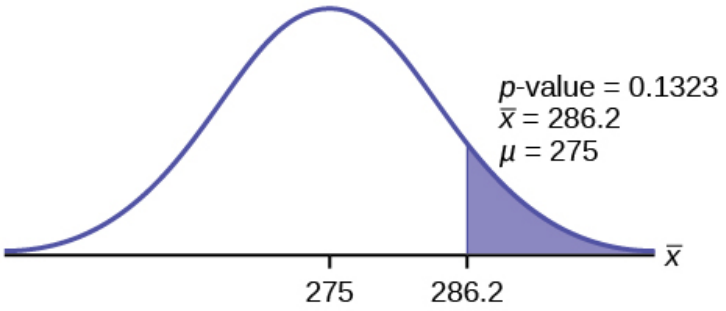


Figure 8.8

Compare α and the p-value:

$\alpha = 0.025$ p-value = 0.1323

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated.

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

Put the data and frequencies into lists. Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Data and press ENTER. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you

put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value ($p = 0.1331$, a little different from the previous calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 8.16

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Solution 8.16

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

H_0: $\mu = 65$ H_a: $\mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's t-distribution:

p-value = $P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.

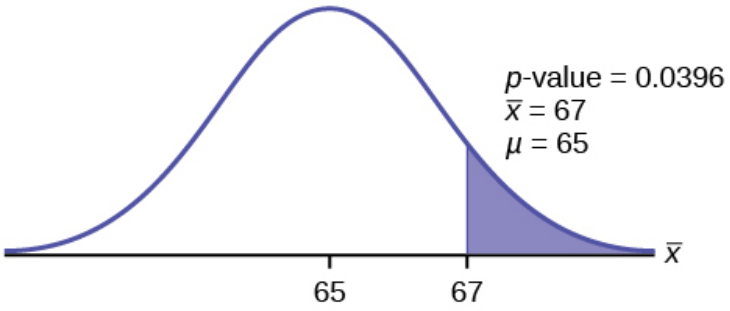


Figure 8.9

Compare α and the p-value:

Since $\alpha = 0.05$ and p-value = 0.0396. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

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

Put the data into a list. Press STAT and arrow over to TESTS. Press 2:T-Test. Arrow over to Data and press ENTER. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for Freq:. Arrow down to μ: and arrow over to $> \mu_0$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value ($p = 0.0396$) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Try It

8.16 It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Example 8.17

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time**

brides and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution 8.17

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

H_0: $p = 0.50$ H_a: $p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percent of of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ Therefore, $P' \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$

where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$

Calculate the p-value using the normal distribution for proportions:

p-value = $P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the *p*-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in **Figure 8.9**).

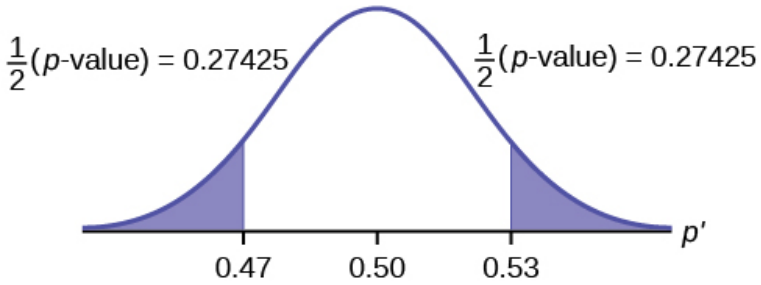


Figure 8.10

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to 0.50 – 0.03 = 0.47 where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value = 0.5485. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated.

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

Press STAT and arrow over to TESTS. Press 5:1-PropZTest. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to Prop and arrow to not equals p_0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). Prop not equals .5 is the alternate hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides who are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Try It

8.17 A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Example 8.18

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution 8.18

Set up the Hypothesis Test:

H_0: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N\left(0.30, \sqrt{\frac{(0.30) \cdot (0.70)}{150}}\right)$

a. The value that helps determine the p-value is p'. Calculate p'.

Solution 8.18

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$x = 43$, $n = 150$

$p' = \frac{43}{150}$

b. What is a **success** for this problem?

Solution 8.18
b. A success is having three cell phones in a household.

c. What is the level of significance?

Solution 8.18
c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.
Calculate the p-value.

Solution 8.18
d. p-value = 0.7216

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Solution 8.18
e. Assuming that $\alpha = 0.05$, $\alpha < p$-value. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Try It

8.18 Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p-value, state your conclusion, and identify the Type I and Type II errors.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example 8.19

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,

Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Use the new shampoo or go without?

Solution 8.19

Set up the hypothesis test:

H_0: $p \leq 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable $\bar{X}$ or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

p-value = 0.0103

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.

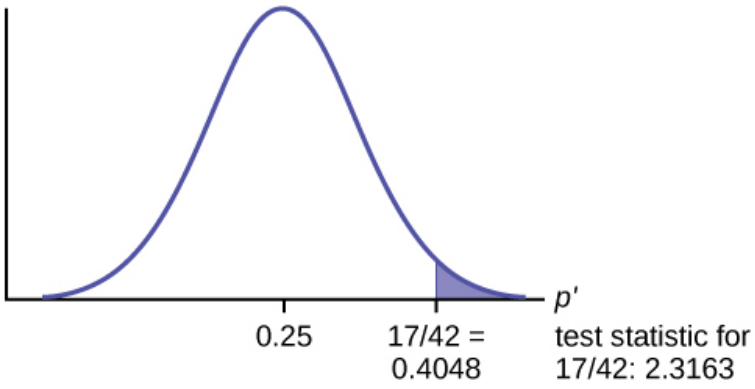


Figure 8.11

Compare α and the p-value:

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	$\alpha < p$-value

Table 8.4

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

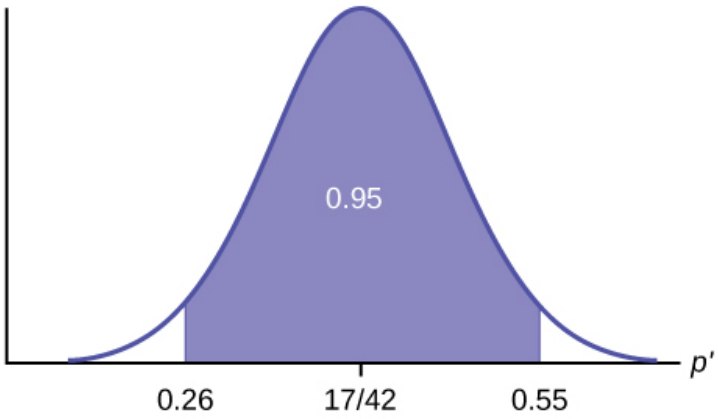


Figure 8.12

Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion p of fleas that are killed by the new shampoo is between 26% and 55%.

NOTE

This test result is not very definitive since the *p*-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example 8.20

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95
Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Solution 8.20

Let's follow a four-step process to answer this statistical question.

1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
   a. H_0: $\mu \leq 1$
   b. H_a: $\mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's-t distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

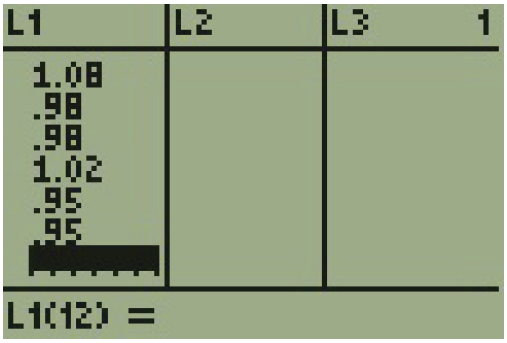


Figure 8.13

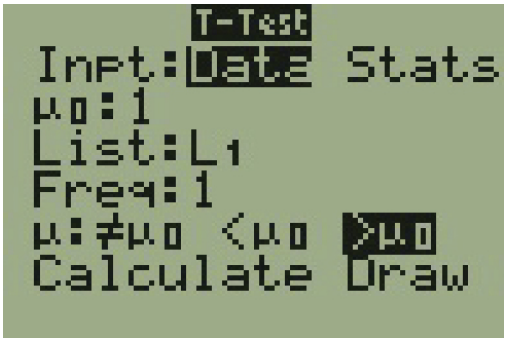


Figure 8.14

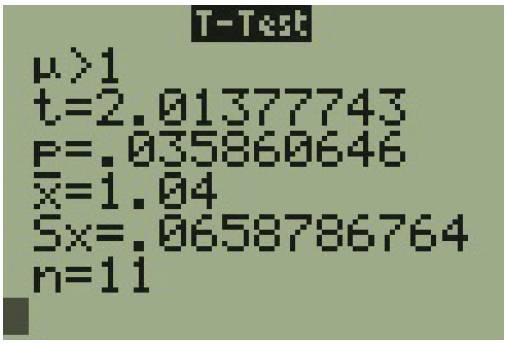


Figure 8.15

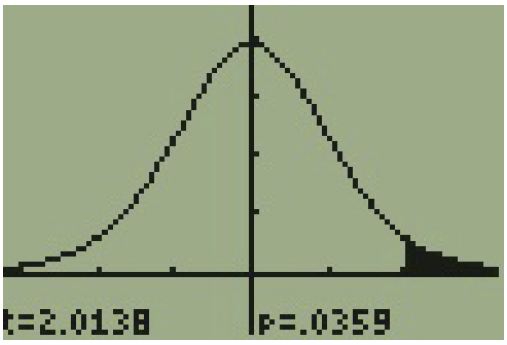


Figure 8.16

4. **State the Conclusions**: Since the *p*-value* ($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example 8.21

