

The Saylor Foundation's Flexbook: Jill Schmidtkofer's *Advanced Probability and Statistics*: partially adapted from David Lane's *Online Statistics Education: A Multimedia Course of Study* and CK-12: *Advanced Probability and Statistics*

Statistics

Jill Schmidlkofer
Larry Ottman, (LarryO)
Ellen Lawske, (EllenL)
CK12 Editor
Raja Almukkahal, (RajaA)
Larry Ottman, (LarryO)
Brenda Meery, (BrendaM)
Danielle DeLancey, (DanielleD)
Ellen Lawske, (EllenL)
Danielle DeLancey, (DanielleD)

Say Thanks to the Authors
Click <http://www.ck12.org/saythanks>
(No sign in required)



To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Copyright © 2013 CK-12 Foundation, www.ck12.org

The names “CK-12” and “CK12” and associated logos and the terms “**FlexBook®**” and “**FlexBook Platform®**” (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state, and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link <http://www.ck12.org/saythanks> (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution/Non-Commercial/Share Alike 3.0 Unported (CC BY-NC-SA) License (<http://creativecommons.org/licenses/by-nc-sa/3.0/>), as amended and updated by Creative Commons from time to time (the “CC License”), which is incorporated herein by this reference.

Complete terms can be found at <http://www.ck12.org/terms>.

Printed: July 14, 2013

flexbook
next generation textbooks



AUTHORS

Jill Schmidlkofer
Larry Ottman, (LarryO)
Ellen Lawsky, (EllenL)
CK12 Editor
Raja Almukkahal, (RajaA)
Larry Ottman, (LarryO)
Brenda Meery, (BrendaM)
Danielle DeLancey, (DanielleD)
Ellen Lawsky, (EllenL)
Danielle DeLancey, (DanielleD)

Contents

1	An Introduction to Analyzing Statistical Data	1
1.1	Statistical Terminology	2
1.2	An Overview of Data	8
1.3	Measures of Center	11
1.4	Measures of Spread	20
1.5	References	34
2	Visualizations of Data	35
2.1	Histograms and Frequency Distributions	36
2.2	Common Graphs and Data Plots	48
2.3	Box-and-Whisker Plots	63
3	An Introduction to Probability	81
3.1	Events, Sample Spaces, and Probability	82
3.2	Compound Events	87
3.3	The Complement of an Event	90
3.4	Conditional Probability	94
3.5	Addition and Multiplication Rules	99
3.6	Basic Counting Rules	109
3.7	References	117
4	Discrete Probability Distributions	118
4.1	Two Types of Random Variables	119
4.2	Probability Distribution for a Discrete Random Variable	122
4.3	Mean and Standard Deviation of Discrete Random Variables	125
4.4	Sums and Differences of Independent Random Variables	130
4.5	The Binomial Probability Distribution	138
4.6	References	146
5	Normal Distribution	147
5.1	The Standard Normal Probability Distribution	148
5.2	The Density Curve of the Normal Distribution	162
5.3	Applications of the Normal Distribution	175
6	Planning and Conducting an Experiment or Study	184
6.1	Surveys and Sampling	185
6.2	Experimental Design	195
7	Sampling Distributions and Estimations	207
7.1	Introduction to Sampling Distributions	208
7.2	The Central Limit Theorem	212
7.3	Confidence Intervals with z-values	220

7.4	References	229
8	Hypothesis Testing	230
8.1	Hypothesis Testing and the P-Value	231
8.2	Testing a Proportion Hypothesis	240
8.3	Testing a Mean Hypothesis	244
8.4	Student's t-Distribution	246
8.5	Testing a Hypothesis for Dependent and Independent Samples	255
8.6	References	267
9	Regression and Correlation	268
9.1	Scatterplots and Linear Correlation	269
9.2	Least-Squares Regression	278
9.3	Inferences about Regression	287
9.4	References	291
10	Chi-Square	292
10.1	The Goodness-of-Fit Test	293
10.2	Test of Independence	298

CHAPTER**1**

An Introduction to Analyzing Statistical Data

Chapter Outline

- 1.1 STATISTICAL TERMINOLOGY
 - 1.2 AN OVERVIEW OF DATA
 - 1.3 MEASURES OF CENTER
 - 1.4 MEASURES OF SPREAD
 - 1.5 REFERENCES
-

1.1 Statistical Terminology

Learning Objectives

- Distinguish between quantitative and categorical variables.
- Describe the difference between a population and a sample and be able to distinguish between a parameter and a statistic.

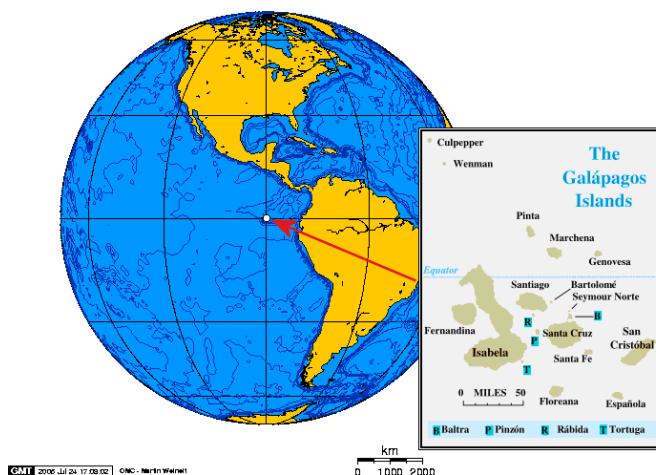
Introduction

In this lesson, you will be introduced to some basic vocabulary of statistics and learn how to distinguish between different types of variables. We will use the real-world example of information about the Giant Galapagos Tortoise.



The Galapagos Tortoises

The Galapagos Islands, off the coast of Ecuador in South America, are famous for the amazing diversity and uniqueness of life they possess. One of the most famous Galapagos residents is the Galapagos Giant Tortoise, which is found nowhere else on earth. Charles Darwin's visit to the islands in the 19th Century and his observations of the tortoises were extremely important in the development of his theory of evolution.



The tortoises lived on nine of the Galapagos Islands, and each island developed its own unique species of tortoise. In fact, on the largest island, there are four volcanoes, and each volcano has its own species. When first discovered, it was estimated that the tortoise population of the islands was around 250,000. Unfortunately, once European ships and settlers started arriving, those numbers began to plummet. Because the tortoises could survive for long periods of time without food or water, expeditions would stop at the islands and take the tortoises to sustain their crews with fresh meat and other supplies for the long voyages. Also, settlers brought in domesticated animals like goats and pigs that destroyed the tortoises' habitat. Today, two of the islands have lost their species, a third island has no remaining tortoises in the wild, and the total tortoise population is estimated to be around 15,000. The good news is there have been massive efforts to protect the tortoises. Extensive programs to eliminate the threats to their habitat, as well as breed and reintroduce populations into the wild, have shown some promise.

Approximate distribution of Giant Galapagos Tortoises in 2004, Estado Actual De Las Poblaciones de Tortugas Terrestres Gigantes en las Islas Galápagos, Marquez, Wiedenfeld, Snell, Fritts, MacFarland, Tapia, y Nanjoa, Scología Aplicada, Vol. 3, Num. 1,2, pp. 98-11.

TABLE 1.1:

Island or Volcano	Species	Climate Type	Shell Shape	Estimate of Total Population	Population Density (per km ²)	Number of Individuals Repatriated*
Wolf	becki	semi-arid	intermediate	1139	228	40
Darwin	microphyes	semi-arid	dome	818	205	0
Alcedo	vandenburghi	humid	dome	6,320	799	0
Sierra Negra	guntheri	humid	flat	694	122	286
Cerro Azul	vicina	humid	dome	2,574	155	357
Santa Cruz	nigrita	humid	dome	3,391	730	210
Española	hoodensis	arid	saddle	869	200	1,293
San Cristóbal	chathamensis	semi-arid	dome	1,824	559	55
Santiago	darwini	humid	intermediate	1,165	124	498
Pinzón	ephippium	arid	saddle	532	134	552
Pinta	abingdoni	arid	saddle	1	Does not apply	0

*Repatriation is the process of raising tortoises and releasing them into the wild when they are grown to avoid local predators that prey on the hatchlings.



Classifying Variables

Statisticians refer to an entire group that is being studied as a *population*. Each member of the population is called a *unit*. In this example, the population is all Galapagos Tortoises, and the units are the individual tortoises. It is not necessary for a population or the units to be living things, like tortoises or people. For example, an airline employee could be studying the population of jet planes in her company by studying individual planes.

A researcher studying Galapagos Tortoises would be interested in collecting information about different characteristics of the tortoises. Those characteristics are called *variables*. Each column of the previous figure contains a variable. In the first column, the tortoises are labeled according to the island (or volcano) where they live, and in the second column, by the scientific name for their species. When a characteristic can be neatly placed into well-defined groups, or categories, that do not depend on order, it is called a *categorical variable*, or *qualitative variable*.

The last three columns of the previous figure provide information in which the count, or quantity, of the characteristic is most important. For example, we are interested in the total number of each species of tortoise, or how many individuals there are per square kilometer. This type of variable is called a *numerical variable*, or *quantitative variable*. The figure below explains the remaining variables in the previous figure and labels them as categorical or numerical.

TABLE 1.2:

Variable	Explanation	Type
Climate Type	Many of the islands and volcanic habitats have three distinct climate types.	Categorical
Shell Shape	Over many years, the different species of tortoises have developed different shaped shells as an adaptation to assist them in eating vegetation that varies in height from island to island.	Categorical
Number of tagged individuals	Tortoises were captured and marked by scientists to study their health and assist in estimating the total population.	Numerical

TABLE 1.2: (continued)

Variable	Explanation	Type
Number of Individuals Repatriated	There are two tortoise breeding centers on the islands. Through these programs, many tortoises have been raised and then reintroduced into the wild.	Numerical

Population vs. Sample

We have already defined a population as the total group being studied. Most of the time, it is extremely difficult or very costly to collect all the information about a population. In the Galapagos, it would be very difficult and perhaps even destructive to search every square meter of the habitat to be sure that you counted every tortoise. In an example closer to home, it is very expensive to get accurate and complete information about all the residents of the United States to help effectively address the needs of a changing population. This is why a complete counting, or *census*, is only attempted every ten years. Because of these problems, it is common to use a smaller, representative group from the population, called a *sample*.

You may recall the tortoise data included a variable for the estimate of the population size. This number was found using a sample and is actually just an approximation of the true number of tortoises. If a researcher wanted to find an estimate for the population of a species of tortoises, she would go into the field and locate and mark a number of tortoises. She would then use statistical techniques that we will discuss later in this text to obtain an estimate for the total number of tortoises in the population. In statistics, we call the actual number of tortoises a *parameter*. Any number that describes the individuals in a sample (length, weight, age) is called a *statistic*. Each statistic is an estimate of a parameter, whose value may or may not be known.

Errors in Sampling

We have to accept that estimates derived from using a sample have a chance of being inaccurate. This cannot be avoided unless we measure the entire population. The researcher has to accept that there could be variations in the sample due to chance that lead to changes in the population estimate. A statistician would report the estimate of the parameter in two ways: as a *point estimate* (e.g., 915) and also as an *interval estimate*. For example, a statistician would report: "I am fairly confident that the true number of tortoises is actually between 561 and 1075." This range of values is the unavoidable result of using a sample, and not due to some mistake that was made in the process of collecting and analyzing the sample. The difference between the true parameter and the statistic obtained by sampling is called *sampling error*. It is also possible that the researcher made mistakes in her sampling methods in a way that led to a sample that does not accurately represent the true population. For example, she could have picked an area to search for tortoises where a large number tend to congregate (near a food or water source, perhaps). If this sample were used to estimate the number of tortoises in all locations, it may lead to a population estimate that is too high. This type of systematic error in sampling is called *bias*. Statisticians go to great lengths to avoid the many potential sources of bias. We will investigate this in more detail in a later chapter.

Lesson Summary

In statistics, the total group being studied is called the **population**. The individuals (people, animals, or things) in the population are called **units**. The characteristics of those individuals of interest to us are called **variables**. Those variables are of two types: **numerical**, or quantitative, and **categorical**, or qualitative.

Because of the difficulties of obtaining information about all units in a population, it is common to use a small, representative subset of the population, called a **sample**. An actual value of a population variable (for example,

number of tortoises, average weight of all tortoises, etc.) is called a **parameter**. An estimate of a parameter derived from a sample is called a **statistic**.

Whenever a sample is used instead of the entire population, we have to accept that our results are merely estimates, and therefore, have some chance of being incorrect. This is called **sampling error**.

Points to Consider

- How do we summarize, display, and compare categorical and numerical data differently?
- What are the best ways to display categorical and numerical data?
- Is it possible for a variable to be considered both categorical and numerical?
- How can you compare the effects of one categorical variable on another or one quantitative variable on another?

Review Questions

1. In each of the following situations, identify the **population, the units, and each variable that is measured, and tell if each variable is categorical or quantitative**.
 - a. A quality control worker with Sweet-Tooth Candy weighs every 100th candy bar to make sure it is very close to the published weight.
 - b. Doris decides to clean her sock drawer out and sorts her socks into piles by color.
 - c. A researcher is studying the effect of a new drug treatment for diabetes patients. She performs an experiment on 200 randomly chosen individuals with type II diabetes. Because she believes that men and women may respond differently, she records each person's gender, as well as the person's change in blood sugar level after taking the drug for a month.
2. A school is studying its students' test scores by grade. Explain how the characteristic 'grade' could be considered either a categorical or a numerical variable.
3. A school administrator wants to determine the number of seniors who drive to school at least 3 days per week. Out of the 420 seniors, she randomly chooses 30 of them, and she asks each if they drive to school at least 3 days per week. A total of 18 of them, which is 60%, respond yes.
 - a. Identify the population
 - b. Identify the sample.
 - c. Is the value 60% a statistic or a parameter? Defend your answer.
 - d. Is the administrator certain that 60% of all seniors drive to school at least 3 days per week? Why or why not?

Answers to Review Questions: 1.a. Population: all candy bars produced. Units: an individual candy bar. Variable: the weight of the candy bar, which is numerical (quantitative).

1.b. Population: all socks in the drawer. Units: each individual sock. Variable: Color of each sock, which is categorical (qualitative).

1.c. Population: All people with Type II diabetes. Units: each person with diabetes. Variable: change in blood sugar level, which is numerical (quantitative).

2. The listing of A, B, C, D, or F is categorical (qualitative). The listing of numerical scores is numerical (quantitative).

3.a. All seniors at the school. 3.b. The 30 randomly-chosen seniors. 3.c. Statistic; it is an estimate of the

proportion of all seniors who drive, but it is based on the 30 students in the sample. 3.d. The administrator is not certain. The value 60% is subject to sampling error.

1.2 An Overview of Data

Learning Objective

- Given a type of measurement, identify the correct level of measurement: nominal, ordinal, interval, or ratio.

Introduction

This lesson is an overview of the basic considerations involved with collecting and analyzing data.

Levels of Measurement

In the first lesson, you learned about the different types of variables that statisticians use to describe the characteristics of a population. Some researchers and social scientists use a more detailed distinction, called the *levels of measurement*, when examining the information that is collected for a variable. This widely accepted (though not universally used) theory was first proposed by the American psychologist Stanley Smith Stevens in 1946. According to Stevens' theory, the four levels of measurement are nominal, ordinal, interval, and ratio.

Each of these four levels refers to the relationship between the values of the variable.

Nominal measurement

A *nominal measurement* is one in which the values of the variable are names. The names of the different species of Galapagos tortoises are an example of a nominal measurement. The dwelling in which you live - house, apartment, townhouse, condominium, treehouse - is another example of nominal measurement.

Ordinal measurement

An *ordinal measurement* involves collecting information of which the **order** is somehow significant. The name of this level is derived from the use of ordinal numbers for ranking (1st, 2nd, 3rd, etc.). If we measured the different species of tortoise from the largest population to the smallest, this would be an example of ordinal measurement. In ordinal measurement, the distance between two consecutive values does not have meaning. The 1st and 2nd largest tortoise populations by species may differ by a few thousand individuals, while the 7th and 8th may only differ by a few hundred.

Another example of ordinal data is the satisfaction rating scale, sometimes called a Likert scale, which orders responses such as 1 = Strongly Agree, 2 = Agree, 3 = Neutral, 4 = Disagree, 5 = Strongly Disagree. Even when numbers are attached to these responses, they are still considered ordinal data.

Interval measurement

With *interval measurement*, there is significance to the distance between any two values. An example commonly cited for interval measurement is temperature (either degrees Celsius or degrees Fahrenheit). A change of 1 degree is the same if the temperature goes from 0° C to 1° C as it is when the temperature goes from 40° C to 41° C. In addition, there is meaning to the values between the numbers. That is, a half of a degree has meaning.

Ratio measurement

A *ratio measurement* is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. A variable measured at this level not only includes the concepts of order and interval, but also adds the idea of 'nothingness', or zero. With the temperature scale of the previous example, 0° C is really an arbitrarily chosen number (the temperature at which water freezes) and does not represent the absence of temperature. As a result, the ratio between temperatures is relative, and 40° C , for example, is *not* twice as hot as 20° C . On the other hand, for the Galapagos tortoises, the idea of a species having a population of 0 individuals is all too real! As a result, the estimates of the populations are measured on a ratio level, and a species with a population of about 3,300 really is approximately three times as large as one with a population near 1,100.