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 8.21

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
   a. H_0: $p \leq 0.00034$
   b. H_a: $p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420{,}019$. The sample is sufficiently large because we have $np = 420{,}019(0.00034) = 142.8$, $nq = 420{,}019(0.99966) = 419{,}876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are

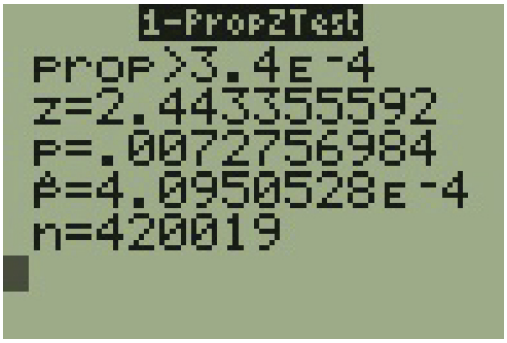


Figure 8.17

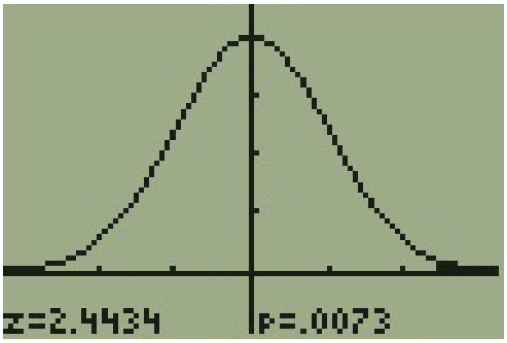


Figure 8.18

4. Since the p-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example 8.22

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess

County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Solution 8.22

We will follow the four-step plan.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion *z*-test. The hypotheses for the test will be
   a. H_0: $p = 0.00078$
   b. H_a: $p \neq 0.00078$
3. The following screen shots display the summary statistics from the hypothesis test.

Figure 8.19

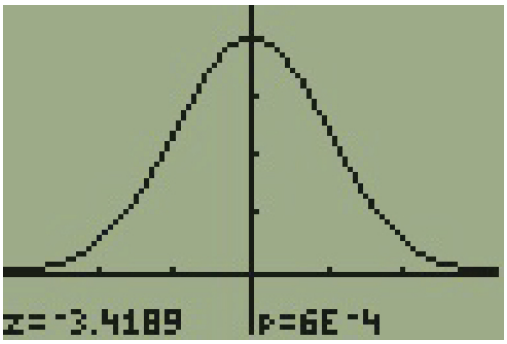


Figure 8.20

4. Since the *p*-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national average proportion.

KEY TERMS

Binomial Distribution a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Central Limit Theorem Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Confidence Interval (CI) an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Critical Value The t or Z value set by the researcher that measures the probability of a Type I error, α.

Hypothesis a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Hypothesis Testing Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

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

Normal Distribution a continuous random variable (RV) with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

***p*-value** the probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

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

Student's *t*-Distribution investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

Test Statistic The formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.

Type I Error The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type II Error The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

CHAPTER REVIEW

8.1 Null and Alternative Hypotheses

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤ or ≥)
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using not equal, less than or greater than symbols, i.e., (≠, <, or >).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

8.2 Outcomes and the Type I and Type II Errors -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

8.3 Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's *t*-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

8.4 Rare Events, the Sample, Decision and Conclusion -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the *p*-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. $\alpha > p$-value, reject the null hypothesis
2. $\alpha \leq p$-value, do not reject the null hypothesis

8.5 Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the *p*-value. (A *z*-score and a *t*-score are examples of test statistics.)
5. Compare the preconceived α with the *p*-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the *p*-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same.If the power is low, the null hypothesis might not be rejected when it should be.

FORMULA REVIEW

8.3 Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

Sample Size	Test Statistic
< 30 (σ unknown)	$t_c = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$
< 30 (σ known)	$Z_c = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

Table 8.5 Test Statistics for Test of Means, Varying Sample Size, Population Known or Unknown

Sample Size	Test Statistic
> 30 (σ unknown)	$Z_c = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$
> 30 (σ known)	$Z_c = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

Table 8.5 Test Statistics for Test of Means, Varying Sample Size, Population Known or Unknown

PRACTICE

8.1 Null and Alternative Hypotheses

1. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

2. You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

3. The American family has an average of two children. What is the random variable? Describe in words.

4. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

5. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

6. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

7. In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

8. Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: _____
b. H_a: _____

9. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

10. 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. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

8.2 Outcomes and the Type I and Type II Errors -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

11. The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

12. A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

13. For **Exercise 9.12**, what are α and β in words?

14. In words, describe $1 - \beta$ For **Exercise 9.12**.

15. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

16. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

17. The power of a test is 0.981. What is the probability of a Type II error?

18. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

19. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

20. A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0, is: the sample contains E-coli. Which is the error with the greater consequence?

8.3 Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

21. Which two distributions can you use for hypothesis testing for this chapter?

22. Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume sample size is large.

23. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

24. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

25. A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

26. It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

27. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?

28. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

29. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

30. You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

31. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

8.4 Rare Events, the Sample, Decision and Conclusion -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

32. When do you reject the null hypothesis?

33. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

34. The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

35. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The *p*-value is almost zero. State the null and alternative hypotheses and interpret the *p*-value.

36. The mean age of graduate students at a University is at most 31 y ears with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The *p*-value is 0.0264. State the null and alternative hypotheses and interpret the *p*-value.

37. Does the shaded region represent a low or a high *p*-value compared to a level of significance of 1%?

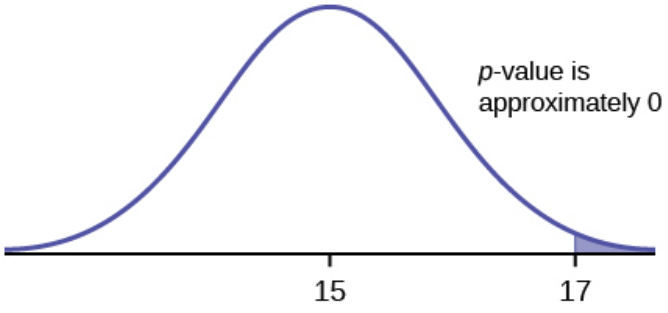


Figure 8.21

38. What should you do when $\alpha > p$-value?

39. What should you do if $\alpha = p$-value?

40. If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

41. Is this a test of means or proportions?

42. What symbol represents the random variable for this test?

43. In words, define the random variable for this test.

44. Is the population standard deviation known and, if so, what is it?

45. Calculate the following:

a. $\bar{x}$ _____
b. σ _____
c. s_x _____
d. n _____

46. Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

47. State the distribution to use for the hypothesis test.

48. A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   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. Calculate the following:
   i. $\bar{x}$ = _____
   ii. s = _____
   iii. n = _____
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

8.5 Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

49. Assume H_0: $\mu = 9$ and H_a: $\mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

50. Assume H_0: $\mu \leq 6$ and H_a: $\mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

51. Assume H_0: $p = 0.25$ and H_a: $p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

52. Draw the general graph of a left-tailed test.

53. Draw the graph of a two-tailed test.

54. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

55. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

56. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

57. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use?

58. If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

59. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

60. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

61. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

HOMEWORK

8.1 Null and Alternative Hypotheses

62. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (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 ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

63. 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. Is there good evidence 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. $p > 0.30$

64. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

65. 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. Conduct a hypothesis test. The null and alternative hypotheses are:

a. H_o: $\bar{x} = 4.5$, H_a : $\bar{x} > 4.5$
b. H_o: $\mu \geq 4.5$, H_a: $\mu < 4.5$
c. H_o: $\mu = 4.75$, H_a: $\mu > 4.75$
d. H_o: $\mu = 4.5$, H_a: $\mu > 4.5$

8.2 Outcomes and the Type I and Type II Errors -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

66. State the Type I and Type II errors in complete sentences given the following statements.

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 ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

67. For statements a-j in **Exercise 9.109**, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

68. 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.

69. 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. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is at least 20%.
d. less than 20%, when in fact, it is less than 20%.

70. 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. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to 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

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

71. 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. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

8.3 Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

72. 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. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ _____

a. $N(7.24, \frac{1.93}{\sqrt{22}})$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

8.4 Rare Events, the Sample, Decision and Conclusion -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

73. 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.
   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. Calculate the following:
   i. x = _____
   ii. n = _____
   iii. p' = _____
g. Calculate σ_x = _____. Show the formula set-up.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, what is your:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

8.5 Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***m47882 (http://legacy.cnx.org/content/m47882/latest/)*** *. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's-*t* distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

74. 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?

75. 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?

76. 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?

77. 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?

78. 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. Should the personnel team believe that the mean number is ten?

79. 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 ten. Does it appear that the mean work week has increased for women at the 5% level?

80. 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?

81. A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5; 4; 7; 3; 6; 4; 5; 3; 6; 3; 8; 5. Conduct a hypothesis test of your belief.

82. Refer to **Exercise 9.119**. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout’s mean I.Q. is **not** four.

83. 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?

84. 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.

85. The mean work 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. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

86. Use the "Lap time" data for Lap 4 (see **m47873 (http://legacy.cnx.org/content/m47873/latest/)**) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

87. Use the "Initial Public Offering" data (see **m47873 (http://legacy.cnx.org/content/m47873/latest/)**) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

NOTE

The following questions were written by past students. They are excellent problems!

88. "Asian Family Reunion," by Chau Nguyen

Every two years it comes around.
We all get together from different towns.
In my honest opinion,
It's not a typical family reunion.
Not forty, or fifty, or sixty,
But how about seventy companions!
The kids would play, scream, and shout
One minute they're happy, another they'll pout.
The teenagers would look, stare, and compare
From how they look to what they wear.
The men would chat about their business
That they make more, but never less.
Money is always their subject
And there's always talk of more new projects.
The women get tired from all of the chats
They head to the kitchen to set out the mats.
Some would sit and some would stand
Eating and talking with plates in their hands.
Then come the games and the songs
And suddenly, everyone gets along!
With all that laughter, it's sad to say
That it always ends in the same old way.
They hug and kiss and say "good-bye"
And then they all begin to cry!
I say that 60 percent shed their tears
But my mom counted 35 people this year.
She said that boys and men will always have their pride,
So we won't ever see them cry.
I myself don't think she's correct,
So could you please try this problem to see if you object?

89. "The Problem with Angels," by Cyndy Dowling

Although this problem is wholly mine,
The catalyst came from the magazine, Time.
On the magazine cover I did find
The realm of angels tickling my mind.
Inside, 69% I found to be
In angels, Americans do believe.
Then, it was time to rise to the task,
Ninety-five high school and college students I did ask.
Viewing all as one group,
Random sampling to get the scoop.
So, I asked each to be true,
"Do you believe in angels?" Tell me, do!
Hypothesizing at the start,
Totally believing in my heart
That the proportion who said yes
Would be equal on this test.
Lo and behold, seventy-three did arrive,
Out of the sample of ninety-five.
Now your job has just begun,
Solve this problem and have some fun.

90. "Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,
I had to find some sane defense.
Some light and lifting simple play
To float my math anxiety away.
Blowing bubbles lifts me high
Takes my troubles to the sky.
POIK! They're gone, with all my stress
Bubble therapy is the best.
The label said each time I blew
The average number of bubbles would be at least 22.
I blew and blew and this I found
From 64 blows, they all are round!
But the number of bubbles in 64 blows
Varied widely, this I know.
20 per blow became the mean
They deviated by 6, and not 16.
From counting bubbles, I sure did relax
But now I give to you your task.
Was 22 a reasonable guess?
Find the answer and pass this test!

91. "Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles
Bred puppies with numerous freckles
The Dalmatians he sought
Possessed spot upon spot
The more spots, he thought, the more shekels.
His competitors did not agree
That freckles would increase the fee.
They said, “Spots are quite nice
But they don't affect price;
One should breed for improved pedigree.”
The breeders decided to prove
This strategy was a wrong move.
Breeding only for spots
Would wreak havoc, they thought.
His theory they want to disprove.
They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.
They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, “Well, shucks,
It's for one hundred one that I aim.”
Said an amateur statistician
Who wanted to help with this mission.
“Twenty-one for the sample
Standard deviation's ample:
They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle and dot
And tallied up every one.
Instead of one hundred one spots
They averaged ninety six dots
Can they muzzle Spreckles’
Obsession with freckles
Based on all the dog data they've got?

92. "Macaroni and Cheese, please!!" by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese:

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ 0.75.

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. (After all, as a poor starving student I can't be expected to feed our class of animals!)

93. "William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of Statistics
- POLONIUS, Hamlet's tutor
- HOROTIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

(The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.)

POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS If thou doest insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What piece of work is man, how noble in reason, how infinite in faculties? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

(Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.)

Act II

(The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.)

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

94. "Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1%.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

95. "Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." (The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation.) "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters which have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names which end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', co-workers, and acquaintances' names that I could remember. Following are the names. (Some are repeats.) Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

96. "Phillip’s Wish," by Suzanne Osorio

My nephew likes to play
Chasing the girls makes his day.
He asked his mother
If it is okay
To get his ear pierced.
She said, “No way!”
To poke a hole through your ear,
Is not what I want for you, dear.
He argued his point quite well,
Says even my macho pal, Mel,
Has gotten this done.
It’s all just for fun.
C’mon please, mom, please, what the hell.
Again Phillip complained to his mother,
Saying half his friends (including their brothers)
Are piercing their ears
And they have no fears
He wants to be like the others.
She said, “I think it’s much less.
We must do a hypothesis test.
And if you are right,
I won’t put up a fight.
But, if not, then my case will rest.”
We proceeded to call fifty guys
To see whose prediction would fly.
Nineteen of the fifty
Said piercing was nifty
And earrings they’d occasionally buy.
Then there’s the other thirty-one,
Who said they’d never have this done.
So now this poem’s finished.
Will his hopes be diminished,
Or will my nephew have his fun?

97. "The Craven," by Mark Salangsang

Once upon a morning dreary
In stats class I was weak and weary.
Pondering over last night's homework
Whose answers were now on the board
This I did and nothing more.
While I nodded nearly napping
Suddenly, there came a tapping.
As someone gently rapping,
Rapping my head as I snore.
Quoth the teacher, "Sleep no more."
"In every class you fall asleep,"
The teacher said, his voice was deep.
"So a tally I've begun to keep
Of every class you nap and snore.
The percentage being forty-four."
"My dear teacher I must confess,
While sleeping is what I do best.
The percentage, I think, must be less,
A percentage less than forty-four."
This I said and nothing more.
"We'll see," he said and walked away,
And fifty classes from that day
He counted till the month of May
The classes in which I napped and snored.
The number he found was twenty-four.
At a significance level of 0.05,
Please tell me am I still alive?
Or did my grade just take a dive
Plunging down beneath the floor?
Upon thee I hereby implore.

98. 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%.

99. Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California's percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California, was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California. NOTE: For more accurate results, use more California community colleges and this past year's data.

100. 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. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

101. 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.

102. Registered nurses earned an average annual salary of $69,110. For that same year, 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 a hypothesis test.

103. La Leche League International reports that the mean age of weaning a child from breastfeeding 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.

104. 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. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?
After conducting the test, your decision and conclusion are

a. Reject H_0: There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0: There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0: There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject H_0: There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

105. A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of 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%.

106. 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. Conduct a hypothesis test.

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

Instructions: For the following ten exercises,
Hypothesis testing: For the following ten exercises, answer each question.

a. State the null and alternate hypothesis.
b. State the *p*-value.
c. State alpha.
d. What is your decision?
e. Write a conclusion.
f. Answer any other questions asked in the problem.

107. 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 a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

108. 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?

109. 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?

110. The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level in Kentucky? Are the results applicable across the country? Why?

111. 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 $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

112. 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.

113. 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?

114. 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?

115. 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 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 the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

116. 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. The class took 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 the student academic group's claim correct?

REFERENCES

8.1 Null and Alternative Hypotheses

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

8.5 Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA

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 June 27, 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).

SOLUTIONS

1 The random variable is the mean Internet speed in Megabits per second.

3 The random variable is the mean number of children an American family has.

5 The random variable is the proportion of people picked at random in Times Square visiting the city.

7

a. H_0: $p = 0.42$

b. H_a: $p < 0.42$

9

a. H_0: $\mu = 15$

b. H_a: $\mu \neq 15$

11 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.

13 α = the probability that you think the bag cannot withstand -15 degrees F, when in fact it can β = the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

15 Type I: The procedure will go well, but the doctors think it will not. Type II: The procedure will not go well, but the doctors think it will.

17 0.019

19 0.998

21 A normal distribution or a Student's *t*-distribution

23 Use a Student's *t*-distribution

25 a normal distribution for a single population mean

27 It must be approximately normally distributed.

29 They must both be greater than five.

31 binomial distribution

33 The outcome of winning is very unlikely.

35 H_0: $\mu >= 73$
H_a: $\mu < 73$
The *p*-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5% level. The data do support the claim.