Comparing the Levels of Measurement

Using Stevens' theory can help make distinctions in the type of data that the numerical/categorical classification could not. Let's use an example from the previous section to help show how you could collect data at different levels of measurement from the same population. Assume your school wants to collect data about all the students in the school.

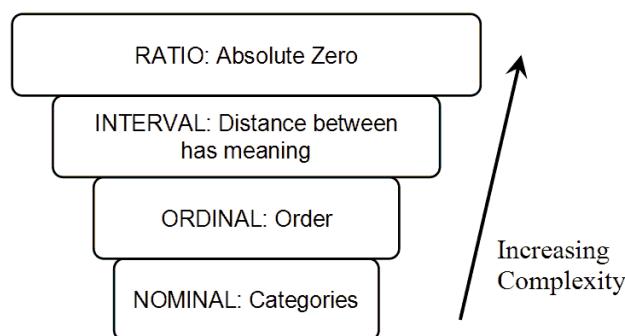
If we collect information about the students' gender, race, political opinions, or the town or sub-division in which they live, we have a nominal measurement.

If we collect data about the students' year in school, we are now ordering that data numerically (9^{th} , 10^{th} , 11^{th} , or 12^{th} grade), and thus, we have an ordinal measurement.

If we gather data for students' SAT math scores, we have an interval measurement. There is no absolute 0, as SAT scores are scaled. The ratio between two scores is also meaningless. A student who scored a 600 did not necessarily do twice as well as a student who scored a 300.

Data collected on a student's age, height, weight, and grades will be measured on the ratio level, so we have a ratio measurement. In each of these cases, there is an absolute zero that has real meaning. Someone who is 18 years old is twice as old as a 9-year-old.

It is also helpful to think of the levels of measurement as building in complexity, from the most basic (nominal) to the most complex (ratio). Each higher level of measurement includes aspects of those before it. The diagram below is a useful way to visualize the different levels of measurement.



Lesson Summary

Data can be measured at different levels, depending on the type of variable and the amount of detail that is collected. A widely used method for categorizing the different types of measurement breaks them down into four groups. Nominal data is measured by classification or categories. Ordinal data uses numerical categories that convey a

meaningful order. Interval measurements show order, and the spaces between the values also have significant meaning. In ratio measurement, the ratio between any two values has meaning, because the data include a true zero value.

Point to Consider

- How do we summarize, display, and compare data measured at different levels?

Review Questions

1. In each of the following situations, identify the level(s) at which each of these measurements has been collected.
 - a. Lois surveys her classmates about their eating preferences by asking them to rank a list of foods from least favorite to most favorite.
 - b. Lois collects similar data, but asks each student what her favorite thing to eat is.
 - c. In math class, Noam collects data on the Celsius temperature of his cup of coffee over a period of several minutes.
 - d. Noam weighs every student's math book.
2. Which of the following statements is **not** true.
 - (a) All ordinal measurements are also nominal.
 - (b) All interval measurements are also ordinal.
 - (c) All ratio measurements are also interval.
 - (d) Steven's levels of measurement is the one theory of measurement that all researchers agree on.

Answers to Review Questions: 1.a. Ordinal 1.b. Nominal 1.c. Interval 1.d. Ratio

2. Choice d. is not true. Stevens' level of measurements theory is widely, but not universally, accepted.

1.3 Measures of Center

Learning Objectives

- Calculate the mean, median, and mode for a set of data, and compare and contrast these measures of center.
- Identify the symbols and know the formulas for sample and population means.
- Calculate the weighted mean, percentiles, and quartiles for a data set.

Introduction

This lesson is an overview of some of the basic statistics used to measure the **center** of a set of data.

Measures of Central Tendency

Once data are collected, it is useful to summarize the data set by identifying a value around which the data are centered. Three commonly used measures of center are the mode, the median, and the mean.

Mode

The *mode* is defined as the most frequently occurring number in a data set. The mode is most useful in situations that involve categorical (qualitative) data that are measured at the nominal level. In the last chapter, we referred to the data with the Galapagos tortoises and noted that the variable 'Climate Type' was such a measurement. For this example, the mode is the value 'humid'.

Example: The students in a statistics class were asked to report the number of children that live in their house (including brothers and sisters temporarily away at college). The data are recorded below:

1, 3, 4, 3, 1, 2, 2, 2, 1, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 6

In this example, the mode could be a useful statistic that would tell us something about the families of statistics students in our school. In this case, 2 is the mode, as it is the most frequently occurring number of children in the sample, telling us that most students in the class come from families where there are 2 children.

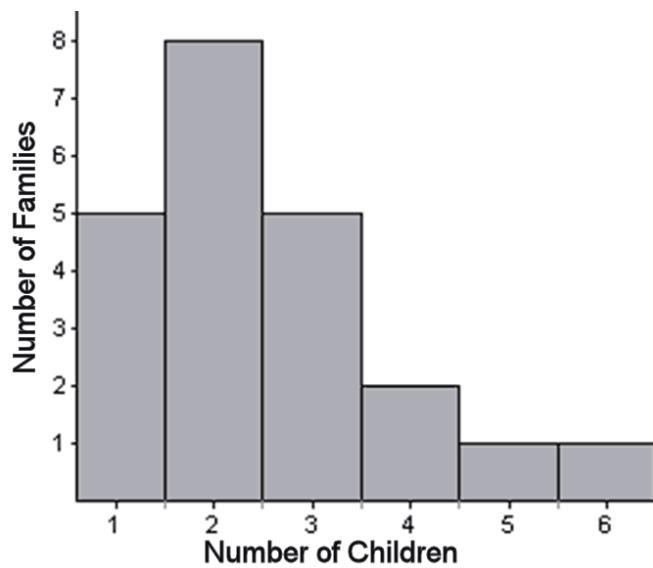
If there were seven 3-child households and seven 2-child households, we would say the data set has two modes. In other words, the data would be *bimodal*. When a data set is described as being bimodal, it is clustered about two different modes. Technically, if there were more than two, they would all be the mode. However, the more of them there are, the more trivial the mode becomes. In these cases, we would most likely search for a different statistic to describe the center of such data.

If there is an equal number of each data value, the mode is not useful in helping us understand the data, and thus, we say the data set has no mode.

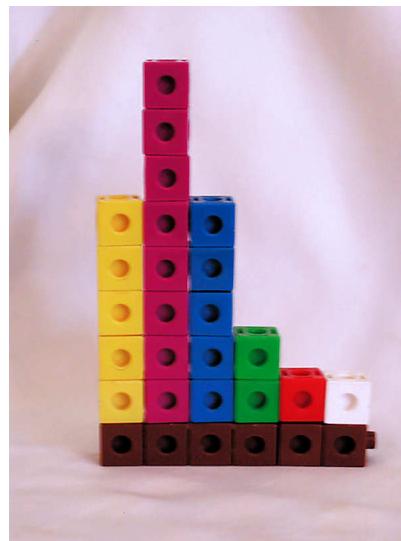
Mean

Another measure of central tendency is the arithmetic average, or *mean*. This value is calculated by adding all the data values and dividing the sum by the total number of data points. The mean is the numerical balancing point of the data set.

We can illustrate this physical interpretation of the mean. Below is a graph of the class data from the last example.

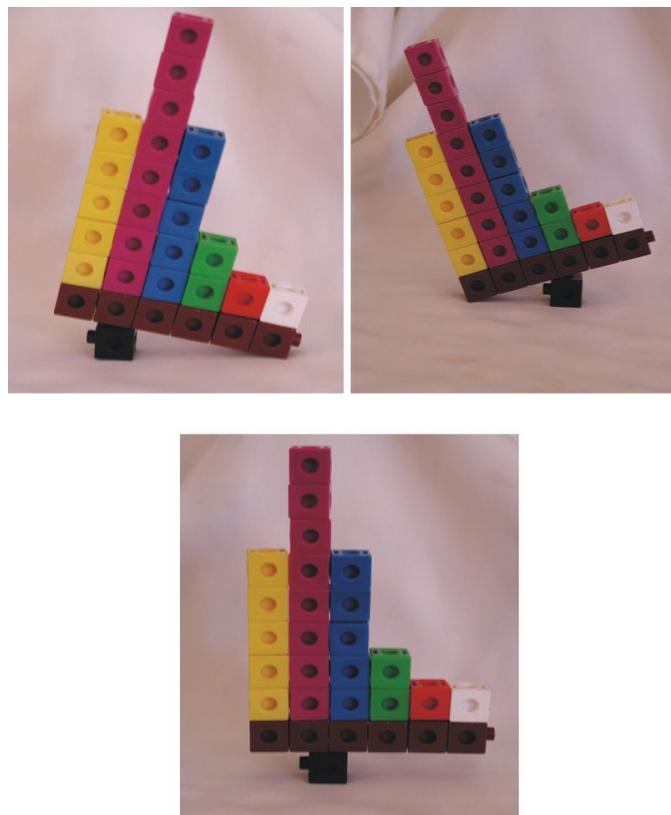


If you have snap cubes like you used to use in elementary school, you can make a physical model of the graph, using one cube to represent each student's family and a row of six cubes at the bottom to hold them together, like this:



There are 22 students in this class, and the total number of children in all of their houses is 55, so the mean of this data is $\frac{55}{22} = 2.5$. Statisticians use the symbol \bar{x} to represent the mean when x is the symbol for a single measurement. Read \bar{x} as “ x bar.”

It turns out that the model that you created balances at 2.5. In the pictures below, you can see that a block placed at 3 causes the graph to tip left, while one placed at 2 causes the graph to tip right. However, if you place the block at 2.5, it balances perfectly!



Symbolically, the formula for the sample mean is as follows

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

where:

\bar{x} is the symbol for the sample mean

n is the sample size.

$\sum x$ is an instruction that tells us to **add** all of the "x" data values.

\bar{x} is a statistic, since it is a measure of a sample, and μ is a parameter, since it is a measure of a population. \bar{x} is an estimate of μ .

Median

The *median* is simply the middle number in an ordered set of data.

Suppose a student took five statistics quizzes and received the following scores:

80, 94, 75, 96, 90

To find the median, you must put the data in order. The median will be the data point that is in the middle. Placing the data in order from least to greatest yields: 75, 80, 90, 94, 96.

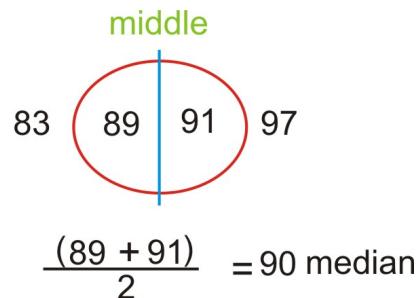
The middle number in this case is the third score, or 90, so the median of this data is 90.

When there is an even number of numbers, no one of the data points will be in the middle. In this case, we take the average (mean) of the two middle numbers.

Example: Consider the following quiz scores: 91, 83, 97, 89

Place them in numeric order: 83, 89, 91, 97.

The second and third numbers straddle the middle of this set. The mean of these two numbers is 90, so the median of the data is 90.



Mean vs. Median

Both the mean and the median are important and widely used measures of center. Consider the following example: Suppose you got an 85 and a 93 on your first two statistics quizzes, but then you had a really bad day and got a 14 on your next quiz!

The mean of your three grades would be 64. Which is a better measure of your performance? As you can see, the middle number in the set is an 85. That middle does not change if the lowest grade is an 84, or if the lowest grade is a 14. However, when you add the three numbers to find the mean, the sum will be much smaller if the lowest grade is a 14.

Outliers and Resistance

The mean and the median are so different in this example because there is one grade that is extremely different from the rest of the data. In statistics, we call such extreme values *outliers*. The mean is affected by the presence of an outlier; however, the median is not. A statistic that is not affected by outliers is called *resistant*. We say that **the median is a resistant measure of center, and the mean is not resistant**. In a sense, the median is able to resist the pull of a far away value, but the mean is drawn to such values. It cannot resist the influence of outlier values. As a result, when we have a data set that contains an outlier, it is often better to use the median to describe the center, rather than the mean.

Example: In 2005, the CEO of Yahoo, Terry Semel, was paid almost \$231,000,000. This is certainly not typical of what the average worker at Yahoo could expect to make. Instead of using the mean salary to describe how Yahoo pays its employees, it would be more appropriate to use the median salary of all the employees.

You will often see medians used to describe the typical value of houses in a given area, as the presence of a very few extremely large and expensive homes could make the mean appear misleadingly large.

Other Measures of Center

Weighted Mean

The *weighted mean* is a method of calculating the mean where instead of each data point contributing equally to the mean, some data points contribute more than others. This could be because they appear more often or because a decision was made to increase their importance (give them more weight). The most common type of weight to use is the frequency, which is the number of times each number is observed in the data. When we calculated the mean for the children living at home, we could have used a weighted mean calculation. The calculation would look like this:

$$\frac{(5)(1) + (8)(2) + (5)(3) + (2)(4) + (1)(5) + (1)(6)}{22}$$

The symbolic representation of this is as follows:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

where:

x is the data value

f is the frequency, how many times a particular data value x occurs

$\sum f$ is the sum of all the frequencies, which is the same as n , the number of pieces of data you have

Percentiles and Quartiles

A *percentile* is a statistic that identifies the percentage of the data that is less than the given value. The most commonly used percentile is the median. Because it is in the numeric middle of the data, half of the data is below the median. Therefore, we could also call the median the 50th percentile. A 40th percentile would be a value in which 40% of the numbers are less than that observation.

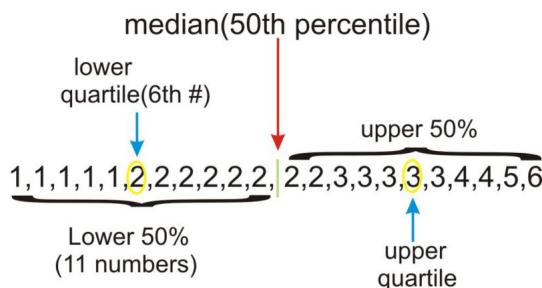
Example: To check a child's physical development, pediatricians use height and weight charts that help them to know how the child compares to children of the same age. A child whose height is in the 70th percentile is taller than 70% of children of the same age.

Two very commonly used percentiles are the 25th and 75th percentiles. The median, 25th, and 75th percentiles divide the data into four parts. Because of this, the 25th percentile is notated as Q_1 and is called the *lower quartile*, and the 75th percentile is notated as Q_3 and is called the *upper quartile*. The median is a middle quartile and is sometimes referred to as Q_2 .

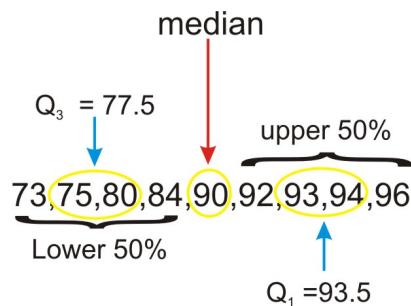
Example: Let's return to the previous data set, which is as follows:

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6

Recall that the median (50th percentile) is 2. The quartiles can be thought of as the medians of the upper and lower halves of the data.



In this case, there are an odd number of values in each half. If there were an even number of values, then we would follow the procedure for medians and average the middle two values of each half. Look at the set of data below:



The median in this set is 90. Because it is the middle number, it is not technically part of either the lower or upper halves of the data, so we do not include it when calculating the quartiles. However, not all statisticians agree that this is the proper way to calculate the quartiles in this case. As we mentioned in the last section, some things in statistics are not quite as universally agreed upon as in other branches of mathematics. The exact method for calculating quartiles is another one of these topics.

Using the TI-84 Calculator to Find the Mean, Median, and Quartiles

Turn power on. Press the [STAT] button. You will see that the EDIT tab at the top is highlighted, and [Option 1: Edit...] is already highlighted. Press the Enter button.

You are now at the list-making screen. Use the arrow buttons on the calculator to scroll to L₁ and make sure the first line is highlighted.

Let's assume we want to input the 4 data values: 4, 8, 7, 5. Enter each of the 4 data values from the first data set into L₁, pressing the Enter button each time you type a piece of data. Your calculator screen should look like this:

L1	L2	L3	Z
4	17		
8	10		
7	9		
5	12		
-----	-----	-----	-----
L2(5) =			

FIGURE 1.1

To leave the list-making screen, press the 2^{nd} button and then the Mode button. This action tells the calculator to quit the list-making screen. The list is saved in the calculator, but you are now free to perform other calculator functions.

Now you want to find the mean and the median. Press [STAT] and then scroll right to CALC. There are thirteen options for the CALC option. Choose option 1, [1-Var Stats]. Press [ENTER].

At the new screen, you will Press [2^{nd}] [1] for "List." This tells the calculator you want the summary stats for the data in L₁. Skip the next line for FreqList. Scroll to Calculate. Your screen display should look like this:

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

FIGURE 1.2

Press [ENTER]. Your screen will look like this:

```
1-Var Stats
x̄=6
Σx=24
Σx²=154
Sx=1.825741858
σx=1.58113883
n=4
```

FIGURE 1.3

This first screen gives the mean , the sample standard deviation , and the population standard deviation , among other

things. If you look at the bottom left of the screen, you will see a small down arrow, and this tells you to scroll down for more information. Scroll all the way down, and the display will look like this:



FIGURE 1.4

These are additional values of interest, including the value of the first and third quartiles and the median.