37 The shaded region shows a low *p*-value.

39 Do not reject H_0.

41 means

43 the mean time spent in jail for 26 first time convicted burglars

45

a. 3

b. 1.5

c. 1.8

d. 26

47 $\bar{X} \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

49 This is a left-tailed test.

51 This is a two-tailed test.

53

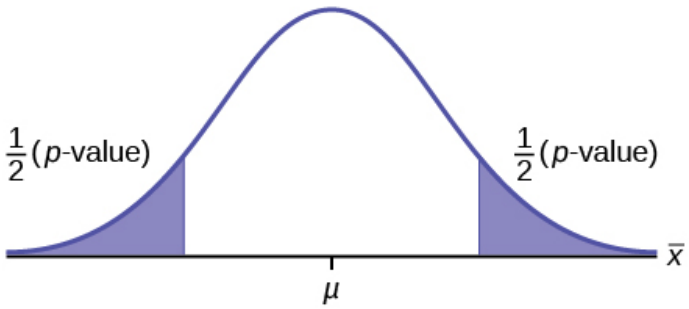


Figure 8.22

55 a right-tailed test

57 a left-tailed test

59 This is a left-tailed test.

61 This is a two-tailed test.

62

a. H_0: $\mu = 34$; H_a: $\mu \neq 34$

b. H_0: $p \leq 0.60$; H_a: $p > 0.60$

c. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$

d. H_0: $p = 0.29$; H_a: $p \neq 0.29$

e. H_0: $p = 0.05$; H_a: $p < 0.05$

f. H_0: $\mu \leq 10$; H_a: $\mu > 10$

g. H_0: $p = 0.50$; H_a: $p \neq 0.50$

h. H_0: $\mu = 6$; H_a: $\mu \neq 6$

i. H_0: $p \geq 0.11$; H_a: $p < 0.11$

j. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

64 c

66

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.

68 b

70 d

72 d

74

a. H_0: $\mu \geq 50{,}000$

b. H_a: $\mu < 50{,}000$

c. Let $\bar{X}$ = the average lifespan of a brand of tires.

d. normal distribution

e. $z = -2.315$

f. p-value = 0.0103

g. Check student's solution.

h.
   i. alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

i. (43,537, 49,463)

76

a. H_0: $\mu = \$1.00$

b. H_a: $\mu \neq \$1.00$

c. Let $\bar{X}$ = the average cost of a daily newspaper.

d. normal distribution

e. $z = -0.866$

f. p-value = 0.3865

g. Check student's solution.

h.
   i. Alpha: 0.01
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is \$1. The mean cost could be \$1.

i. (\$0.84, \$1.06)

78

a. H_0: $\mu = 10$

b. H_a: $\mu \neq 10$

c. Let $\bar{X}$ the mean number of sick days an employee takes per year.

d. Student's t-distribution

e. $t = -1.12$

f. p-value = 0.300

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.

iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of sick days is not ten.

i. (4.9443, 11.806)

80

a. H_0: $p \geq 0.6$

b. H_a: $p < 0.6$

c. Let P' = the proportion of students who feel more enriched as a result of taking Elementary Statistics.

d. normal for a single proportion

e. 1.12

f. p-value = 0.1308

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The p-value is greater than 0.05.

iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence Interval: (0.409, 0.654)
The "plus-4s" confidence interval is (0.411, 0.648)

82

a. H_0: $\mu = 4$

b. H_a: $\mu \neq 4$

c. Let $\bar{X}$ the average I.Q. of a set of brown trout.

d. two-tailed Student's t-test

e. $t = 1.95$

f. p-value = 0.076

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The p-value is greater than 0.05

iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. (3.8865,5.9468)

84

a. H_0: $p \geq 0.13$

b. H_a: $p < 0.13$

c. Let P' = the proportion of Americans who have seen or sensed angels

d. normal for a single proportion

e. –2.688

f. p-value = 0.0036

g. Check student's solution.

h. i. alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The *p*-value is less than 0.05.

iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

i. (0, 0.0623).
The"plus-4s" confidence interval is (0.0022, 0.0978)

86

a. H_0: $\mu \geq 129$

b. H_a: $\mu < 129$

c. Let $\bar{X}$ = the average time in seconds that Terri finishes Lap 4.

d. Student's *t*-distribution

e. $t = 1.209$

f. 0.8792

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The *p*-value is greater than 0.05.

iv. Conclusion: There is insufficient evidence to conclude that Terri's mean lap time is less than 129 seconds.

i. (128.63, 130.37)

88

a. H_0: $p = 0.60$

b. H_a: $p < 0.60$

c. Let P' = the proportion of family members who shed tears at a reunion.

d. normal for a single proportion

e. –1.71

f. 0.0438

g. Check student's solution.

h. i. alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: *p*-value < alpha

iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the *p*-value and alpha are quite close, so other tests should be done.

i. We are 95% confident that between 38.29% and 61.71% of family members will shed tears at a family reunion. (0.3829, 0.6171). The"plus-4s" confidence interval (see chapter 8) is (0.3861, 0.6139)

Note that here the "large-sample" 1 – PropZTest provides the approximate *p*-value of 0.0438. Whenever a *p*-value based on a normal approximation is close to the level of significance, the exact *p*-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

90

a. H_0: $\mu \geq 22$

b. H_a: $\mu < 22$

c. Let $\bar{X}$ = the mean number of bubbles per blow.

d. Student's *t*-distribution

e. –2.667

f. *p*-value = 0.00486

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The *p*-value is less than 0.05.

iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.

i. (18.501, 21.499)

92

a. H_0: $\mu \leq 1$

b. H_a: $\mu > 1$

c. Let $\bar{X}$ = the mean cost in dollars of macaroni and cheese in a certain town.

d. Student's *t*-distribution

e. $t = 0.340$

f. *p*-value = 0.36756

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The *p*-value is greater than 0.05

iv. Conclusion: The mean cost could be $1, or less. At the 5% significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.

i. (0.8291, 1.241)

94

a. H_0: $p = 0.01$

b. H_a: $p > 0.01$

c. Let P' = the proportion of errors generated

d. Normal for a single proportion

e. 2.13

f. 0.0165

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis

iii. Reason for decision: The *p*-value is less than 0.05.

iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.

i. Confidence interval: (0, 0.094).
The"plus-4s" confidence interval is (0.004, 0.144).

96

a. H_0: $p = 0.50$
b. H_a: $p < 0.50$
c. Let P' = the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. p-value = 0.0448
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: The p-value is less than 0.05. (However, they are very close.)
   iv. Conclusion: There is sufficient evidence to support the claim that less than 50% of his friends have pierced ears.
i. Confidence Interval: (0.245, 0.515): The "plus-4s" confidence interval is (0.259, 0.519).

98

a. H_0: $p = 0.40$
b. H_a: $p < 0.40$
c. Let P' = the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. p-value = 0.1563
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.
i. Confidence Interval: (0.3241, 0.4240): The "plus-4s" confidence interval is (0.3257, 0.4250).

100

a. H_0: $p = 0.14$
b. H_a: $p < 0.14$
c. Let P' = the proportion of NYC residents that smoke.
d. normal for a single proportion
e. –0.2756
f. p-value = 0.3914
g. Check student's solution.
h. i. alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.
i. Confidence Interval: (0.0502, 0.2070): The "plus-4s" confidence interval (see chapter 8) is (0.0676, 0.2297).

102

a. H_0: $\mu = 69{,}110$

b. H_a: $\mu > 69{,}110$

c. Let $\bar{X}$ = the mean salary in dollars for California registered nurses.

d. Student's *t*-distribution

e. $t = 1.719$

f. *p*-value: 0.0466

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

i. ($68,757, $73,485)

104 c

106 c

108

a. H_0: $p = 0.488$ H_a: $p \neq 0.488$

b. *p*-value = 0.0114

c. alpha = 0.05

d. Reject the null hypothesis.

e. At the 5% level of significance, there is enough evidence to conclude that 48.8% of families own stocks.

f. The survey does not appear to be accurate.

110

a. H_0: $p = 0.517$ H_a: $p \neq 0.517$

b. *p*-value = 0.9203.

c. alpha = 0.05.

d. Do not reject the null hypothesis.

e. 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.

f. 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. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

112

a. H_0: $\mu \geq 11.52$ H_a: $\mu < 11.52$

b. *p*-value = 0.000002 which is almost 0.

c. alpha = 0.05.

d. Reject the null hypothesis.

e. 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.

f. We would make the same conclusion if alpha was 1% because the *p*-value is almost 0.

114

a. H_0: $\mu \leq 5.8$ H_a: $\mu > 5.8$

b. p-value = 0.9987

c. alpha = 0.05

d. Do not reject the null hypothesis.

e. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

116

a. H_0: $\mu \geq 150$ H_a: $\mu < 150$

b. p-value = 0.0622

c. alpha = 0.01

d. Do not reject the null hypothesis.

e. 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.

f. The student academic group's claim appears to be correct.

9 | LINEAR REGRESSION AND CORRELATION

Figure 9.1 Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

Introduction

Chapter Objectives

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

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

9.1 | Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + \text{bx}$$

where a and b are constant numbers.