Lesson Summary

When examining a set of data, we use descriptive statistics to provide information about where the data are centered. The **mode** is a measure of the most frequently occurring number in a data set and is most useful for categorical data and data measured at the nominal level. The mean and median are two of the most commonly used measures of center. The **mean**, or average, is the sum of the data points divided by the total number of data points in the set. In a data set that is a sample from a population, the sample mean is denoted by \bar{x} . The population mean is denoted by μ . The **median** is the numeric middle of a data set. If there are an odd number of data points, this middle value is easy to find. If there is an even number of data values, the median is the mean of the middle two values. An **outlier** is a number that has an extreme value when compared with most of the data. The median is resistant. That is, it is not affected by the presence of outliers. The mean is not resistant, and therefore, the median tends to be a more appropriate measure of center to use in examples that contain outliers. Because the mean is the numerical balancing point for the data, it is an extremely important measure of center that is the basis for many other calculations and processes necessary for making useful conclusions about a set of data.

A **weighted mean** involves multiplying individual data values by their frequencies or percentages before adding them and then dividing by the total of the frequencies (weights).

A **percentile** is a data value for which the specified percentage of the data is below that value. The median is the 50th percentile. Two well-known percentiles are the 25th percentile, which is called the **lower quartile**, Q_1 , and the 75th percentile, which is called the **upper quartile**, Q_3 .

Points to Consider

- How do you determine which measure of center best describes a particular data set?
- What are the effects of outliers on the various measures of spread?
- How can we represent data visually using the various measures of center?

Review Questions

1. In Lois' 2nd grade class, all of the students are between 45 and 52 inches tall, except one boy, Lucas, who is 62 inches tall. Which of the following statements is true about the heights of all of the students?
 - a. The mean height and the median height are about the same.
 - b. The mean height is greater than the median height.

- c. The mean height is less than the median height.
 - d. More information is needed to answer this question.
 - e. None of the above is true.
2. Enrique has a 91, 87, and 95 for his statistics grades for the first three quarters. His mean grade for the year must be a 93 in order for him to be exempt from taking the final exam. Assuming grades are rounded following valid mathematical procedures, what is the lowest whole number grade he can get for the 4th quarter and still be exempt from taking the exam?
3. The chart below shows the data from the Galapagos tortoise preservation program with just the number of individual tortoises that were bred in captivity and reintroduced into their native habitat.

TABLE 1.3:

Island or Volcano	Number of Individuals Repatriated
Wolf	40
Darwin	0
Alcedo	0
Sierra Negra	286
Cerro Azul	357
Santa Cruz	210
Españaola	1293
San Cristóbal	55
Santiago	498
Pinzón	552
Pinta	0

Figure: Approximate Distribution of Giant Galapagos Tortoises in 2004 (“Estado Actual De Las Poblaciones de Tortugas Terrestres Gigantes en las Islas Galápagos,” Marquez, Wiedenfeld, Snell, Fritts, MacFarland, Tapia, y Nanjoa, Scología Aplicada, Vol. 3, Num. 1, 2, pp. 98-11).

For this data, calculate each of the following:

- (a) mode
 - (b) median
 - (c) mean
 - (d) upper and lower quartiles
 - (e) Why is the answer to (c) significantly higher than the answer to (b)?
4. There are 2 sections of Biology 101. Section A has 20 students, and their average on the last test was 80. Section B has 32 students, and their average on the last test was 90. What is the course average?

Answers to Review Questions: (1.) B (2.) 99 (3.a.) 0 (3.b.) 210 (3.c.) 299.18 (3.d.) $Q_1 = 0$ and $Q_3 = 498$ (3.e.) There is an outlier. (4.) 86.15

1.4 Measures of Spread

Learning Objectives

- Calculate the range, the interquartile range, the standard deviation, and the variance for a population and a sample, and know the symbols, formulas, and uses of these measures of spread.

Introduction

In the last lesson, we studied measures of central tendency. Another important feature that can help us understand more about a data set is the manner in which the data are distributed, or spread. Variation and dispersion are words that are also commonly used to describe this feature. There are several commonly used statistical measures of spread that we will investigate in this lesson.

Range

One measure of spread is the range. The *range* is simply the difference between the smallest value (minimum) and the largest value (maximum) in the data.

Example: Return to the data set used in the previous lesson, which is shown below:

75, 80, 90, 94, 96

The range of this data set is $96 - 75 = 21$. This is telling us the distance between the maximum and minimum values in the data set.

The range is useful because it requires very little calculation, and therefore, gives a quick and easy snapshot of how the data are spread. However, it is limited, because it only involves two values in the data set, and it is not resistant to outliers.

Interquartile Range

The *interquartile range* is the difference between the Q_3 and Q_1 , and it is abbreviated *IQR*. Thus, $IQR = Q_3 - Q_1$. The *IQR* gives information about how the middle 50% of the data are spread. Fifty percent of the data values are always between Q_3 and Q_1 .

Example: A recent study proclaimed Mobile, Alabama the wettest city in America (http://www.livescience.com/environment/070518_rainy_cities.html). The following table lists measurements of the approximate annual rainfall in Mobile for the last 10 years. Find the range and *IQR* for this data.

TABLE 1.4:

	Rainfall (inches)
1998	90
1999	56
2000	60
2001	59
2002	74

TABLE 1.4: (continued)

	Rainfall (inches)
2003	76
2004	81
2005	91
2006	47
2007	59

Figure: Approximate Total Annual Rainfall, Mobile, Alabama. *Source:* <http://www.cwop1353.com/CoopGaugeData.htm>

First, place the data in order from smallest to largest. The range is the difference between the minimum and maximum rainfall amounts.

$$\text{RANGE: } 91 - 47 = 44$$

To find the *IQR*, first identify the quartiles, and then compute $Q_3 - Q_1$.

$$\text{IQR: } 81 - 59 = 22$$

In this example, the range tells us that there is a difference of 44 inches of rainfall between the wettest and driest years in Mobile. The *IQR* shows that there is a difference of 22 inches of rainfall, even in the middle 50% of the data. It appears that Mobile experiences wide fluctuations in yearly rainfall totals, which might be explained by its position near the Gulf of Mexico and its exposure to tropical storms and hurricanes.

Standard Deviation

The standard deviation is an extremely important measure of spread that is based on the mean. Recall that the mean is the numerical balancing point of the data. One way to measure how the data are spread is to look at how far away each of the values is from the mean. The difference between a data value and the mean is called the *deviation*. Written symbolically, it would be as follows:

$$\text{Deviation} = x - \bar{x}$$

Let's take the simple data set of three measurements shown below:

9.5, 11.5, 12

The mean of this data set is 11. The deviations are as follows:

TABLE 1.5: Table of Deviations

x	$x - \bar{x}$
9.5	$9.5 - 11 = -1.5$
11.5	$11.5 - 11 = 0.5$
12	$12 - 11 = 1$

Notice that if a data value is less than the mean, the deviation of that value is negative. Points that are above the mean have positive deviations.

The *standard deviation* is a measure of the typical, or average, deviation for all of the data points from the mean. However, the very property that makes the mean so special also makes it tricky to calculate a standard deviation. Because the mean is the balancing point of the data, when you add the deviations, they always sum to 0.

TABLE 1.6: Table of Deviations, Including the Sum.

Observed Data	Deviations
9.5	$9.5 - 11 = -1.5$
11.5	$11.5 - 11 = 0.5$
12	$12 - 11 = 1$
Sum of deviations	$-1.5 + 0.5 + 1 = 0$

Therefore, we need all the deviations to be positive before we add them up. One way to do this would be to make them positive by taking their absolute values. This is a technique we use for a similar measure called the *mean absolute deviation*. For the standard deviation, though, we square all the deviations. The square of any real number is always positive.

TABLE 1.7:

Observed Data x	Deviation $x - \bar{x}$	$(x - \bar{x})^2$
9.5	-1.5	$(-1.5)^2 = 2.25$
11.5	0.5	$(0.5)^2 = 0.25$
12	1	1

$$\text{Sum of the squared deviations} = 2.25 + 0.25 + 1 = 3.5$$

We want to find the average of the squared deviations. Usually, to find an average, you divide by the number of terms in your sum. In finding the standard deviation, however, we divide by $n - 1$. In this example, since $n = 3$, we divide by 2. The result, which is called the *variance*, is 1.75. The variance of a sample is denoted by s^2 and is a measure of how closely the data are clustered around the mean. Because we squared the deviations before we added them, the units we were working in were also squared. To return to the original units, we must take the square root of our result: $\sqrt{1.75} \approx 1.32$. **This quantity is the sample standard deviation and is denoted by s .** The number indicates that in our sample, the typical data value is approximately 1.32 units away from the mean. It is a measure of how closely the data are clustered around the mean. A small standard deviation means that the data points are clustered close to the mean, while a large standard deviation means that the data points are spread out from the mean.

Example: The following are scores for two different students on two quizzes:

Student 1: 100; 0

Student 2: 50; 50

Note that the mean score for each of these students is 50.

Student 1: Deviations: $100 - 50 = 50$; $0 - 50 = -50$

Squared deviations: 2500; 2500

Variance = 5000

Standard Deviation = 70.7

Student 2: Deviations: $50 - 50 = 0$; $50 - 50 = 0$

Squared Deviations: 0; 0

Variance = 0

Standard Deviation = 0

Student 2 has scores that are tightly clustered around the mean. In fact, the standard deviation of zero indicates that there is no variability. The student is absolutely consistent.

So, while the average of each of these students is the same (50), one of them is consistent in the work he/she does, and the other is not. This raises questions: Why did student 1 get a zero on the second quiz when he/she had a perfect paper on the first quiz? Was the student sick? Did the student forget about the quiz and not study? Or was the second quiz indicative of the work the student can do, and was the first quiz the one that was questionable? Did the student cheat on the first quiz?

There is one more question that we haven't answered regarding standard deviation, and that is, "Why $n - 1$?" Dividing by $n - 1$ is only necessary for the calculation of the standard deviation of a **sample**. When you are calculating the standard deviation of a population, you divide by N , the number of data points in your population. When you have a sample, you are not getting data for the entire population, and there is bound to be random variation due to sampling (remember that this is called sampling error).

When we claim to have the standard deviation, we are making the following statement:

"The typical distance of a point from the mean is ..."

But we might be off by a little from using a sample, so it would be better to overestimate s to represent the standard deviation.

Before we close, look at this chart closely. It summarizes the symbols used in formulas for the parameters and the statistics for the mean, the standard deviation, and the variance. Commit these to memory!

TABLE 1.8: Symbols

	Sample	Population
Mean	\bar{x}	μ
Standard Deviation	s	σ
Variance	s^2	σ^2

Formulas

Sample Standard Deviation:

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

where:

Σ is the instruction to "add up"

x is each individual data value.

\bar{x} is the mean of the sample.

n is the sample size.

Variance of a sample:

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

where:

x is each data value.

\bar{x} is the mean of the sample.

n is the sample size.

Using the TI-84 to Calculate the Standard Deviation

The calculation of the sample standard deviation is quite irksome for more than 3 data values. The use of the calculator will ease the computational burden. (In Section 1.3, detailed instructions were given for making lists, but a short review of listmaking is provided here as review.)

Let's say you want to calculate the standard deviation of 6 IQ scores: 80, 85, 100, 104, 116, and 135.

First, enter the data values into the List by pressing [STAT] and choosing (EDIT)(Edit); then press [ENTER]. If you have already entered data previously into List 1, you can clear the entire column by scrolling all the way up to the "L1" heading of the first column; then press [CLEAR] and then [ENTER]. Enter each of the 6 data values into L₁.

Now press [STAT], and scroll to [CALC]. Choose Option 1: [1-Var Stats] and press [ENTER]. Instruct the calculator that you want to use data from L₁ by pressing the [2nd] and [1] buttons. Scroll down to "Calculate" and press [ENTER]. Your screen should show:

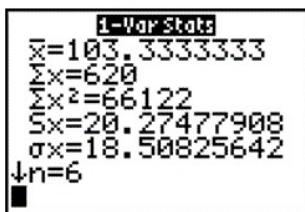


FIGURE 1.5

The calculator symbol for the sample standard deviation s is s_x , and it is shown as 20.27477908. The population standard deviation has a symbol of σ_x , and its value is 18.50825642. We will seldom use the population standard deviation value, because in statistics we almost always use samples.

Look at the small arrow on the calculator display. It is an indicator that if you scroll down, there will be more data summary information. Scroll down, and your display will look like this:

You can easily calculate the interquartile range (IQR) by using the formula $IQR = Q_3 - Q_1$, which are both provided to you on this screen. For the IQ data, the IQR is $(116 - 85) = 31$.

Lesson Summary

When examining a set of data, we use descriptive statistics to provide information about how the data are spread out. The **range** is a measure of the difference between the smallest and largest numbers in a data set. The **interquartile range** is the difference between the upper and lower quartiles. A more informative measure of spread is based on

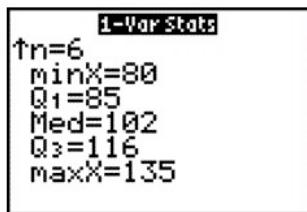


FIGURE 1.6

the mean. We can look at how individual points vary from the mean by subtracting the mean from the data value. This is called the deviation. The standard deviation is a measure of the average deviation for the entire data set. Because the deviations always sum to zero, we find the standard deviation by adding the squared deviations. When we have the entire population, the sum of the squared deviations is divided by the population size. This value is called the **variance**. Taking the square root of the variance gives the **standard deviation**. *For a population, the standard deviation is denoted by σ . Because a sample is prone to random variation (sampling error), we adjust the sample standard deviation to make it a little larger by dividing the sum of the squared deviations by one less than the number of observations. The result of that division is the sample variance, and the square root of the sample variance is the sample standard deviation, usually notated as s .*

Points to Consider

- How do you determine which measure of spread best describes a particular data set?
- What information does the standard deviation tell us about the specific, real data being observed?
- What are the effects of outliers on the various measures of spread?
- How does altering the spread of a data set affect its visual representation(s)?

Review Questions

1. Use the rainfall data in the chart below to answer this question.
 - a. Calculate and record the sample mean:
 - b. Complete the chart to calculate the variance and the standard deviation.

TABLE 1.9:

Year	Rainfall (inches)	Deviation	Squared Deviations
1999	56		
2000	60		
2001	54		
2002	74		
2003	76		

Variance:

Standard Deviation:

2. Use the Galapagos Tortoise data below to answer parts a. and b. Use your calculator !

TABLE 1.10:

TABLE 1.10: (continued)

Island or Volcano	Number of Individuals Repatriated
Santa Cruz	210
Españaola	1293
San Cristóbal	55
Santiago	498
Pinzón	552
Pinta	0

- a. Calculate the range and the *IQR* for this data.
- b. Calculate the standard deviation for this data.
3. If $\sigma^2 = 9$, then the population standard deviation is:
- a. 3
 - b. 8
 - c. 9
 - d. 81
4. Which data set has the largest standard deviation?
- a. 10 10 10 10 10
 - b. 0 0 10 10 10
 - c. 0 9 10 11 20
 - d. 20 20 20 20 20

Answers to Review Questions: (1.a.) mean is 64 (1.b.) variance = 106 and standard deviation is 10.3.
 (2.a.) range = 1293 (2.b.) *IQR* = 498 (3.) A (4.) C

Part One: Multiple Choice

1. Which of the following is true for any set of data?
 - a. The range is a resistant measure of spread.
 - b. The standard deviation is not resistant.
 - c. The range can be greater than the standard deviation.
 - d. The *IQR* is always greater than the range.
 - e. The range can be negative.
2. The following shows the mean number of days of precipitation by month in Juneau, Alaska:

TABLE 1.11: Mean Number of Days With Precipitation >0.1 inches

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
18	17	18	17	17	15	17	18	20	24	20	21

Source: <http://www.met.utah.edu/jhorel/html/wx/climate/daysrain.html> (2/06/08)

Which month contains the median number of days of rain?

- (a) January
- (b) February
- (c) June
- (d) July
- (e) September

1. Given the data 2, 10, 14, 6, which of the following is equivalent to \bar{x} ?

- a. mode
- b. median
- c. midrange
- d. range
- e. none of these

2. Place the following in order from smallest to largest.

I. Range

II. Standard Deviation

III. Variance

- a. I, II, III
- b. I, III, II
- c. II, III, I
- d. II, I, III
- e. It is not possible to determine the correct answer.

3. On the first day of school, a teacher asks her students to fill out a survey with their name, gender, age, and homeroom number. How many quantitative variables are there in this example?

- a. 0
- b. 1
- c. 2
- d. 3
- e. 4

4. You collect data on the shoe sizes of the students in your school by recording the sizes of 50 randomly selected males' shoes. What is the highest level of measurement that you have demonstrated?

- a. nominal
- b. ordinal
- c. interval
- d. ratio

5. According to a 2002 study, the mean height of Chinese men between the ages of 30 and 65 is 164.8 cm, with a standard deviation of 6.4 cm (<http://aje.oxfordjournals.org/cgi/reprint/155/4/346.pdf> accessed Feb 6, 2008). Which of the following statements is true based on this study?

- a. The interquartile range is 12.8 cm.
- b. All Chinese men are between 158.4 cm and 171.2 cm.
- c. At least 75% of Chinese men between 30 and 65 are between 158.4 and 171.2 cm.
- d. At least 75% of Chinese men between 30 and 65 are between 152 and 177.6 cm.

- e. All Chinese men between 30 and 65 are between 152 and 177.6 cm.
6. Sampling error is best described as:
- The unintentional mistakes a researcher makes when collecting information
 - The natural variation that is present when you do not get data from the entire population
 - A researcher intentionally asking a misleading question, hoping for a particular response
 - When a drug company does its own experiment that proves its medication is the best
 - When individuals in a sample answer a survey untruthfully
7. If the sum of the squared deviations for a sample of 20 individuals is 277, the standard deviation is closest to:
- 3.82
 - 3.85
 - 13.72
 - 14.58
 - 191.82

Part Two: Open-Ended Questions

1. Erica's grades in her statistics classes are as follows: Quizzes: 62, 88, 82 Labs: 89, 96 Tests: 87, 99
- In this class, quizzes count once, labs count twice as much as a quiz, and tests count three times as much as a quiz. Determine the following:
 - mode
 - mean
 - median
 - upper and lower quartiles
 - midrange
 - range
 - If Erica's quiz grade of 62 was removed from the data, briefly describe (without recalculating) the anticipated effect on the statistics you calculated in part (a).
2. Mr. Crunchy's sells small bags of potato chips that are advertised to contain 12 ounces of potato chips. To minimize complaints from their customers, the factory sets the machines to fill bags with an average weight of 13 ounces. For an experiment in his statistics class, Spud goes to 5 different stores, purchases 1 bag from each store, and then weighs the contents. The weights of the bags are: 13, 18, 12, 65, 12, 87, 13, 32, and 12.93 grams.

(a) Calculate the sample mean.

(b) Complete the chart below to calculate the standard deviation of Spud's sample.

TABLE 1.12:

Observed Data	$(x - \bar{x})$	$(x - \bar{x})^2$
13.18		
12.65		
12.87		
13.32		
12.93		
Sum of the squared deviations		

- (c) Calculate the variance.
- (d) Calculate the standard deviation.
- (e) Explain what the standard deviation means in the context of the problem.
1. The following table includes data on the number of square kilometers of the more substantial islands of the Galapagos Archipelago. (There are actually many more islands if you count all the small volcanic rock outcroppings as islands.)
- TABLE 1.13:**
- | Island | Approximate Area (sq. km) |
|---------------|----------------------------------|
| Baltra | 8 |
| Darwin | 1.1 |
| Españaola | 60 |
| Fernandina | 642 |
| Floreana | 173 |
| Genovesa | 14 |
| Isabela | 4640 |
| Marchena | 130 |
| North Seymour | 1.9 |
| Pinta | 60 |
| Pinzón | 18 |
| Rabida | 4.9 |
| San Cristóbal | 558 |
| Santa Cruz | 986 |
| Santa Fe | 24 |
| Santiago | 585 |
| South Plaza | 0.13 |
| Wolf | 1.3 |
- Source: http://en.wikipedia.org/wiki/Gal%C3%A1pagos_Islands*
- (a) Calculate each of the following for the above data:
- mode
 - mean
 - median
 - upper quartile
 - lower quartile
 - range
 - standard deviation
- (b) Explain why the mean is so much larger than the median in the context of this data.
- (c) Explain why the standard deviation is so large.
1. At <http://content.usatoday.com/sports/baseball/salaries/default.aspx>, USA Today keeps a database of major league baseball salaries. Pick a team and look at the salary statistics for that team. Next to the average salary, you will see the median salary. If this site is not available, a web search will most likely locate similar data.

- (a) Record the median and verify that it is correct by clicking on the team and looking at the salaries of the individual players.
- (b) Find the other measures of center and record them.
- (i) mean
 - (ii) mode
 - (iii) midrange
 - (iv) lower quartile
 - (v) upper quartile
 - (vi) IQR
- (c) Explain the real-world meaning of each measure of center in the context of this data.
- (i) mean
 - (ii) median
 - (iii) mode
 - (iv) midrange
 - (v) lower quartile
 - (vi) upper quartile
 - (vii) IQR
- (d) Find the following measures of spread:
- (i) range
 - (ii) standard deviation
- (e) Explain the real-world meaning of each measure of spread in the context of this situation.
- (i) range
 - (ii) standard deviation
- (f) Write two sentences commenting on two interesting features about the way the salary data are distributed for this team.

Keywords

Bias

The systematic error in sampling is called *bias*.

Bimodal

When data set is clustered about two different modes, it is described as being bimodal.

Categorical variable

When a characteristic can be neatly placed into well-defined groups, or categories, that do not depend on order, it is called a *categorical variable*, or *qualitative variable*.

Census

to get accurate and complete information about all the residents of the United States to help effectively address the needs of a changing population. This is why a complete counting, or census, is only attempted every ten years.

Chebyshev's Theorem

The Probability that any random variable that lies within k standard deviations of its mean is atleast $1 - \frac{1}{k^2}$. It emphasizes the fact that the variance and the standard deviation measure the variability of a random variable about its mean.

Deviation

The difference between the data value and the mean

Interquartile range(IQR)

The range is a measure of the difference between the smallest and largest numbers in a data set. The interquartile range is the difference between the upper and lower quartiles.

Interval

The distance between any two values.

Interval estimate

A statistician would report the estimate of the parameter in two ways: as a *point estimate* (e.g., 915) and also as an *interval estimate*.

Levels of measurement

Some researchers and social scientists use a more detailed distinction, called the *levels of measurement*,

Lower quartile

The 25th percentile is notated as Q_1 and is called the *lower quartile*,

Mean

The mean is the numerical balancing point of the data set.

Mean absolute deviation

This is a technique we use for a similar measure called the *mean absolute deviation*.

Median

The *median* is simply the middle number in an ordered set of data.

Midrange

The *midrange* (sometimes called the midextreme) is found by taking the mean of the maximum and minimum values of the data set.

Mode

The *mode* is defined as the most frequently occurring number in a data set.

 $n\%$ trimmed mean

a statistician may choose to remove a certain percentage of the extreme values. This is called an $n\%$ *trimmed mean*.

Nominal

Nominal data is measured by classification or categories.

Numerical variable

how many individuals there are per square kilometer. This type of variable is called a *numerical variable*, or *quantitative variable*.

Ordinal

Ordinal data uses numerical categories that convey a meaningful order.

Outliers

Extreme values in a Dataset are referred to as *outliers*. The mean is affected by the presence of an outlier;

Parameter

An actual value of a population variable is called a parameter.

Percentile

A percentile is a data value for which the specified percentage of the data is below that value.

Point estimate

A statistician would report the estimate of the parameter in two ways: as a *point estimate*

Population

the total group being studied is called the population.

Qualitative variable

that do not depend on order, it is called a *categorical variable*, or *qualitative variable*..

Quantitative variable

quantity, of the characteristic is most important. how many individuals there are per square kilometer. This type of variable is called a *numerical variable*, or *quantitative variable*.

Range

The *range* is the difference between the smallest value (minimum) and the largest value (maximum) in the data.

Ratio

the estimates of the populations are measured on a ratio level,

Resistant

A statistic that is not affected by outliers is called *resistant*.

Sample

representative group from the population, called a *sample*.

Sampling error

The difference between the true parameter and the statistic obtained by sampling is called *sampling error*.

Standard deviation

The standard deviation is an extremely important measure of spread that is based on the mean.

Statistic

Any number that describes the individuals in a sample (length, weight, age) is called a *statistic*.

Trimmed mean

Recall that the mean is not resistant to the effects of outliers.

Unit

Each member of the population is called a *unit*.

Upper quartile

The 75th percentile is notated as Q_3 and is called the *upper quartile*.

Variables

A researcher studying Galapagos Tortoises would be interested in collecting information about different characteristics of the tortoises. Those characteristics are called *variables*.

Variance

When we have the entire population, the sum of the squared deviations is divided by the population size. This value is called the variance.

Weighted mean

The *weighted mean* is a method of calculating the mean where instead of each data point contributing equally to the mean, some data points contribute more than others.

1.5 References

1. CK-12 Foundation. . CCSA
2. CK-12 Foundation. Input for List 1. CCSA
3. CK-12 Foundation. 1-Var Stats Display. CCSA
4. CK-12 Foundation. 1-Var Stats Display (cont'd). CCSA
5. CK-12 Foundation. . CCSA
6. CK-12 Foundation. . CCSA

CHAPTER**2****Visualizations of Data****Chapter Outline**

- 2.1 HISTOGRAMS AND FREQUENCY DISTRIBUTIONS**
 - 2.2 COMMON GRAPHS AND DATA PLOTS**
 - 2.3 BOX-AND-WHISKER PLOTS**
-

2.1 Histograms and Frequency Distributions

Learning Objectives

- Read and make frequency tables for a data set.
- Identify and translate data sets to and from a histogram, a relative frequency histogram, and a frequency polygon.
- Identify histogram distribution shapes as skewed or symmetric and understand the basic implications of these shapes.
- Identify and translate data sets to and from an ogive plot (cumulative distribution function).

Introduction

Charts and graphs of various types, when created carefully, can provide instantaneous important information about a data set without calculating, or even having knowledge of, various statistical measures. This chapter will concentrate on some of the more common visual presentations of data.

Frequency Tables

The earth has seemed so large in scope for thousands of years that it is only recently that many people have begun to take seriously the idea that we live on a planet of limited and dwindling resources. This is something that residents of the Galapagos Islands are also beginning to understand. Because of its isolation and lack of resources to support large, modernized populations of humans, the problems that we face on a global level are magnified in the Galapagos. Basic human resources such as water, food, fuel, and building materials must all be brought in to the islands. More problematically, the waste products must either be disposed of in the islands, or shipped somewhere else at a prohibitive cost. As the human population grows exponentially, the Islands are confronted with the problem of what to do with all the waste. In most communities in the United States, it is easy for many to put out the trash on the street corner each week and perhaps never worry about where that trash is going. In the Galapagos, the desire to protect the fragile ecosystem from the impacts of human waste is more urgent and is resulting in a new focus on renewing, reducing, and reusing materials as much as possible. There have been recent positive efforts to encourage recycling programs.

It is not easy to bury tons of trash in solid volcanic rock. The sooner we realize that we are in the same position of limited space and that we have a need to preserve our global ecosystem, the more chance we have to save not only the uniqueness of the Galapagos Islands, but that of our own communities. All of the information in this chapter is focused around the issues and consequences of our recycling habits, or lack thereof!

Example: Water, Water, Everywhere!

Bottled water consumption worldwide has grown, and continues to grow at a phenomenal rate. According to the Earth Policy Institute, 154 billion gallons were produced in 2004. While there are places in the world where safe water supplies are unavailable, most of the growth in consumption has been due to other reasons. The largest consumer of bottled water is the United States, which arguably could be the country with the best access to safe, convenient, and reliable sources of tap water. The large volume of toxic waste that is generated by the plastic bottles and the small fraction of the plastic that is recycled create a considerable environmental hazard. In addition, huge volumes of carbon emissions are created when these bottles are manufactured using oil and transported great distances by oil-burning vehicles.

**FIGURE 2.1**

The Recycling Center on Santa Cruz in the Galapagos turns all the recycled glass into pavers that are used for the streets in Puerto Ayora.

Example: Take an informal poll of your class. Ask each member of the class, on average, how many beverage bottles they use in a week. Once you collect this data, the first step is to organize it so it is easier to understand. A frequency table is a common starting point. *Frequency tables* simply display each value of the variable, and the number of occurrences (the frequency) of each of those values. In this example, the variable is the number of plastic beverage bottles of water consumed each week.

Consider the following raw data:

6, 4, 7, 7, 8, 5, 3, 6, 8, 6, 5, 7, 7, 5, 2, 6, 1, 3, 5, 4, 7, 4, 6, 7, 6, 6, 7, 5, 4, 6, 5, 3

When creating a frequency table, it is often helpful to use tally marks as a running total to avoid missing a value or over-representing another.

TABLE 2.1: Frequency table using tally marks

Number of Plastic Beverage Bottles per Week	Tally	Frequency
1		1
2		1
3		3
4		4
5		5
6		6
		8

TABLE 2.1: (continued)

Number of Plastic Beverage Bottles per Week	Tally	Frequency
7		5
8		2

The following data set shows the countries in the world that consume the most bottled water per person per year.

TABLE 2.2:

Liters of Bottled Water Consumed per Person per Year	
Italy	183.6
Mexico	168.5
United Arab Emirates	163.5
Belgium and Luxembourg	148.0
France	141.6
Spain	136.7
Germany	124.9
Lebanon	101.4
Switzerland	99.6
Cyprus	92.0
United States	90.5
Saudi Arabia	87.8
Czech Republic	87.1
Austria	82.1
Portugal	80.3

Figure: Bottled Water Consumption per Person in Leading Countries in 2004. Source: http://www.earth-policy.org/Updates/2006/Update51_data.htm

These data values have been measured at the ratio level. There is some flexibility required in order to create meaningful and useful categories for a frequency table. The values range from 80.3 liters to 183 liters. By examining the data, it seems appropriate for us to create our frequency table in groups of 10. We will skip the tally marks in this case, because the data values are already in numerical order, and it is easy to see how many are in each classification.

A bracket, '[' or ']', indicates that the endpoint of the interval is included in the class. A parenthesis, '(' or ')', indicates that the endpoint is not included. It is common practice in statistics to include a number that borders two classes as the larger of the two numbers in an interval. For example, $[80 - 90)$ means this classification includes everything from 80 and gets infinitely close to, but not equal to, 90. 90 is included in the next class, $[90 - 100)$.

TABLE 2.3:

Liters per Person	Frequency
$[80 - 90)$	4
$[90 - 100)$	3
$[100 - 110)$	1
$[110 - 120)$	0
$[120 - 130)$	1
$[130 - 140)$	1
$[140 - 150)$	2
$[150 - 160)$	0

TABLE 2.3: (continued)

Liters per Person	Frequency
[160 – 170)	2
[170 – 180)	0
[180 – 190)	1

Figure: Completed Frequency Table for World Bottled Water Consumption Data (2004)

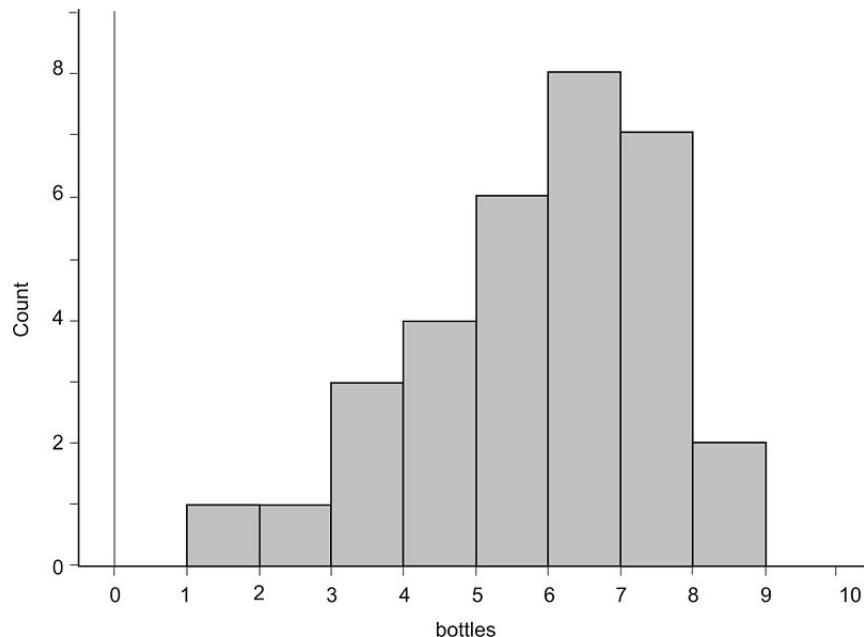
Histograms

Once you can create a frequency table, you are ready to create our first graphical representation, called a *histogram*. Let's revisit our data about student bottled beverage habits.

TABLE 2.4: Completed Frequency Table for Water Bottle Data

Number of Plastic Beverage Bottles per Week	Frequency
1	1
2	1
3	3
4	4
5	6
6	8
7	7
8	2

Here is the same data in a histogram:

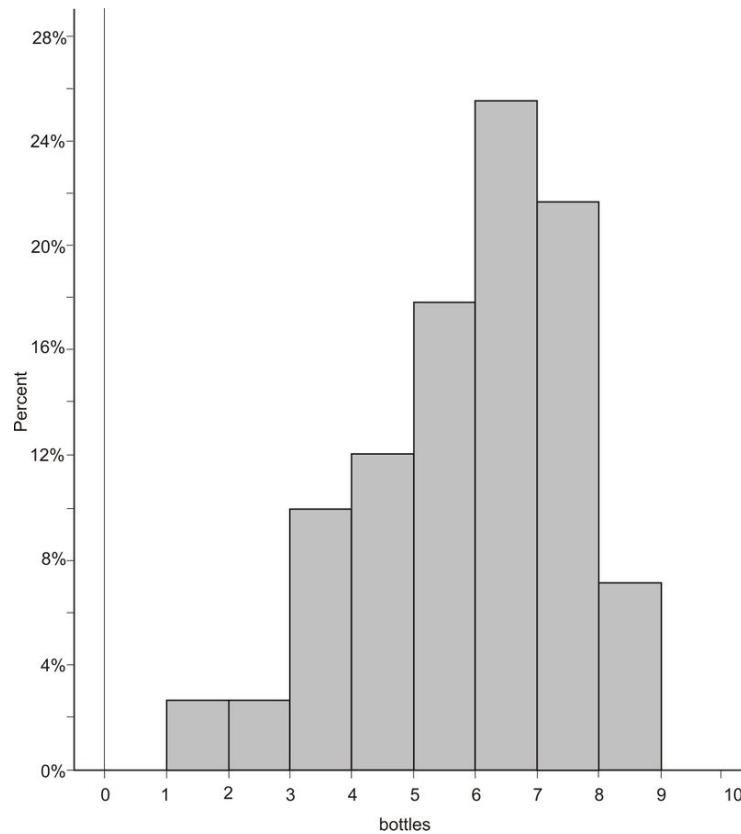


In this case, the horizontal axis represents the variable (number of plastic bottles of water consumed), and the vertical axis is the frequency, or count. Each vertical bar represents the number of people in each class of ranges of bottles. For example, in the range of consuming [1 – 2) bottles, there is only one person, so the height of the bar is at 1. We can see from the graph that the most common class of bottles used by people each week is the [6 – 7) range, or six bottles per week.