The variable ***x* is the independent variable, and *y* is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 9.1

The following examples are linear equations.

$$y = 3 + 2\text{x}$$
$$y = -0.01 + 1.2\text{x}$$

Try It Σ

9.1 Is the following an example of a linear equation?

$y = -0.125 - 3.5x$

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example 9.2

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

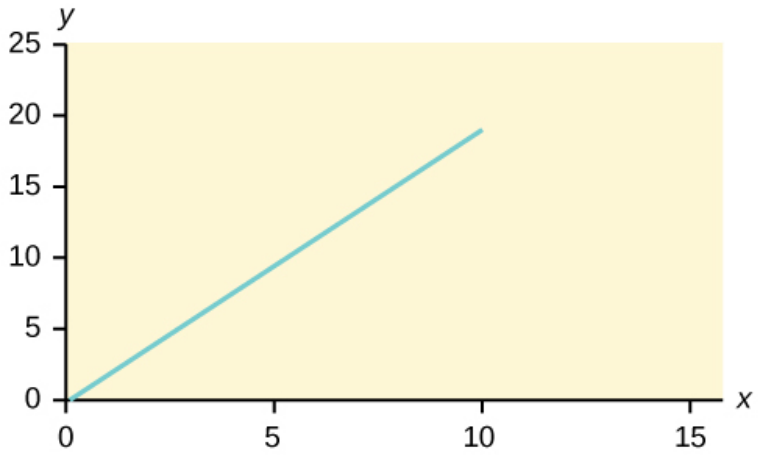


Figure 9.2

Try It

9.2 Is the following an example of a linear equation? Why or why not?

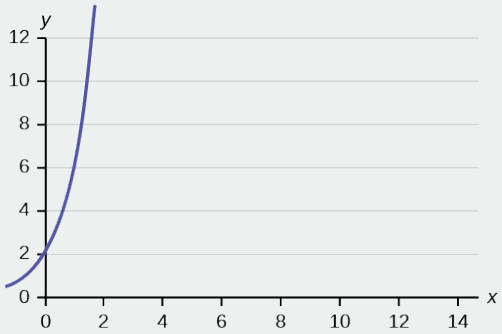


Figure 9.3

Example 9.3

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Solution 9.3

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

9.3 Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

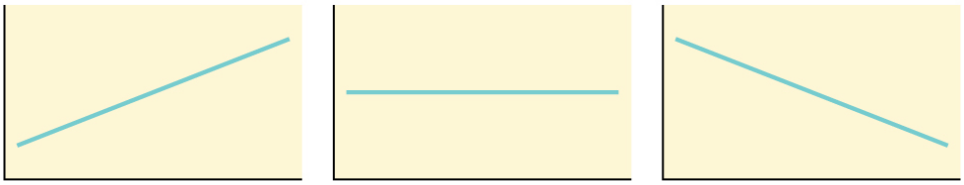


Figure 9.4 Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 9.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution 9.4

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Try It

9.4 Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

9.2 | Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

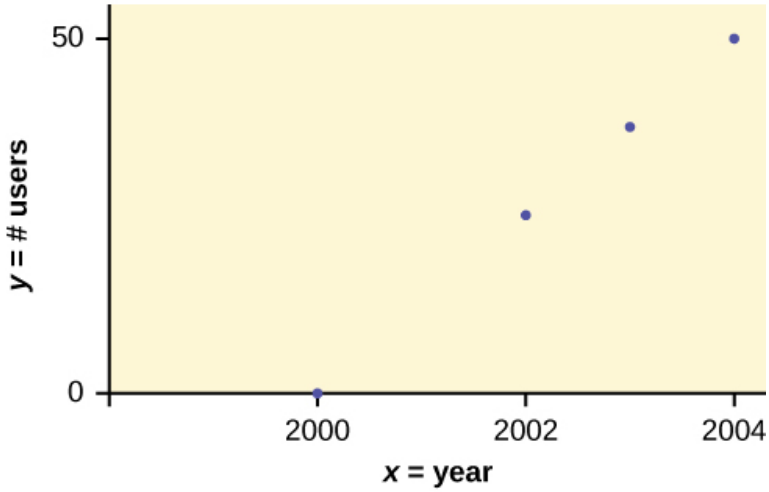
Example 9.5

In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let *x* = the year and let *y* = the number of m-commerce users, in millions.

x (year)	*y* (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Table 9.1

(a) Table showing the number of m-commerce users (in millions) by year.



(b) Scatter plot showing the number of m-commerce users (in millions) by year.

Figure 9.5

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

To create a scatter plot:

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. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
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 the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Try It Σ

9.5 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

Table 9.2

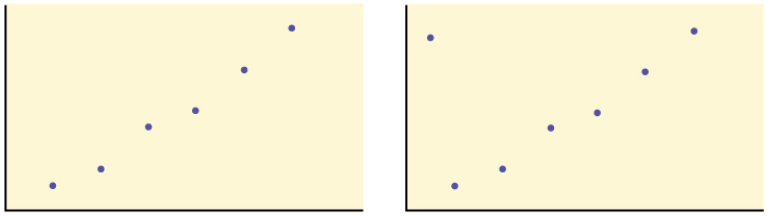
Construct a scatter plot and state if what Amelia thinks appears to be true.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

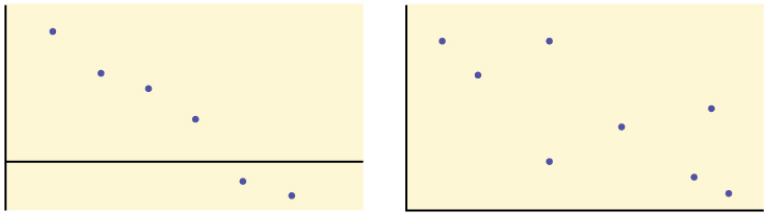
You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



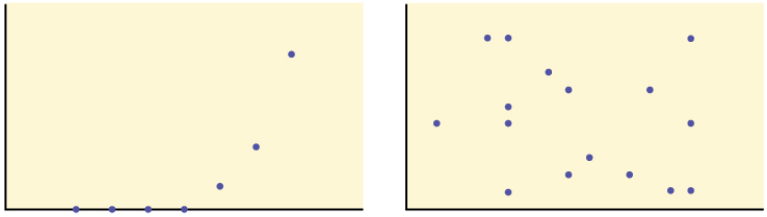
(a) Positive linear pattern (strong) (b) Linear pattern w/ one deviation

Figure 9.6



(a) Negative linear pattern (strong) (b) Negative linear pattern (weak)

Figure 9.7



(a) Exponential growth pattern (b) No pattern

Figure 9.8

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x

CHAPTER REVIEW

9.1 Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\boldsymbol{y = mx + b}$, where $\boldsymbol{m}$ and $\boldsymbol{b}$ are constants, $\boldsymbol{x}$ is the independent variable, $\boldsymbol{y}$ is the dependent variable. In a statistical context, a linear equation is written in the form $\boldsymbol{y = a + bx}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the $\boldsymbol{x}$ variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The ***y*-intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

9.2 Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

FORMULA REVIEW

9.1 Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

$y = a + bx$ where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

PRACTICE

9.1 Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

1. What are the dependent and independent variables?

2. Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

3. Graph the equation from **Exercise 9.2**.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

4. Find the equation that expresses the total fee in terms of the number of days the payment is late.

5. Graph the equation from **Exercise 9.4**.

6. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

7. Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

c. $y - x = 8x^2$

d. $4y = 8$

8. Does the graph show a linear equation? Why or why not?

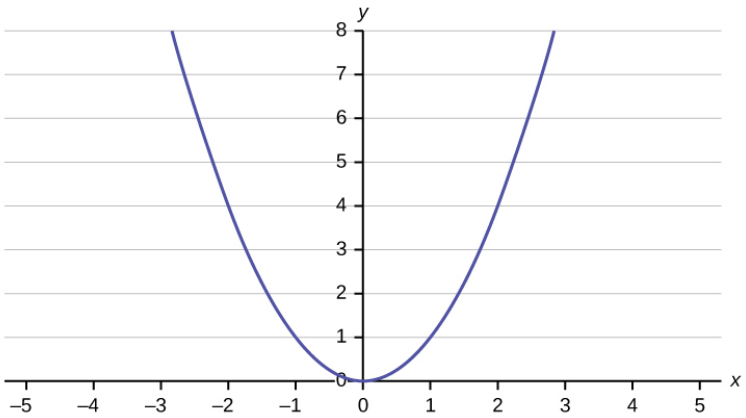


Figure 9.9

Table 9.3 contains real data for the first two decades of AIDS reporting.

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347

2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 9.3 Adults and Adolescents only, United States

9. Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

10. What are the independent and dependent variables?

11. What is the y-intercept and what is the slope? Interpret them using complete sentences.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