A histogram is for numerical data. With histograms, the different sections are referred to as *bins*. Think of a column, or bin, as a vertical container that collects all the data for that range of values. If a value occurs on the border between two bins, it is commonly agreed that this value will go in the larger class, or the bin to the right. It is important when drawing a histogram to be certain that there are enough bins so that the last data value is included. Often this means you have to extend the horizontal axis beyond the value of the last data point. In this example, if we had stopped the graph at 8, we would have missed that data, because the 8's actually appear in the bin between 8 and 9. Very often, when you see histograms in newspapers, magazines, or online, they may instead label the midpoint of each bin. Some graphing software will also label the midpoint of each bin, unless you specify otherwise.

Relative Frequency Histogram

A *relative frequency histogram* is just like a regular histogram, but instead of labeling the frequencies on the vertical axis, we use the percentage of the total data that is present in that bin. For example, there is only one data value in the first bin. This represents $\frac{1}{32}$, or approximately 3%, of the total data. Thus, the vertical bar for the bin extends upward to 3%.

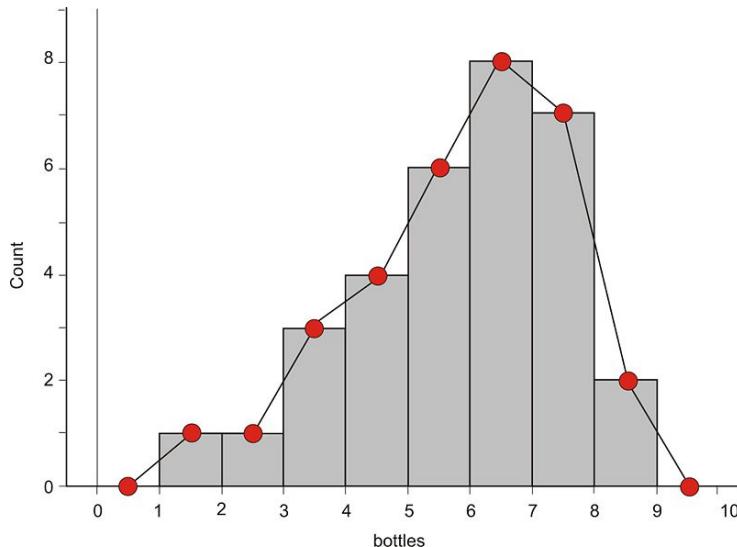


Frequency Polygons

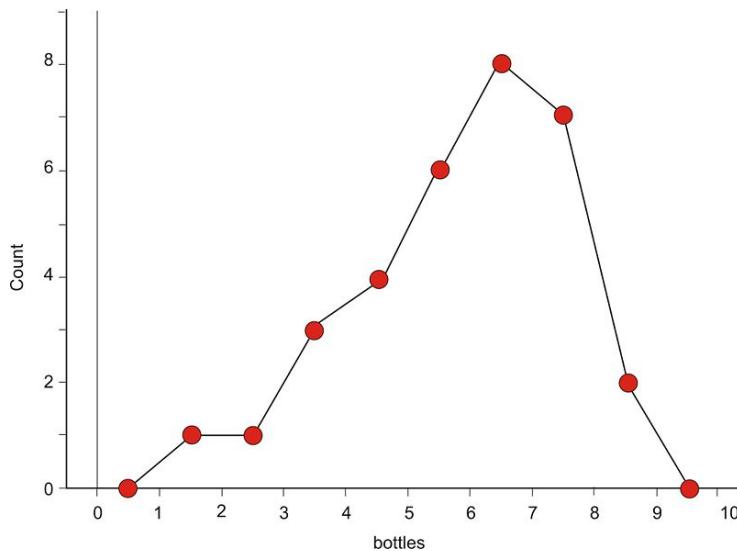
A *frequency polygon* is similar to a histogram, but instead of using bins, a polygon is created by plotting the frequencies and connecting those points with a series of line segments.

To create a frequency polygon for the bottle data, we first find the midpoints of each classification, plot a point at the frequency for each bin at the midpoint, and then connect the points with line segments. To make a polygon with the horizontal axis, plot the midpoint for the class one greater than the maximum for the data, and one less than the minimum.

Here is a frequency polygon constructed directly from the previously-shown histogram:

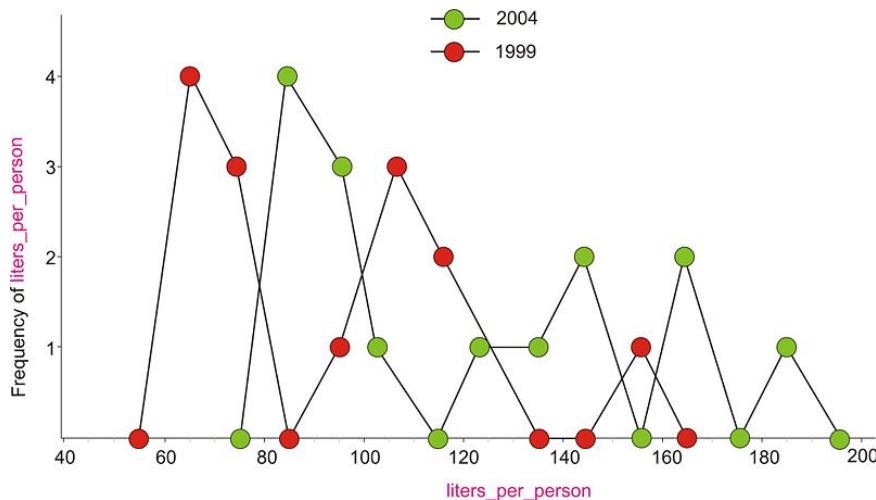


Here is the frequency polygon in finished form:



Frequency polygons are helpful in showing the general overall shape of a distribution of data. They can also be useful for comparing two sets of data. Imagine how confusing two histograms would look graphed on top of each other!

Example: It would be interesting to compare bottled water consumption in two different years. Two frequency polygons would help give an overall picture of how the years are similar, and how they are different. In the following graph, two frequency polygons, one representing 1999, and the other representing 2004, are overlaid. 1999 is in red, and 2004 is in green.



It appears there was a shift to the right in all the data, which is explained by realizing that all of the countries have significantly increased their consumption. The first peak in the lower-consuming countries is almost identical in the two frequency polygons, but it increased by 20 liters per person in 2004. In 1999, there was a middle peak, but that group shifted significantly to the right in 2004 (by between 40 and 60 liters per person). The frequency polygon is the first type of graph we have learned about that makes this type of comparison easier.

Cumulative Frequency Histograms and Ogive Plots

Very often, it is helpful to know how the data accumulate over the range of the distribution. To do this, we will add to our frequency table by including the cumulative frequency, which is how many of the data points are in all the classes up to and including a particular class.

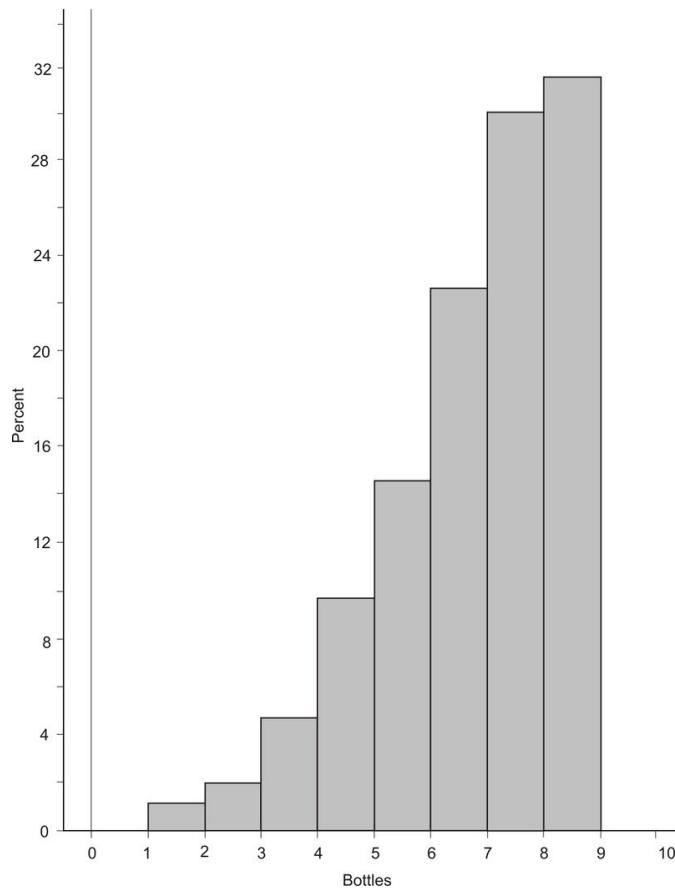
TABLE 2.5:

Number of Plastic Beverage Bottles per Week	Frequency	Cumulative Frequency
1	1	1
2	1	2
3	3	5
4	4	9
5	6	15
6	8	23
7	7	30
8	2	32

Figure: Cumulative Frequency Table for Bottle Data

For example, the cumulative frequency for 5 bottles per week is 15, because 15 students consumed 5 or fewer bottles per week. Notice that the cumulative frequency for the last class is the same as the total number of students in the data. This should always be the case.

If we drew a histogram of the cumulative frequencies, or a *cumulative frequency histogram*, it would look as follows:

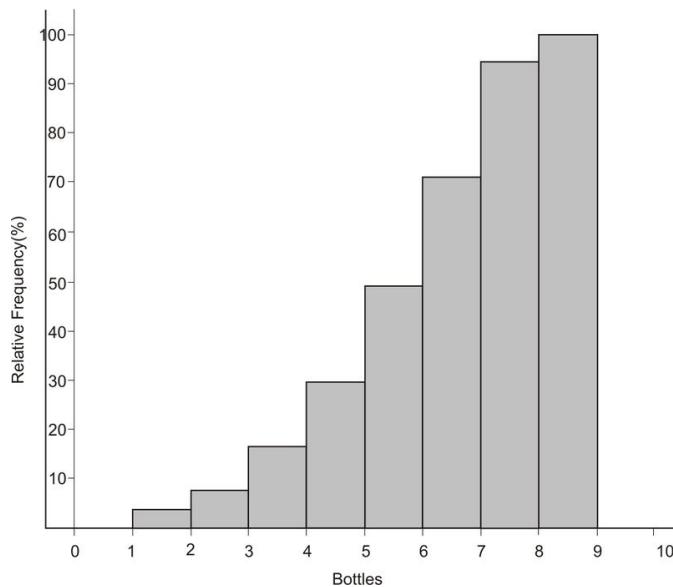


A *relative cumulative frequency histogram* would be the same, except that the vertical bars would represent the relative cumulative frequencies of the data:

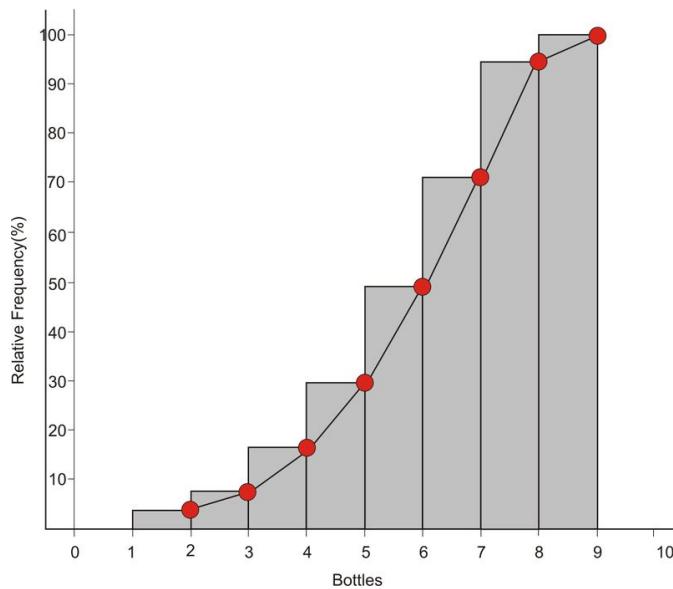
TABLE 2.6:

Number of Plastic Beverage Bottles per Week	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Frequency (%)
1	1	1	3.1	3.1
2	1	2	6.3	6.3
3	3	5	15.6	15.6
4	4	9	28.1	28.1
5	6	15	46.9	46.9
6	8	23	71.9	71.9
7	7	30	93.8	93.8
8	2	32	100	100

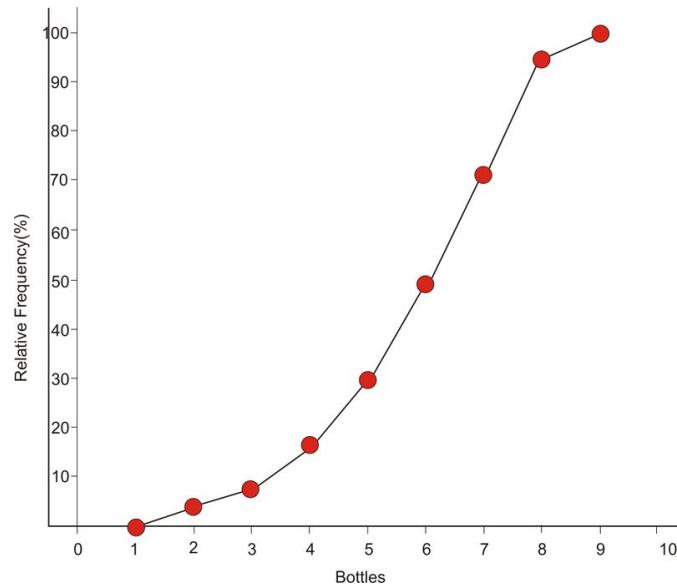
Figure: Relative Cumulative Frequency Table for Bottle Data



Remembering what we did with the frequency polygon, we can remove the bins to create a new type of plot. In the frequency polygon, we connected the midpoints of the bins. In a *relative cumulative frequency plot*, we use the point on the right side of each bin.



The reason for this should make a lot of sense: when we read this plot, each point should represent the percentage of the total data that is less than or equal to a particular value, just like in the frequency table. For example, the point that is plotted at 4 corresponds to 15.6%, because that is the percentage of the data that is less than or equal to 3. It does not include the 4's, because they are in the bin to the right of that point. This is why we plot a point at 1 on the horizontal axis and at 0% on the vertical axis. None of the data is lower than 1, and similarly, all of the data is below 9. Here is the final version of the plot:



This plot is commonly referred to as an **ogive plot**. The name ogive comes from a particular pointed arch originally present in Arabic architecture and later incorporated in Gothic cathedrals. Here is a picture of a cathedral in Ecuador with a close-up of an ogive-type arch:



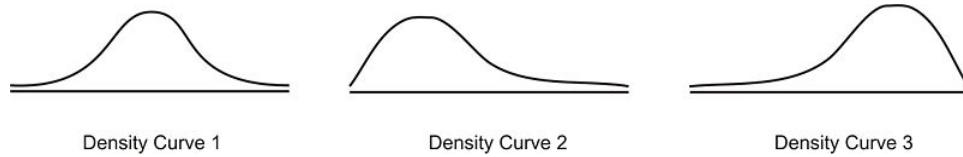
If a distribution is symmetric and mound shaped, then its ogive plot will look just like the shape of one half of such an arch.

Shape, Center, Spread

In the first chapter, we introduced measures of center and spread as important descriptors of a data set. The shape of a distribution of data is very important as well. Shape, center, and spread should always be your starting point when describing a data set.

Referring to our imaginary student poll on using plastic beverage containers, we notice that the data are spread out from 0 to 9. The graph for the data illustrates this concept, and the range quantifies it. Look back at the graph and notice that there is a large concentration of students in the 5, 6, and 7 region. This would lead us to believe that the center of this data set is somewhere in this area. We use the mean and/or median to measure central tendency, but it is also important that you *see* that the center of the distribution is near the large concentration of data. This is done with shape.

Shape is harder to describe with a single statistical measure, so we will describe it in less quantitative terms. A very important feature of this data set, as well as many that you will encounter, is that it has a single large concentration of data that appears like a mountain. A data set that is shaped in this way is typically referred to as *mound-shaped*. Mound-shaped data will usually look like one of the following three pictures:



Think of these graphs as frequency polygons that have been smoothed into curves. In statistics, we refer to these graphs as *density curves*. The most important feature of a density curve is symmetry. The first density curve above is *symmetric* and mound-shaped. Notice the second curve is mound-shaped, but the center of the data is concentrated on the left side of the distribution. The right side of the data is spread out across a wider area. This type of distribution is referred to as *skewed right*. It is the direction of the long, spread out section of data, called the *tail*, that determines the direction of the skewing. For example, in the 3rd curve, the left tail of the distribution is stretched out, so this distribution is *skewed left*. Our student bottle data set has this skewed-left shape.

Lesson Summary

A **frequency table** is useful to organize data into classes according to the number of occurrences, or frequency, of each class. **Relative frequency** shows the percentage of data in each class. A **histogram** is a graphical representation of a frequency table (either actual or relative frequency). A **frequency polygon** is created by plotting the midpoint of each bin at its frequency and connecting the points with line segments. Frequency polygons are useful for viewing the overall shape of a distribution of data, as well as comparing multiple data sets. For any distribution of data, you should always be able to describe the **shape, center, and spread**. A data set that is mound shaped can be classified as either symmetric or skewed. Distributions that are **skewed left** have the bulk of the data concentrated on the higher end of the distribution, and the lower end, or tail, of the distribution is spread out to the left. A **skewed-right** distribution has a large portion of the data concentrated in the lower values of the variable, with the tail spread out to the right. A relative cumulative frequency plot, or **ogive** plot, shows how the data accumulate across the different values of the variable.

Points to Consider

- What characteristics of a data set make it easier or harder to represent it using frequency tables, histograms, or frequency polygons?
- What characteristics of a data set make representing it using frequency tables, histograms, frequency polygons, or ogive plots more or less useful?
- What effects does the shape of a data set have on the statistical measures of center and spread?
- How do you determine the most appropriate classification to use for a frequency table or the bin width to use for a histogram?

Review Questions

1. Lois was gathering data on the plastic beverage bottle consumption habits of her classmates. She prepared a frequency table, but she now wants to create a variety of charts for displaying the data.

TABLE 2.7:

Plastic Used per Week	Bottles	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumula- tive Frequency
1	2				
2	2				
3	3				
4	2				
5	3				
6	7				
7	6				
8	1				

- (a) Complete the table above, including the relative frequency (round to the nearest tenth of a percent), the cumulative frequency, and the relative cumulative frequency.
- (b) Create a relative frequency histogram from your table.
- (c) Draw the corresponding frequency polygon.
- (d) Create the ogive plot.
- (e) Comment on the shape, center, and spread of the distribution displayed by the histogram. (Do not actually calculate any specific statistics).
- (f) Add up the relative frequency column. What is the total? What should it be? Why might the total not be what you would expect?
- (g) Use the ogive to determine the median, remembering that the median is the 50th percentile.
- (h) What does the steepest part of an ogive plot tell you about the distribution?

Short Answers to Review Questions:

1.a, 1.b, 1.c, 1.d (see detailed answers). (1) (e). Center: 5 or 6 Shape: unimodal and skewed left. Spread: 7 units (1)(f). 1 (1)(g) About 5.2 (1)(h) the category with the largest frequency.

2.2 Common Graphs and Data Plots

Learning Objectives

- Identify and translate data sets to and from a bar graph and a pie graph.
- Identify and translate data sets to and from a dot plot.
- Identify and translate data sets to and from a stem-and-leaf plot.
- Identify and translate data sets to and from a scatterplot.
- Identify graph distribution shapes as skewed or symmetric, and understand the basic implication of these shapes.
- Compare distributions of univariate data (shape, center, spread, and outliers).

Introduction

In this section, we will continue to investigate the different types of graphs that can be used to interpret a data set. In addition to a few more ways to represent single numerical variables, we will also study methods for displaying categorical variables. You will also be introduced to using a scatterplot and a line graph to show the relationship between two variables.

Categorical Variables: Bar Graphs and Pie Graphs

Example: E-Waste and Bar Graphs

We live in an age of unprecedented access to increasingly sophisticated and affordable personal technology. Cell phones, computers, and televisions now improve so rapidly that, while they may still be in working condition, the drive to make use of the latest technological breakthroughs leads many to discard usable electronic equipment. Much of that ends up in a landfill, where the chemicals from batteries and other electronics add toxins to the environment. Approximately 80% of the electronics discarded in the United States is also exported to third world countries, where it is disposed of under generally hazardous conditions by unprotected workers.¹ The following table shows the amount of tonnage of the most common types of electronic equipment discarded in the United States in 2005.

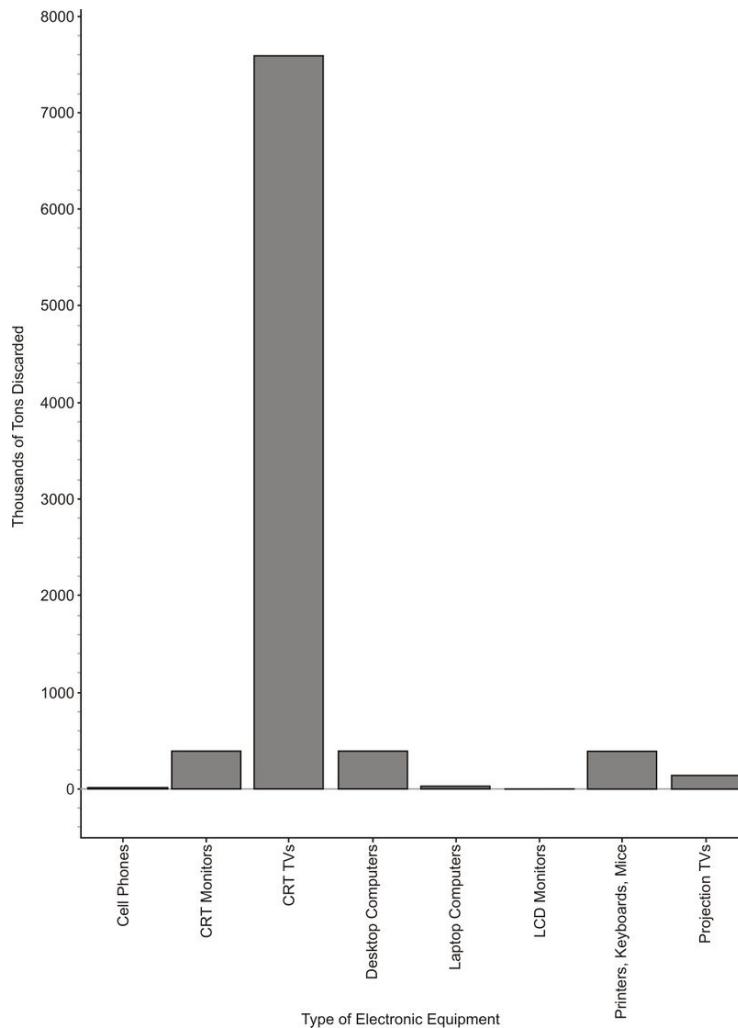
TABLE 2.8:

Electronic Equipment	Thousands of Tons Discarded
Cathode Ray Tube (CRT) TV's	7591.1
CRT Monitors	389.8
Printers, Keyboards, Mice	324.9
Desktop Computers	259.5
Laptop Computers	30.8
Projection TV's	132.8
Cell Phones	11.7
LCD Monitors	4.9

Figure: Electronics Discarded in the US (2005). *Source:* National Geographic, January 2008. Volume 213 No. 1,

pg 73.

The type of electronic equipment is a categorical variable, and therefore, this data can easily be represented using the *bar graph* below:



While this looks very similar to a histogram, the bars in a bar graph usually are separated slightly. The graph is just a series of disjoint categories.

Please note that discussions of shape, center, and spread have no meaning for a bar graph, and it is not, in fact, even appropriate to refer to this graph as a distribution. For example, some students misinterpret a graph like this by saying it is skewed right. If we rearranged the categories in a different order, the same data set could be made to look skewed left. Do not try to infer any of these concepts from a bar graph!

Pie Graphs

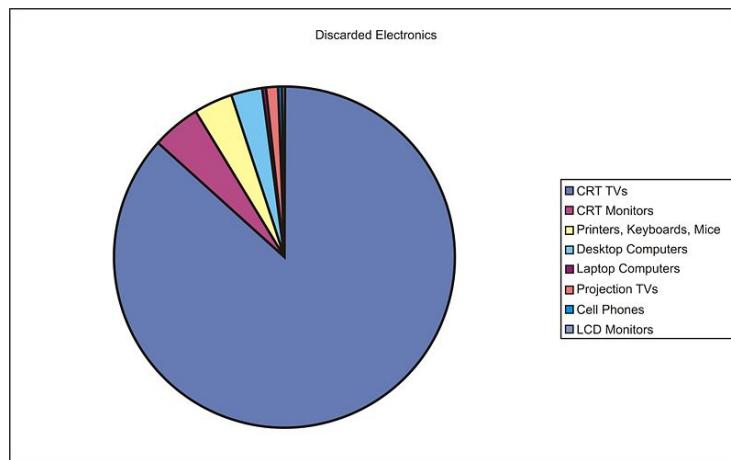
Usually, data that can be represented in a bar graph can also be shown using a *pie graph* (also commonly called a circle graph or pie chart). In this representation, we convert the count into a percentage so we can show each category relative to the total. Each percentage is then converted into a proportionate sector of the circle. To make this conversion, simply multiply the percentage by 360, which is the total number of degrees in a circle.

Here is a table with the percentages and the approximate angle measure of each sector:

TABLE 2.9:

Electronic Equipment	Thousands of Tons Discarded	Percentage of Total Discarded	Angle Measure of Circle Sector
Cathode Ray Tube (CRT) TV's	7591.1	86.8	312.5
CRT Monitors	389.8	4.5	16.2
Printers, Keyboards, Mice	324.9	3.7	13.4
Desktop Computers	259.5	3.0	10.7
Laptop Computers	30.8	0.4	1.3
Projection TV's	132.8	1.5	5.5
Cell Phones	11.7	0.1	0.5
LCD Monitors	4.9	~ 0	0.2

And here is the completed pie graph:



Displaying Univariate Data

Dot Plots

A *dot plot* is one of the simplest ways to represent numerical data. After choosing an appropriate scale on the axes, each data point is plotted as a single dot. Multiple points at the same value are stacked on top of each other using equal spacing to help convey the shape and center.

Example: The following is a data set representing the percentage of paper packaging manufactured from recycled materials for a select group of countries.

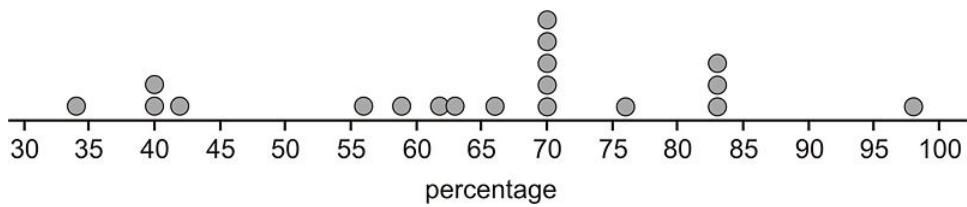
TABLE 2.10: Percentage of the paper packaging used in a country that is recycled. Source: National Geographic, January 2008. Volume 213 No. 1, pp. 86-87.

Country	% of Paper Packaging Recycled
Estonia	34
New Zealand	40
Poland	40
Cyprus	42
Portugal	56
United States	59
Italy	62

TABLE 2.10: (continued)

Country	% of Paper Packaging Recycled
Spain	63
Australia	66
Greece	70
Finland	70
Ireland	70
Netherlands	70
Sweden	76
France	76
Germany	83
Austria	83
Belgium	83
Japan	98

The dot plot for this data would look like this:



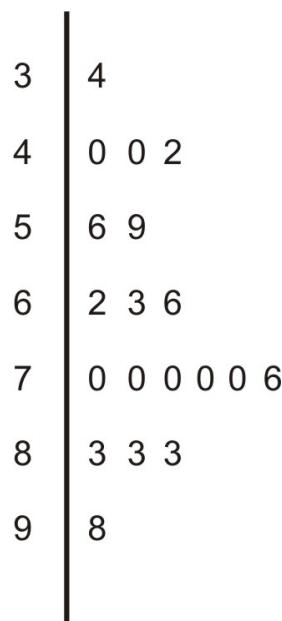
Notice that this data set is centered at a manufacturing rate for using recycled materials of between 65 and 70 percent. It is spread from 34% to 98%, and appears very roughly symmetric, perhaps even slightly skewed left. Dot plots have the advantage of showing all the data points and giving a quick and easy snapshot of the shape, center, and spread. Dot plots are not much help when there is little repetition in the data. They can also be very tedious if you are creating them by hand with large data sets, though computer software can make quick and easy work of creating dot plots from such data sets.

Stem-and-Leaf Plots

One of the shortcomings of dot plots is that they do not show the actual values of the data. You have to read or infer them from the graph. From the previous example, you might have been able to guess that the lowest value is 34%, but you would have to look in the data table itself to know for sure. A *stem-and-leaf plot* is a similar plot in which it is much easier to read the actual data values. In a stem-and-leaf plot, each data value is represented by two digits: the stem and the leaf. In this example, it makes sense to use the ten's digits for the stems and the one's digits for the leaves. The stems are on the left of a dividing line as follows:



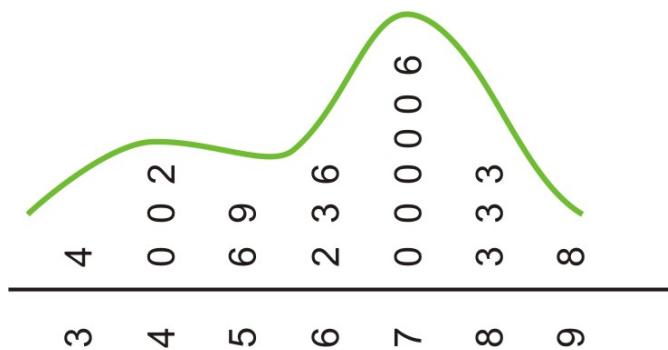
Once the stems are decided, the leaves representing the one's digits are listed in numerical order from left to right:



It is important to explain the meaning of the data in the plot for someone who is viewing it without seeing the original data. For example, you could place the following sentence at the bottom of the chart:

Note: 5|69 means 56% and 59% are the two values in the 50's.

If you could rotate this plot on its side, you would see the similarities with the dot plot. The general shape and center of the plot is easily found, and we know exactly what each point represents. This plot also shows the slight skewing to the left that we suspected from the dot plot. Stem plots can be difficult to create, depending on the numerical qualities and the spread of the data. If the data values contain more than two digits, you will need to remove some of the information by rounding. A data set that has large gaps between values can also make the stem plot hard to create and less useful when interpreting the data.



Back-to-Back Stem Plots

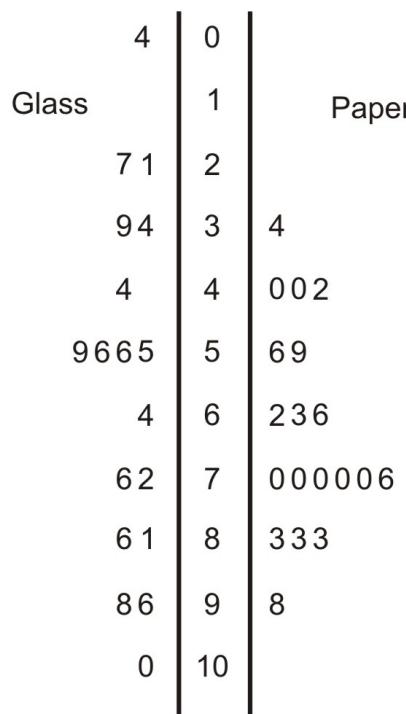
Stem plots can also be a useful tool for comparing two distributions when placed next to each other. These are commonly called *back-to-back stem plots*.

In the previous example, we looked at recycling in paper packaging. Here are the same countries and their percentages of recycled material used to manufacture glass packaging:

TABLE 2.11: Percentage of the glass packaging used in a country that is recycled. Source: National Geographic, January 2008. Volume 213 No. 1, pp. 86-87.

Country	% of Glass Packaging Recycled
Cyprus	4
United States	21
Poland	27
Greece	34
Portugal	39
Spain	41
Australia	44
Ireland	56
Italy	56
Finland	56
France	59
Estonia	64
New Zealand	72
Netherlands	76
Germany	81
Austria	86
Japan	96
Belgium	98
Sweden	100

In a back-to-back stem plot, one of the distributions simply works off the left side of the stems. In this case, the spread of the glass distribution is wider, so we will have to add a few extra stems. Even if there are no data values in a stem, you must include it to preserve the spacing, or you will not get an accurate picture of the shape and spread.



We have already mentioned that the spread was larger in the glass distribution, and it is easy to see this in the comparison plot. You can also see that the glass distribution is more symmetric and is centered lower (around the mid-50's), which seems to indicate that overall, these countries manufacture a smaller percentage of glass from recycled material than they do paper. It is interesting to note in this data set that Sweden actually imports glass from other countries for recycling, so its effective percentage is actually more than 100.

Displaying Bivariate Data

Scatterplots and Line Plots

Bivariate simply means two variables. All our previous work was with univariate, or single-variable data. The goal of examining *bivariate data* is usually to show some sort of relationship or association between the two variables.

Example: We have looked at recycling rates for paper packaging and glass. It would be interesting to see if there is a predictable relationship between the percentages of each material that a country recycles. Following is a data table that includes both percentages.