12. What are the independent and dependent variables?

13. How many pounds of soil does the shoreline lose in a year?

14. What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

15. What are the slope and y-intercept? Interpret their meaning.

16. If you owned this stock, would you want a positive or negative slope? Why?

9.2 Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

17. Does the scatter plot appear linear? Strong or weak? Positive or negative?

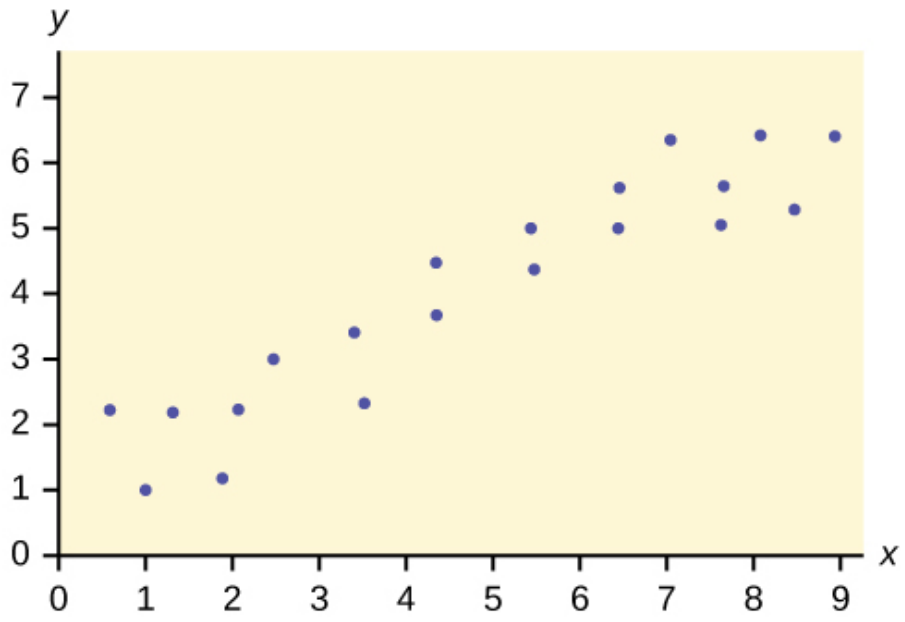


Figure 9.10

18. Does the scatter plot appear linear? Strong or weak? Positive or negative?

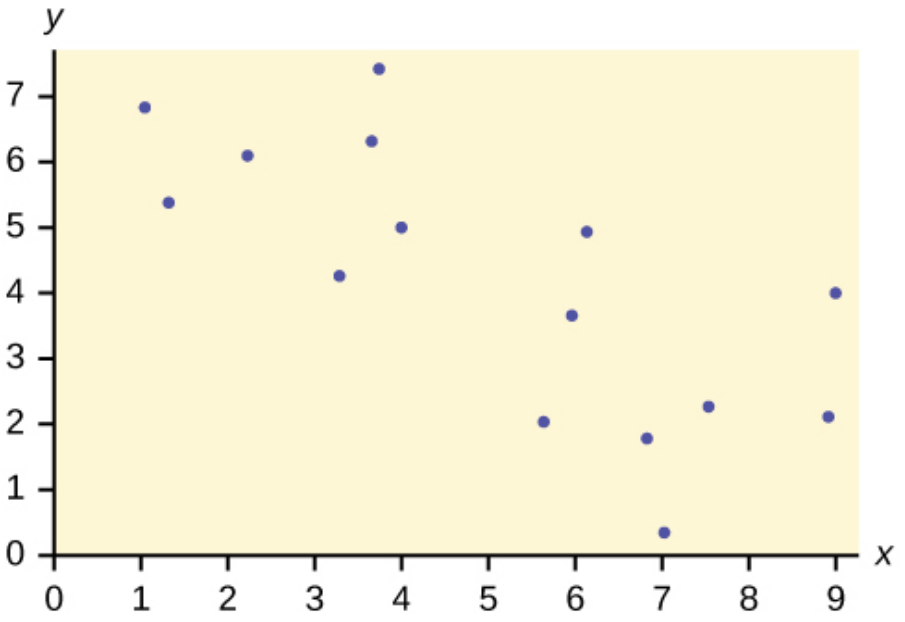


Figure 9.11

19. Does the scatter plot appear linear? Strong or weak? Positive or negative?

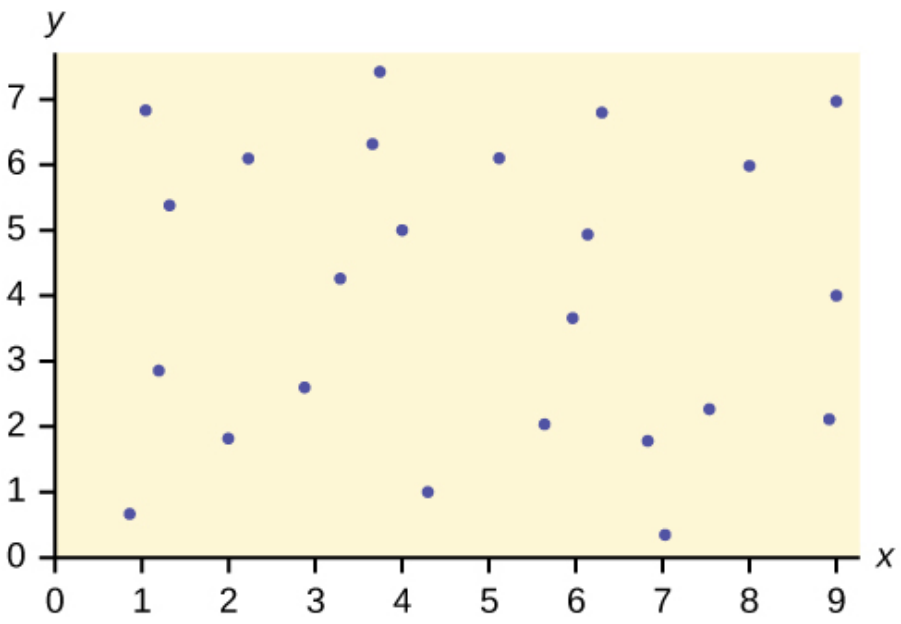


Figure 9.12

HOMEWORK

9.1 Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

20. For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

21. Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

Table 9.4

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

9.2 Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

22. The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. **Table 9.5** shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Table 9.5

23. The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Table 9.6

24. 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. Construct a scatter plot of the data.

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

Table 9.7

25. If the level of significance is 0.05 and the *p*-value is 0.06, what conclusion can you draw?

26. If there are 15 data points in a set of data, what is the number of degree of freedom?

REFERENCES

9.1 Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

SOLUTIONS

1 dependent variable: fee amount; independent variable: time

3

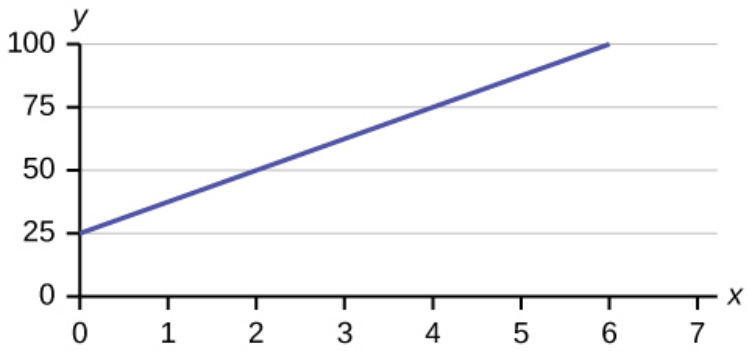


Figure 9.13

5

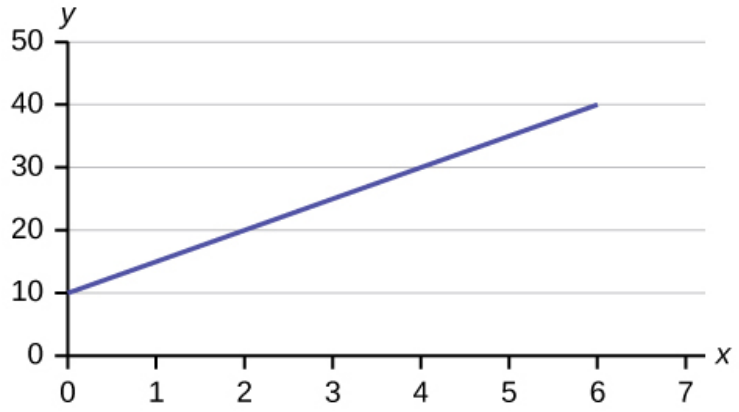


Figure 9.14

7 $y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

9 The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

11 The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

13 12,000 pounds of soil

15 The slope is –1.5 ($b = –1.5$). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

17 The data appear to be linear with a strong, positive correlation.

19 The data appear to have no correlation.

20

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

22 Check student's solution.

24 For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

26 13

APPENDIX A: STATISTICAL TABLES

Please see the following page.

APPENDIX B: MATHEMATICAL PHRASES, SYMBOLS, AND FORMULAS

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of *X* is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of *X* is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer *X* than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Table B1

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\ }$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	*IQR*	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	$\bar{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	*s*	s	sample standard deviation
Descriptive Statistics	s^2	s squared	sample variance
Descriptive Statistics	σ	sigma	population standard deviation
Descriptive Statistics	σ^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \cup B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \cap B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	*A′*	A-prime, complement of A	complement of A, not A
Probability Topics	*P*(*A*′)	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	*PDF*	prob. density function	same
Discrete Random Variables	*X*	X	the random variable X
Discrete Random Variables	*X* ~	the distribution of X	same
Discrete Random Variables	$\geq$	greater than or equal to	same
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	=	equal to	same
Discrete Random Variables	≠	not equal to	same

Table B2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	*f*(*x*)	*f* of *x*	function of *x*
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	*U*	uniform distribution	same
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	*f*(*x*) =	*f* of *x* equals	same
Continuous Random Variables	*m*	*m*	decay rate (for exp. dist.)
The Normal Distribution	*N*	normal distribution	same
The Normal Distribution	*z*	*z*-score	same
The Normal Distribution	*Z*	standard normal dist.	same
The Central Limit Theorem	$\bar{X}$	*X*-bar	the random variable *X*-bar
The Central Limit Theorem	$\mu_{\bar{x}}$	mean of *X*-bars	the average of *X*-bars
The Central Limit Theorem	$\sigma_{\bar{x}}$	standard deviation of *X*-bars	same
Confidence Intervals	*CL*	confidence level	same
Confidence Intervals	*CI*	confidence interval	same
Confidence Intervals	*EBM*	error bound for a mean	same
Confidence Intervals	*EBP*	error bound for a proportion	same
Confidence Intervals	*t*	Student's *t*-distribution	same
Confidence Intervals	*df*	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *α*/2 area in right tail	same
Confidence Intervals	p'	*p*-prime	sample proportion of success
Confidence Intervals	q'	*q*-prime	sample proportion of failure
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\bar{X}1 - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions

Table B2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	*y* equals a plus *b*-*x*	equation of a straight line
Linear Regression and Correlation	$\hat{y}$	*y*-hat	estimated value of *y*
Linear Regression and Correlation	r	sample correlation coefficient	same
Linear Regression and Correlation	ε	error term for a regression line	same
Linear Regression and Correlation	*SSE*	Sum of Squared Errors	same
F-Distribution and ANOVA	*F*	*F*-ratio	*F*-ratio

Table B2 Symbols and their Meanings

Formulas

Symbols You Must Know		
Population		**Sample**
N	Size	n
μ	Mean	$\bar{x}$
σ^2	Variance	s^2
σ	Standard Deviation	s
p	Proportion	p'
Single Data Set Formulae		
Population		**Sample**
$\mu = E(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i)$	Arithmetic Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$
	Geometric Mean	$\tilde{x} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$
$Q_3 = \frac{3(n+1)}{4}$, $Q_1 = \frac{(n+1)}{4}$	Inter-Quartile Range $IQR = Q_3 - Q_1$	$Q_3 = \frac{3(n+1)}{4}$, $Q_1 = \frac{(n+1)}{4}$

Table B3

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	Variance	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Single Data Set Formulae		
Population		**Sample**
$\mu = E(x) = \frac{1}{N}\sum_{i=1}^{N}(m_i * f_i)$	Arithmetic Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(m_i * f_i)$
	Geometric Mean	$\tilde{x} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(m_i - \mu)^2 * f_i$	Variance	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{x})^2 * f_i$
$CV = \frac{\sigma}{\mu} * 100$	Coefficient of Variation	$CV = \frac{s}{\bar{x}} * 100$

Table B3

Basic Probability Rules			
$P(A \cap B) = P(A\|B) * P(B)$			**Multiplication Rule**
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$			**Addition Rule**
$P(A \cap B) = P(A) * P(B)$ or $P(A\|B) = P(A)$			**Independence Test**
Hypergeometric Distribution Formulae			
$nCx = \binom{n}{x} = \frac{n!}{x!(n-x)!}$		**Combinatorial Equation**	
$P(x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$		**Probability Equation**	
$E(X) = \mu = np$		**Mean**	
$\sigma^2 = \left(\frac{N-n}{N-1}\right)np(q)$		**Variance**	
Binomial Distribution Formulae			
$P(x) = \frac{n!}{x!(n-x)!}p^x(q)^{n-x}$		**Probability Density Function**	
$E(X) = \mu = np$		**Arithmetic Mean**	
$\sigma^2 = np\left(q\right)$		**Variance**	
Geometric Distribution Formulae			
$P(X = x) = (1-p)^{x-1}(p)$	**Probability when x is the first success.**	**Probability when x is the number of failures before first success**	$P(X = x) = (1-p)^x(p)$

Table B4

$\mu = \frac{1}{p}$	**Mean**	**Mean**	$\mu = \frac{1-p}{p}$
$\sigma^2 = \frac{(1-p)}{p^2}$	**Variance**	**Variance**	$\sigma^2 = \frac{(1-p)}{p^2}$
Poisson Distribution Formulae			
$P(x) = \frac{e^{-\mu}\mu^x}{x!}$	**Probability Equation**		
$E(X) = \mu$	**Mean**		
$\sigma^2 = \mu$	**Variance**		
Uniform Distribution Formulae			
$f(x) = \frac{1}{b-a}$ for $a \le x \le b$	**PDF**		
$E(X) = \mu = \frac{a+b}{2}$	**Mean**		
$\sigma^2 = \frac{(b-a)^2}{12}$	**Variance**		
Exponential Distribution Formulae			
$P(X \le x) = 1 - e^{-mx}$	**Cumulative Probability**		
$E(X) = \mu = \frac{1}{m}$ or $m = \frac{1}{\mu}$	**Mean and Decay Factor**		
$\sigma^2 = \frac{1}{m^2} = \mu^2$	**Variance**		

Table B4

The following page of formulae requires the use of the " Z ", " t ", and " χ^2 " tables.	
$Z = \frac{x-\mu}{\sigma}$	**Z-transformation for Normal Distribution**
$Z = \frac{x-np'}{\sqrt{np'(q')}}$	**Normal Approximation to the Binomial**
Probability (ignores subscripts) **Hypothesis Testing**	**Confidence Intervals** [bracketed symbols equal margin of error] (subscripts denote locations on respective distribution tables)
$Z_c = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	*Interval for the population mean when sigma is known* $\bar{x} \pm \left[Z_{(\alpha/2)} \frac{\sigma}{\sqrt{n}} \right]$
$Z_c = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	*Interval for the population mean when sigma is unknown but* $n > 30$ $\bar{x} \pm \left[Z_{(\alpha/2)} \frac{s}{\sqrt{n}} \right]$
$t_c = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	*Interval for the population mean when sigma is unknown but* $n < 30$ $\bar{x} \pm \left[t_{(n-1),(\alpha/2)} \frac{s}{\sqrt{n}} \right]$

Table B5

$Z_c = \frac{p' - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$	*Interval for the population proportion* $p' \pm \left[Z_{(\alpha/2)} \sqrt{\frac{p'q'}{n}} \right]$
$t_c = \frac{\bar{d} - \delta_0}{s_d}$	*Interval for difference between two means with matched pairs* $\bar{d} \pm \left[t_{(n-1),(\alpha/2)} \frac{s_d}{\sqrt{n}} \right]$ *where* s_d *is the deviation of the differences*
$Z_c = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	*Interval for difference between two means when sigmas are known* $(\bar{x}_1 - \bar{x}_2) \pm \left[Z_{(\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$
$t_c = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \delta_0}{\sqrt{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)}}$	*Interval for difference between two means with equal variances when sigmas are unknown* $(\bar{x}_1 - \bar{x}_2) \pm \left[t_{df,(\alpha/2)} \sqrt{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)} \right]$ *where* $df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right) + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)}$
$Z_c = \frac{(p'_1 - p'_2) - \delta_0}{\sqrt{\frac{p'_1(q'_1)}{n_1} + \frac{p'_2(q'_2)}{n_2}}}$	*Interval for difference between two population proportions* $(p'_1 - p'_2) \pm \left[Z_{(\alpha/2)} \sqrt{\frac{p'_1(q'_1)}{n_1} + \frac{p'_2(q'_2)}{n_2}} \right]$
$\chi_c^2 = \frac{(n-1)s^2}{\sigma_0^2}$	*Tests for GOF, Independence, and Homogeneity* $\chi_c^2 = \Sigma \frac{(O-E)^2}{E}$ *where O = observed values and E = expected values*

The Next 3 Formulae are for Determining Sample Size with Confidence Intervals
(note: E represents the margin of error)

$n = \frac{Z^2_{\left(\frac{a}{2}\right)} \sigma^2}{E^2}$	$n = \frac{Z^2_{\left(\frac{a}{2}\right)} (0.25)}{E^2}$	$n = \frac{Z^2_{\left(\frac{a}{2}\right)} [p'(q')]}{E^2}$
Use when sigma is known	*Use when* $p\prime$ *is unknown*	*Use when* $p\prime$ *is uknown*
$E = \bar{x} - \mu$	$E = p' - p$	$E = p' - p$

Table B5

Simple Linear Regression Formulae for $y = a + b(x)$	
$r = \frac{\Sigma \left[\left(x - \bar{x}\right)\left(y - \bar{y}\right)\right]}{\sqrt{\Sigma \left(x - \bar{x}\right)^2 * \Sigma \left(y - \bar{y}\right)^2}} = \frac{S_{xy}}{S_x S_y} = \sqrt{\frac{SSR}{SST}}$	**Correlation Coefficient**
$b = \frac{\Sigma \left[\left(x - \bar{x}\right)\left(y - \bar{y}\right)\right]}{\Sigma \left(x - \bar{x}\right)^2} = \frac{S_{xy}}{SS_x} = r_{y,x}\left(\frac{s_y}{s_x}\right)$	**Coefficient *b* (slope)**

Table B6

$a = \bar{y} - b(\bar{x})$	**y-intercept**
$s_e^2 = \frac{\Sigma (y_i - \hat{y}_i)^2}{n-k} = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$	**Estimate of the Error Variance**
$S_b = \frac{s_e^2}{\sqrt{(x_i - \bar{x})^2}} = \frac{s_e^2}{(n-1)s_x^2}$	**Standard Error for Coefficient *b***
$t_c = \frac{b - \beta_0}{s_b}$	**Hypothesis Test for Coefficient *β***
$b \pm [t_{n-2,\ \alpha/\ \ 2} S_b]$	**Interval for Coefficient *β***
$\hat{y} \pm \left[t_{\alpha/\ \ 2} * s_e \left(\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_x}} \right) \right]$	**Interval for Expected alue of *y***
$\hat{y} \pm \left[t_{\alpha/\ \ 2} * s_e \left(\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_x}} \right) \right]$	**Prediction Interval for an Individual *y***
ANOVA Formulae	
$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	**Sum of Squares Regression**
$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$	**Sum of Squares Error**
$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	**Sum of Squares Total**
$R^2 = \frac{SSR}{SST}$	**Coefficient of Determination**

Table B6

The following is the breakdown of a one-way ANOVA table for linear regression.				
Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Squares*	$F - Ratio$
Regression	SSR	1 *or* $k-1$	$MSR = \frac{SSR}{df_R}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n-k$	$MSE = \frac{SSE}{df_E}$	
Total	SST	$n-1$		

Table B7

INDEX

ATTRIBUTIONS

Collection: **Business Statistics I -- MGMT 2262 -- Mt Royal University -- Version 2016 Revision A**
Edited by: Claude Laflamme
URL: **http://legacy.cnx.org/content/col11990/1.5/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Business Statistics -- BSTA 200 -- Humber College -- Version 2016RevA -- DRAFT 2016-04-04 <**http://legacy.cnx.org/content/col11969/1.5**> arranged by Claude Laflamme.