TABLE 2.12:

Country	% of Paper Packaging Recycled	% of Glass Packaging Recycled
Estonia	34	64
New Zealand	40	72
Poland	40	27
Cyprus	42	4
Portugal	56	39
United States	59	21
Italy	62	56
Spain	63	41
Australia	66	44
Greece	70	34
Finland	70	56

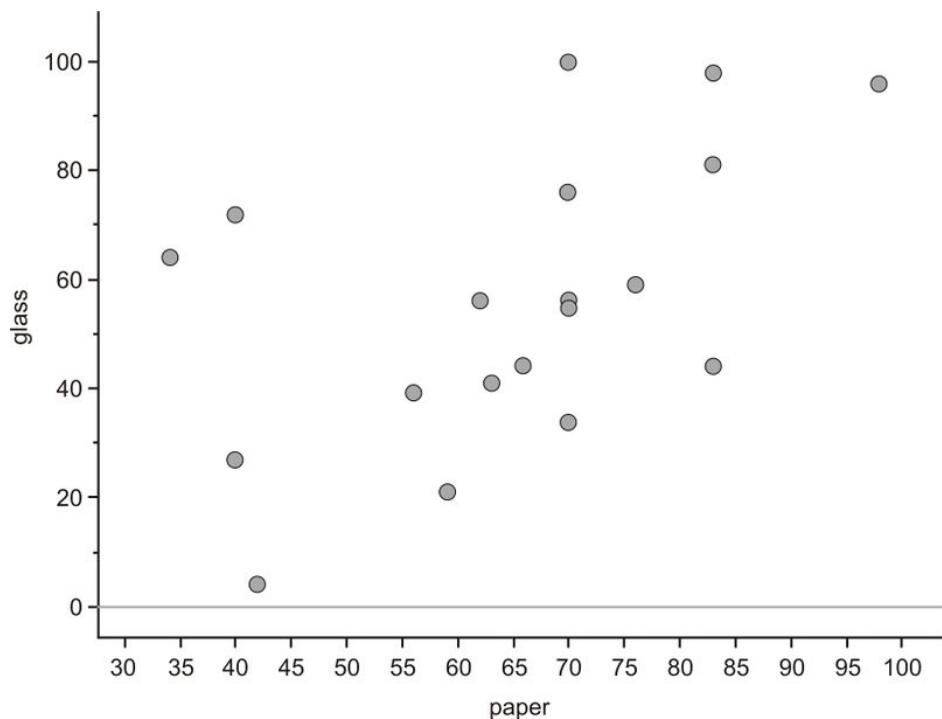
TABLE 2.12: (continued)

Country	% of Paper Packaging Recycled	% of Glass Packaging Recycled
Ireland	70	55
Netherlands	70	76
Sweden	70	100
France	76	59
Germany	83	81
Austria	83	44
Belgium	83	98
Japan	98	96

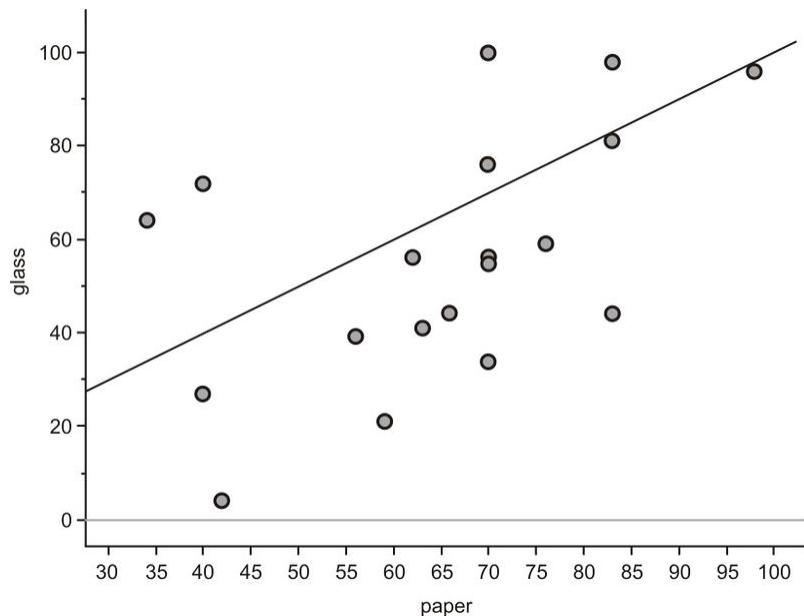
Figure: Paper and Glass Packaging Recycling Rates for 19 countries

Scatterplots

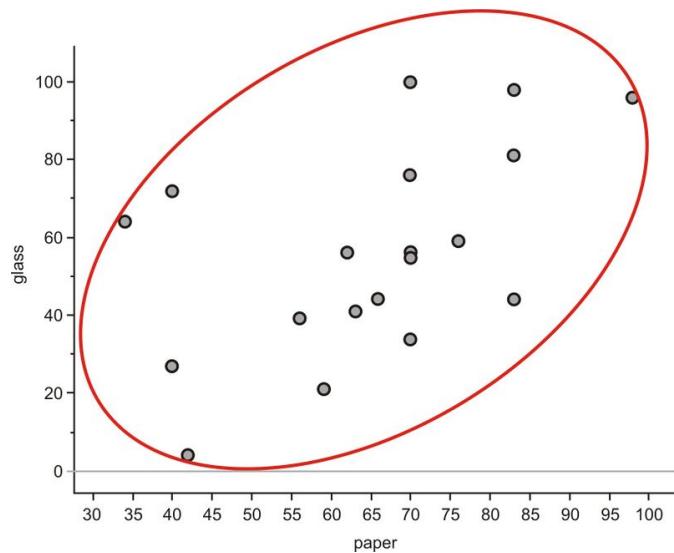
We will place the paper recycling rates on the horizontal axis and those for glass on the vertical axis. Next, we will plot a point that shows each country's rate of recycling for the two materials. This series of disconnected points is referred to as a *scatterplot*.



Recall that one of the things you saw from the stem-and-leaf plot is that, in general, a country's recycling rate for glass is lower than its paper recycling rate. On the next graph, we have plotted a line that represents the paper and glass recycling rates being equal. If all the countries had the same paper and glass recycling rates, each point in the scatterplot would be on the line. Because most of the points are actually below this line, you can see that the glass rate is lower than would be expected if they were similar.



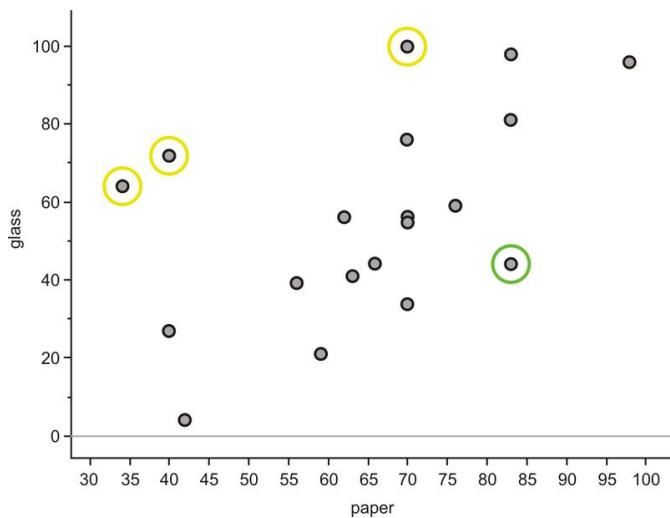
With univariate data, we initially characterize a data set by describing its shape, center, and spread. For bivariate data, we will also discuss three important characteristics: shape, direction, and strength. These characteristics will inform us about the association between the two variables. The easiest way to describe these traits for this scatterplot is to think of the data as a cloud. If you draw an ellipse around the data, the general trend is that the ellipse is rising from left to right.



Data that are oriented in this manner are said to have a *positive linear association*. That is, as one variable increases, the other variable also increases. In this example, it is mostly true that countries with higher paper recycling rates have higher glass recycling rates. Lines that rise in this direction have a positive slope, and lines that trend downward from left to right have a negative slope. If the ellipse cloud were trending down in this manner, we would say the data had a *negative linear association*. For example, we might expect this type of relationship if we graphed a country's glass recycling rate with the percentage of glass that ends up in a landfill. As the recycling rate increases, the landfill percentage would have to decrease.

The ellipse cloud also gives us some information about the strength of the linear association. If there were a strong linear relationship between the glass and paper recycling rates, the cloud of data would be much longer than it is wide. Long and narrow ellipses mean a strong linear association, while shorter and wider ones show a weaker linear

relationship. In this example, there are some countries for which the glass and paper recycling rates do not seem to be related.



New Zealand, Estonia, and Sweden (circled in yellow) have much lower paper recycling rates than their glass recycling rates, and Austria (circled in green) is an example of a country with a much lower glass recycling rate than its paper recycling rate. These data points are spread away from the rest of the data enough to make the ellipse much wider, weakening the association between the variables.

Lesson Summary

Bar graphs are used to represent categorical data in a manner that looks similar to, but is not the same as, a histogram. **Pie (or circle) graphs** are also useful ways to display categorical variables, especially when it is important to show how percentages of an entire data set fit into individual categories. A **dot plot** is a convenient way to represent univariate numerical data by plotting individual dots along a single number line to represent each value. They are especially useful in giving a quick impression of the shape, center, and spread of the data set, but are tedious to create by hand when dealing with large data sets.

Stem-and-leaf plots show similar information with the added benefit of showing the actual data values. Bivariate data can be represented using a **scatterplot** to show what, if any, association there is between the two variables. Usually one of the variables, the **explanatory (independent) variable**, can be identified as having an impact on the value of the other variable, the **response (dependent) variable**. The explanatory variable should be placed on the horizontal axis, and the response variable should be on the vertical axis. Each point is plotted individually on a scatterplot. If there is an association between the two variables, it can be identified as being strong if the points form a very distinct shape with little variation from that shape in the individual points. It can be identified as being weak if the points appear more randomly scattered. If the values of the response variable generally increase as the values of the explanatory variable increase, the data have a **positive association**. If the response variable generally decreases as the explanatory variable increases, the data have a **negative association**.

Points to Consider

- What characteristics of a data set make it easier or harder to represent using dot plots, stem-and-leaf plots, or histograms?
- Which plots are most useful to interpret the ideas of shape, center, and spread?
- What effects does the shape of a data set have on the statistical measures of center and spread?

Review Questions

- Computer equipment contains many elements and chemicals that are either hazardous, or potentially valuable when recycled. The following data set shows the contents of a typical desktop computer weighing approximately 27 kg. Some of the more hazardous substances, like Mercury, have been included in the 'other' category, because they occur in relatively small amounts that are still dangerous and toxic.

TABLE 2.13:

Material	Kilograms
Plastics	6.21
Lead	1.71
Aluminum	3.83
Iron	5.54
Copper	2.12
Tin	0.27
Zinc	0.60
Nickel	0.23
Barium	0.05
Other elements and chemicals	6.44

Figure: Weight of materials that make up the total weight of a typical desktop computer. *Source:* <http://dste.pudu.cherry.gov.in/envisnew/INDUSTRIAL%20SOLID%20WASTE.htm>

- Create a bar graph for this data.
- Complete the chart below to show the approximate percentage of the total weight for each material.

TABLE 2.14:

Material	Kilograms	Approximate Percentage of Total Weight
Plastics	6.21	
Lead	1.71	
Aluminum	3.83	
Iron	5.54	
Copper	2.12	
Tin	0.27	
Zinc	0.60	
Nickel	0.23	
Barium	0.05	
Other elements and chemicals	6.44	

- Create a circle graph for this data.
- The following table gives the percentages of municipal waste recycled by state in the United States, including the District of Columbia, in 1998. Data was not available for Idaho or Texas.

TABLE 2.15:

State	Percentage
Alabama	23
Alaska	7

TABLE 2.15: (continued)

State	Percentage
Arizona	18
Arkansas	36
California	30
Colorado	18
Connecticut	23
Delaware	31
District of Columbia	8
Florida	40
Georgia	33
Hawaii	25
Illinois	28
Indiana	23
Iowa	32
Kansas	11
Kentucky	28
Louisiana	14
Maine	41
Maryland	29
Massachusetts	33
Michigan	25
Minnesota	42
Mississippi	13
Missouri	33
Montana	5
Nebraska	27
Nevada	15
New Hampshire	25
New Jersey	45
New Mexico	12
New York	39
North Carolina	26
North Dakota	21
Ohio	19
Oklahoma	12
Oregon	28
Pennsylvania	26
Rhode Island	23
South Carolina	34
South Dakota	42
Tennessee	40
Utah	19
Vermont	30
Virginia	35
Washington	48
West Virginia	20
Wisconsin	36
Wyoming	5

Source: <http://www.zerowasteamerica.org/MunicipalWasteManagementReport1998.htm>

(a) Create a stem-and-leaf plot for the data. (Note that you will have a stem labeled 0 for the states with less than 10% of their waste that is recycled.)

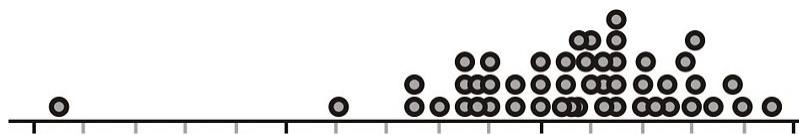
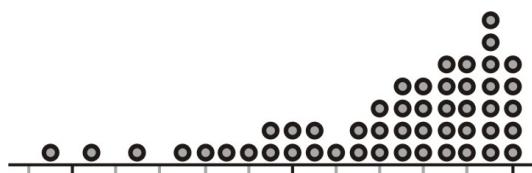
(b) Discuss the shape, center, and spread of this distribution.

(c) Use your stem-and-leaf plot to find the median percentage for this data.

3. Identify the important features of the shape (symmetric, skewed left or skewed right) of each of the following distributions. Are there outliers?

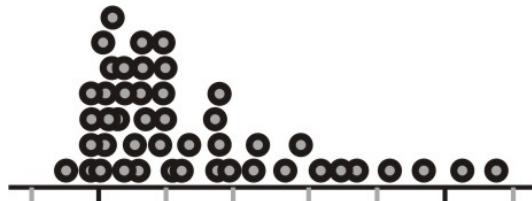
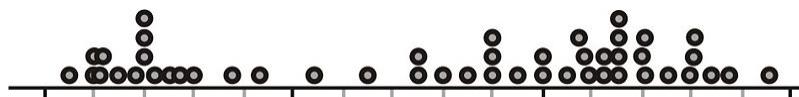
Dotplot A

Dotplot B



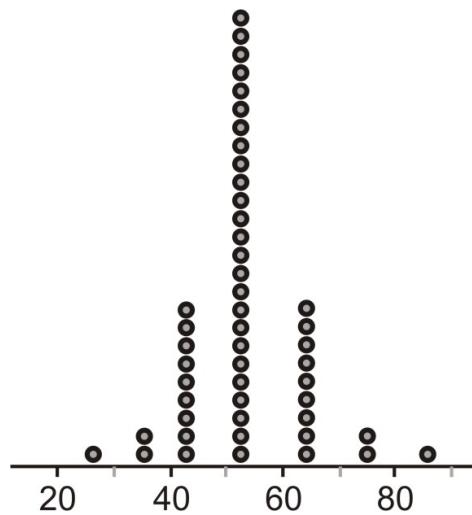
Dotplot C

Dotplot D

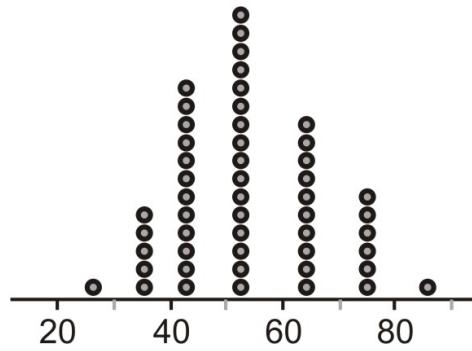


Questions 4-6 refer to the following dot plots:

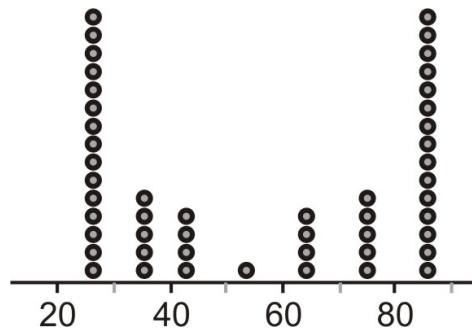
Dotplot A



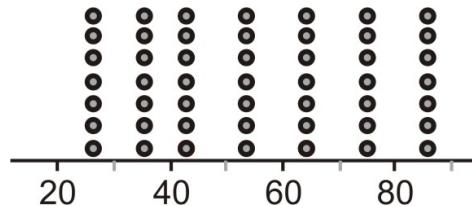
Dotplot B



Dotplot C



Dotplot D



4. How would you characterize the center(s) of these distributions?
5. Which of these distributions has the smallest standard deviation?
6. Which of these distributions has the largest standard deviation?
7. Carl has installed a solar panel on his roof. He measures the number of hours of daily sunshine and the resulting number of kilowatt hours of electricity generated by his solar panel. His results are provided in the chart below.

TABLE 2.16:

Hours of Sunshine	Kwh Generated
0	0
1	0.9
2	4
3	6.1
4	7.7
5	10.4
6	11.9

- (a) Identify the variables in this example, and specify which one is the explanatory variable and which one is the response variable.
- (b) How many Kwh were generated when there were 2 hours of sunshine?
- (c) Draw a scatterplot for this data.
- (d) Describe the direction and strength of the association between the two variables.

Short Answers to Review Questions. (1)(a), (1)(b), (1)(c) (see detailed answers). (2)(a) see detailed answers (2)(b) Shape: Mound-shaped and roughly symmetrical; Center: category 20% - 30% Spread: 43% (2)(c) Median = 26 (3)(a) Dotplot A not symmetric, skewed left; possible outliers. Dotplot B relatively symmetric, with a definite outlier. Dotplot C – bimodal, not symmetric, not skewed. Dotplot D - not symmetric, skewed right, possible outliers. (4) 52 (5) A (6) C (7)(a) Explanatory – hours of sunshine, Response – kwh generated. (7)(b) 4 (7)(c) See detailed answers (7)(d) Positive and strong

References

National Geographic, January 2008. Volume 213 No.1

¹http://www.etoxics.org/site/PageServer?pagename=svtc_global_ewaste_crisis

http://www.earth-policy.org/Updates/2006/Update51_data.htm

2.3 Box-and-Whisker Plots

Learning Objectives

- Calculate the values of the five-number summary.
- Draw and translate data sets to and from a box-and-whisker plot.
- Interpret the shape of a box-and-whisker plot.
- Describe the effects of changing units on summary measures.

Introduction

In this section, the box-and-whisker plot will be introduced.

The Five-Number Summary

The *five-number summary* is a numerical description of a data set comprised of the following measures (in order): minimum value, lower quartile, median, upper quartile, maximum value.

Example: The huge population growth in the western United States in recent years, along with a trend toward less annual rainfall in many areas and even drought conditions in others, has put tremendous strain on the water resources available now and the need to protect them in the years to come. Here is a listing of the reservoir capacities of the major water sources for Arizona:

TABLE 2.17:

Lake/Reservoir	% of Capacity
Salt River System	59
Lake Pleasant	49
Verde River System	33
San Carlos	9
Lyman Reservoir	3
Show Low Lake	51
Lake Havasu	98
Lake Mohave	85
Lake Mead	95
Lake Powell	89

Figure: Arizona Reservoir Capacity, 12 / 31 / 98. Source: <http://www.seattlecentral.edu/qelp/sets/008/008.html>

This data set was collected in 1998, and the water levels in many states have taken a dramatic turn for the worse. For example, Lake Powell is currently at less than 50% of capacity ¹.

Placing the data in order from smallest to largest gives the following:

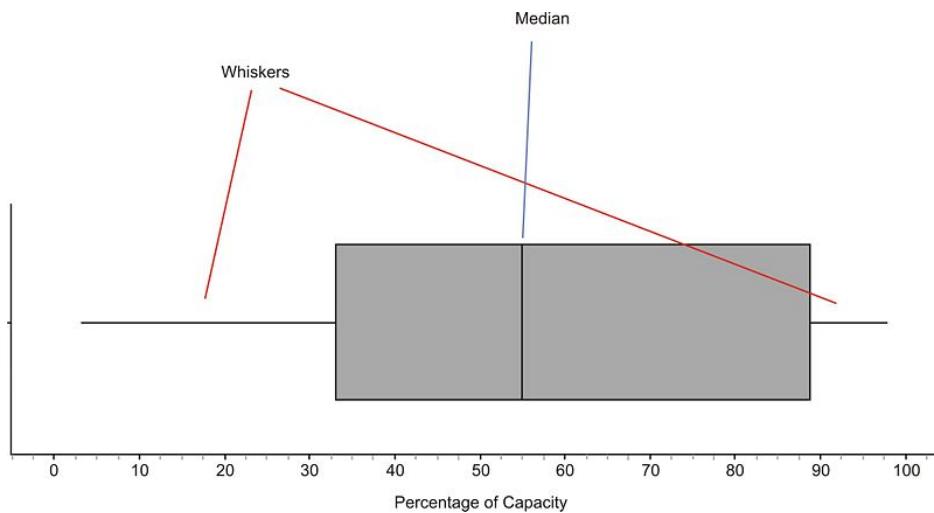
3, 9, 33, 49, 51, 59, 85, 89, 95, 98

Since there are 10 numbers, the median is the average of 51 and 59, which is 55. Recall that the lower quartile Q_1 is the 25th percentile, or where 25% of the data is below that value. In this data set, that number is 33. Also, the upper quartile Q_3 is 89. Therefore, the **five-number summary** is as shown:

$$\{3, 33, 55, 89, 98\}$$

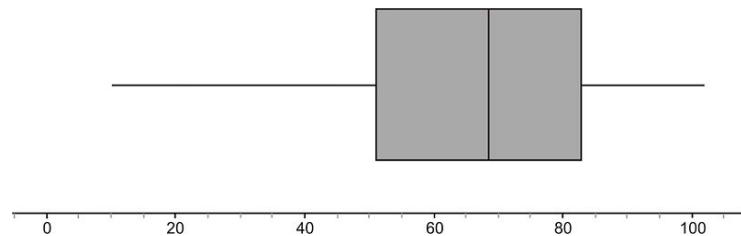
Box-and-Whisker Plots

A *box-and-whisker plot* is a very convenient and informative way to represent single-variable data. To create the 'box' part of the plot, draw a rectangle that extends from the lower quartile Q_1 to the upper quartile Q_3 . Draw a line through the interior of the rectangle at the median. Then connect the ends of the box to the minimum and maximum values using line segments to form the 'whiskers'. Here is the box plot for this data:

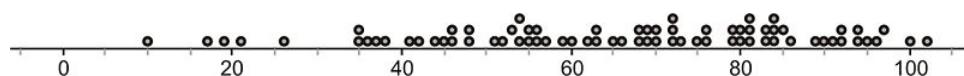


The plot divides the data into quarters; about 25% of the data appears in each section. You can also usually learn something about the shape of the distribution from the sections of the plot. If each of the four sections of the plot is about the same length, then the data will be symmetric. In this example, the different sections are not exactly the same length. The left whisker is slightly longer than the right, and the right half of the box is slightly longer than the left. We would most likely say that this distribution is moderately symmetric. In other words, there is roughly the same amount of data in each section. The different lengths of the sections tell us how the data are spread in each section. The numbers in the left whisker (lowest 25% of the data) are spread more widely than those in the right whisker.

Here is the box plot (as the name is sometimes shortened) for reservoirs and lakes in Colorado:

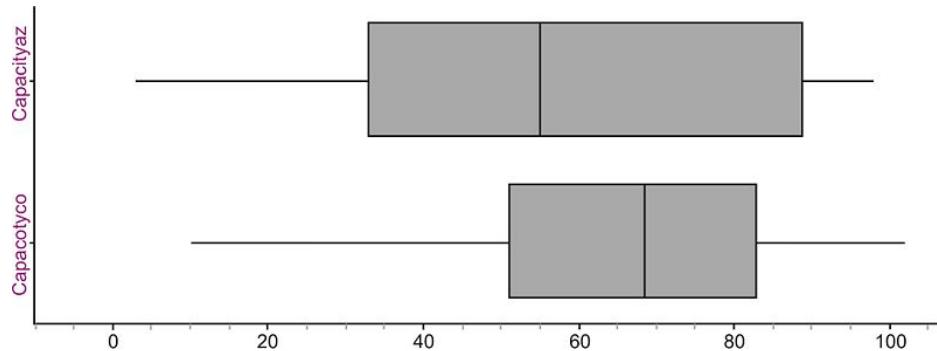


In this case, the third quarter of data (between the median and upper quartile), appears to be a bit more densely concentrated in a smaller area. The data values in the lower whisker also appear to be much more widely spread than in the other sections. Looking at the dot plot for the same data shows that this spread in the lower whisker gives the data a slightly skewed-left appearance (though it is still roughly symmetric).



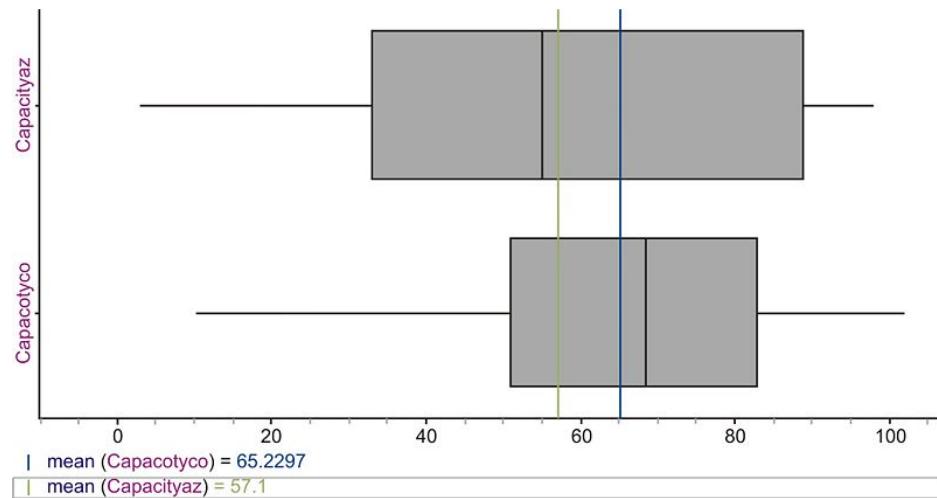
Comparing Multiple Box Plots

Box-and-whisker plots are often used to get a quick and efficient comparison of the general features of multiple data sets. In the previous example, we looked at data for both Arizona and Colorado. How do their reservoir capacities compare? You will often see multiple box plots either stacked on top of each other, or drawn side-by-side for easy comparison. Here are the two box plots:



The plots seem to be spread the same if we just look at the range, but with the box plots, we have an additional indicator of spread if we examine the length of the box (or interquartile range). This tells us how the middle 50% of the data is spread, and Arizona's data values appear to have a wider spread. The center of the Colorado data (as evidenced by the location of the median) is higher, which would tend to indicate that, in general, Arizona's capacities are lower. Recall that the median is a resistant measure of center, because it is not affected by outliers. The mean is not resistant, because it will be pulled toward outlying points. When a data set is skewed strongly in a particular direction, the mean will be pulled in the direction of the skewing, but the median will not be affected. For this reason, the median is a more appropriate measure of center to use for strongly skewed data.

Even though we wouldn't characterize either of these data sets as strongly skewed, this effect is still visible. Here are both distributions with the means plotted for each.



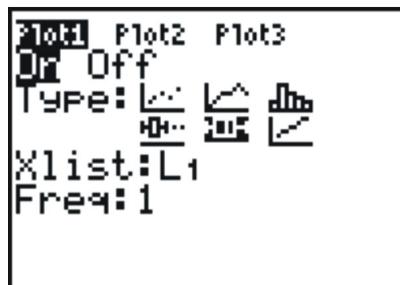
Notice that the long left whisker in the Colorado data causes the mean to be pulled toward the left, making it lower than the median. In the Arizona plot, you can see that the mean is slightly higher than the median, due to the slightly elongated right side of the box. If these data sets were perfectly symmetric, the mean would be equal to the median in each case.

Outliers in Box-and-Whisker Plots

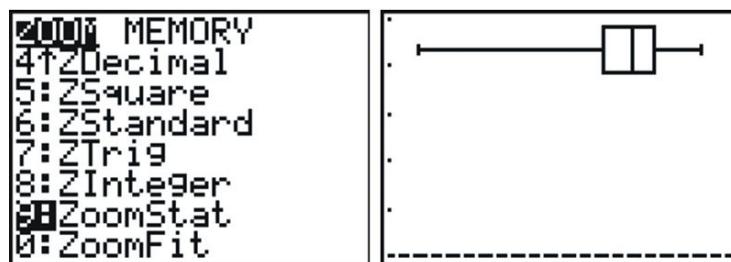
Here are the reservoir data for California (the names of the lakes and reservoirs have been omitted):

80, 83, 77, 95, 85, 74, 34, 68, 90, 82, 75

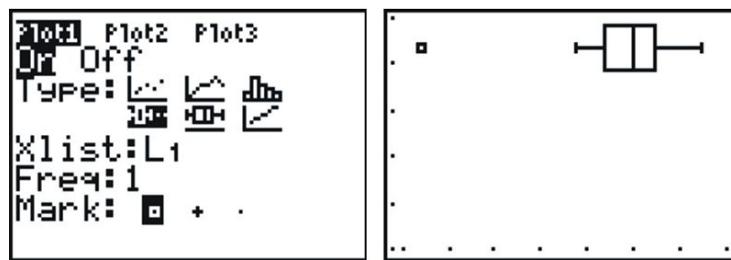
At first glance, the 34 should stand out. It appears as if this point is significantly different from the rest of the data. Let's use a graphing calculator to investigate this plot. Enter your data into a list as we have done before, and then choose a plot. Under 'Type', you will notice what looks like two different box and whisker plots. For now choose the second one (even though it appears on the second line, you must press the right arrow to select these plots).



Setting a window is not as important for a box plot, so we will use the calculator's ability to automatically scale a window to our data by pressing [ZOOM] and selecting '9:Zoom Stat'.

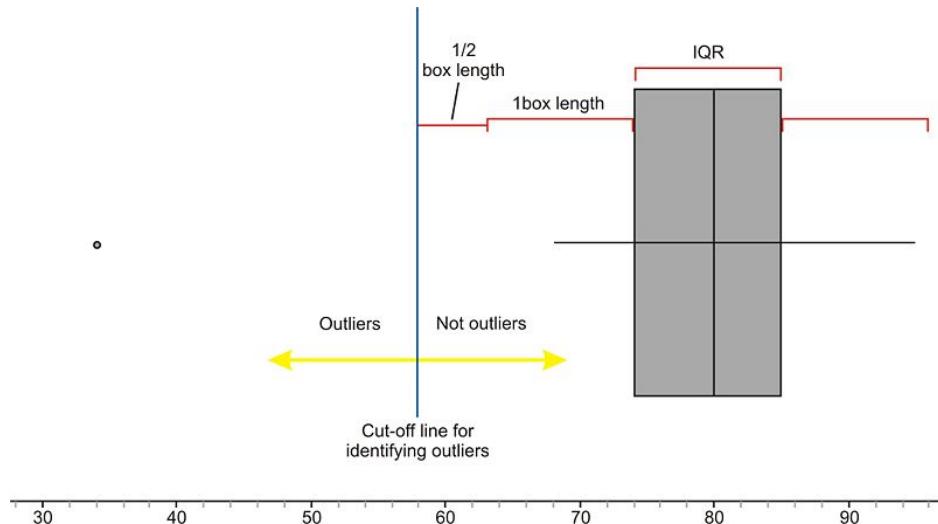


While box plots give us a nice summary of the important features of a distribution, we lose the ability to identify individual points. The left whisker is elongated, but if we did not have the data, we would not know if all the points in that section of the data were spread out, or if it were just the result of the one outlier. It is more typical to use a *modified box plot*. This box plot will show an outlier as a single, disconnected point and will stop the whisker at the previous point. Go back and change your plot to the first box plot option, which is the modified box plot, and then graph it.



Notice that without the outlier, the distribution is really roughly symmetric.

This data set had one obvious outlier, but when is a point far enough away to be called an outlier? We need a standard accepted practice for defining an outlier in a box plot. This rather arbitrary definition is that any point that is more than 1.5 times the interquartile range will be considered an outlier. Because the *IQR* is the same as the length of the box, any point that is more than one-and-a-half box lengths from either quartile is plotted as an outlier.



A common misconception of students is that you stop the whisker at this boundary line. In fact, the last point on the whisker that is not an outlier is where the whisker stops.

The calculations for determining the outlier in this case are as follows:

Lower Quartile Q₁: 74

Upper Quartile Q₃: 85

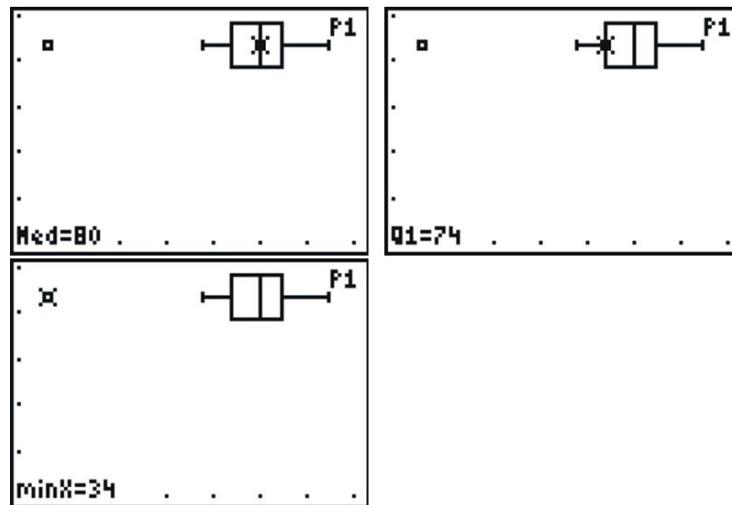
Interquartile range (IQR) : $85 - 74 = 11$

$1.5 * IQR = 16.5$

Cut-off for outliers in left whisker: $74 - 16.5 = 57.5$. Thus, any value less than 57.5 is considered an outlier. For the reservoir data, the value 34 is an outlier.

Cut-off for outliers in right whisker: $85 + 16.5 = 101.5$. Any value greater than 101.5 is considered an outlier. There is no outlier on the high side, because no data value is greater than 101.5.

If you press [TRACE] and use the left or right arrows, the calculator will trace the values of the five-number summary, as well as the outlier.



The Effects of Changing Units on Shape, Center, and Spread

In the previous lesson, we looked at data for the materials in a typical desktop computer.

TABLE 2.18:

Material	Kilograms
Plastics	6.21
Lead	1.71
Aluminum	3.83
Iron	5.54
Copper	2.12
Tin	0.27
Zinc	0.60
Nickel	0.23
Barium	0.05
Other elements and chemicals	6.44