Module: **Preface-- MGMT 2262 -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62372/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Preface <**http://legacy.cnx.org/content/m53808/1.10**> by Alexander Holmes.

Module: **Introduction -- Sampling and Data -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62095/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m54035/1.2**> by Alexander Holmes.

Module: **Definitions of Statistics, Probability, and Key Terms -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62097/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Definitions of Statistics, Probability, and Key Terms <**http://legacy.cnx.org/content/m54033/1.2**> by Alexander Holmes.

Module: **Data, Sampling, and Variation -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62319/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Data, Sampling, and Variation in Data and Sampling <**http://legacy.cnx.org/content/m54036/1.4**> by Alexander Holmes.

Module: **Experimental Design and Ethics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62317/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Experimental Design and Ethics <**http://legacy.cnx.org/content/m54056/1.3**> by Alexander Holmes.

Module: **Introduction -- Descriptive Statistics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62325/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m53786/1.1**> by Alexander Holmes.

Module: **Display Data -- Descriptive Statistics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62323/1.1/**

Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Display Data <**http://legacy.cnx.org/content/m53791/1.10**> by Alexander Holmes.

Module: **Box Plots -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m61599/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Box Plots <**http://legacy.cnx.org/content/m53800/1.1**> by Alexander Holmes.

Module: **Measures of the Location of the Data -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62326/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Measures of the Location of the Data <**http://legacy.cnx.org/content/m53795/1.4**> by Alexander Holmes.

Module: **Measures of the Center of the Data -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62320/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Measures of the Center of the Data <**http://legacy.cnx.org/content/m53802/1.5**> by Alexander Holmes.

Module: **Distribution -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62324/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Skewness and the Mean, Median, and Mode <**http://legacy.cnx.org/content/m53804/1.4**> by Alexander Holmes.

Module: **Measures of Variaton -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62327/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Measures of the Spread of the Data <**http://legacy.cnx.org/content/m53806/1.12**> by Alexander Holmes.

Module: **Introduction -- Probability Topics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62328/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m54216/1.1**> by Alexander Holmes.

Module: **Terminology -- Probability Topics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62337/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Terminology <**http://legacy.cnx.org/content/m54218/1.8**> by Alexander Holmes.

Module: **Independent and Mutually Exclusive Events -- Probaility Topics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62329/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**

Based on: Independent and Mutually Exclusive Events <**http://legacy.cnx.org/content/m54219/1.2**> by Alexander Holmes.

Module: **Two Basic Rules of Probability**
By: Alexander Holmes
URL: **http://legacy.cnx.org/content/m54220/1.7/**
Copyright: Alexander Holmes
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Two Basic Rules of Probability <**http://legacy.cnx.org/content/m46947/1.4**> by OpenStax.

Module: **Contingency Tables and Tree Diagrams -- Probability Topics -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62330/1.2/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Contingency Tables and Probability Trees <**http://legacy.cnx.org/content/m54259/1.7**> by Alexander Holmes.

Module: **Introduction -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62343/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m54262/1.11**> by Alexander Holmes.

Module: **Binomial Distribution -- Discrete Random Variables -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62344/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Binomial Distribution <**http://legacy.cnx.org/content/m54267/1.10**> by Alexander Holmes.

Module: **Introduction -- The Normal Distribution -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62345/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m54495/1.6**> by Alexander Holmes.

Module: **The Standard Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62346/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Standard Normal Distribution <**http://legacy.cnx.org/content/m54496/1.8**> by Alexander Holmes.

Module: **Using the Normal Distribution-- The Normal Distribution -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62347/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Using the Normal Distribution <**http://legacy.cnx.org/content/m54497/1.13**> by Alexander Holmes.

Module: **Introduction -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62358/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**

Based on: Introduction <**http://legacy.cnx.org/content/m54530/1.2**> by Alexander Holmes.

Module: **The Central Limit Theorem for Sample Means (Averages)-- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62359/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Central Limit Theorem for Sample Means (Averages) <**http://legacy.cnx.org/content/m54534/1.11**> by Alexander Holmes.

Module: **Using the Central Limit Theorem -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62348/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Using the Central Limit Theorem <**http://legacy.cnx.org/content/m54537/1.15**> by Alexander Holmes.

Module: **Central Limit Theorem (Pocket Change) -- The Central Limit Theorem -- Mt Royal University -- Version 2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62350/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Central Limit Theorem (Pocket Change) <**http://legacy.cnx.org/content/m50268/1.2**> by OpenStax.

Module: **Introduction -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62353/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m55460/1.6**> by Alexander Holmes.

Module: **A Single Population Mean using the Normal Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62362/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Single Population Mean using the Normal Distribution <**http://legacy.cnx.org/content/m47002/1.9**> by OpenStax.

Module: **A Single Population Mean using the Student t Distribution -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62365/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Single Population Mean using the Student t Distribution <**http://legacy.cnx.org/content/m55462/1.23**> by Alexander Holmes.

Module: **A Population Proportion -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62361/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Population Proportion <**http://legacy.cnx.org/content/m55463/1.16**> by Alexander Holmes.

Module: **Calculating the Sample Size n: Means and Proportions -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62360/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Calculating the Sample Size n: Continuous Random Variables and Binary Random Variables <**http://legacy.cnx.org/content/m56644/1.10**> by Alexander Holmes.

Module: **Confidence Interval (Home Costs) -- Confidence Intervals -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62281/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Confidence Interval (Home Costs) <**http://legacy.cnx.org/content/m50270/1.2**> by OpenStax.

Module: **Introduction -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62366/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m55603/1.3**> by Alexander Holmes.

Module: **Null and Alternative Hypotheses**
By: Alexander Holmes
URL: **http://legacy.cnx.org/content/m55606/1.5/**
Copyright: Alexander Holmes
License: **http://creativecommons.org/licenses/by/4.0/**

Module: **Outcomes and the Type I and Type II Errors -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62369/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Outcomes and the Type I and Type II Errors <**http://legacy.cnx.org/content/m55607/1.10**> by Alexander Holmes.

Module: **Distribution Needed for Hypothesis Testing -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62286/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Distribution Needed for Hypothesis Testing <**http://legacy.cnx.org/content/m55608/1.12**> by Alexander Holmes.

Module: **Rare Events, the Sample, Decision and Conclusion -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62282/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Rare Events, the Sample, Decision and Conclusion <**http://legacy.cnx.org/content/m47044/1.4**> by OpenStax.

Module: **Additional Information and Full Hypothesis Test Examples -- Hypothesis Testing with One Sample -- MtRoyal - Version2016RevA**
By: Claude Laflamme

URL: **http://legacy.cnx.org/content/m62288/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Additional Information and Full Hypothesis Test Examples <**http://legacy.cnx.org/content/m47019/1.8**> by OpenStax.

Module: **Introduction -- Linear Regression and Correlation -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62290/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m47098/1.4**> by OpenStax.

Module: **Linear Equations -- Linear Regression and Correlation -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62297/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Linear Equations <**http://legacy.cnx.org/content/m47100/1.2**> by OpenStax.

Module: **Scatter Plots -- Linear Regression and Correlation -- MtRoyal - Version2016RevA**
By: Claude Laflamme
URL: **http://legacy.cnx.org/content/m62293/1.1/**
Copyright: Claude Laflamme
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Scatter Plots <**http://legacy.cnx.org/content/m47109/1.2**> by OpenStax.

Module: **Statistical Tables**
By: Alexander Holmes
URL: **http://legacy.cnx.org/content/m56641/1.5/**
Copyright: Alexander Holmes
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Appendix H: Tables <**http://legacy.cnx.org/content/m47875/1.7**> by OpenStax.

Module: **Mathematical Phrases, Symbols, and Formulas**
By: Alexander Holmes
URL: **http://legacy.cnx.org/content/m56640/1.7/**
Copyright: Alexander Holmes
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Appendix F: Mathematical Phrases, Symbols, and Formulas <**http://legacy.cnx.org/content/m47891/1.4**> by OpenStax.

ABOUT CONNEXIONS

Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities. Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai.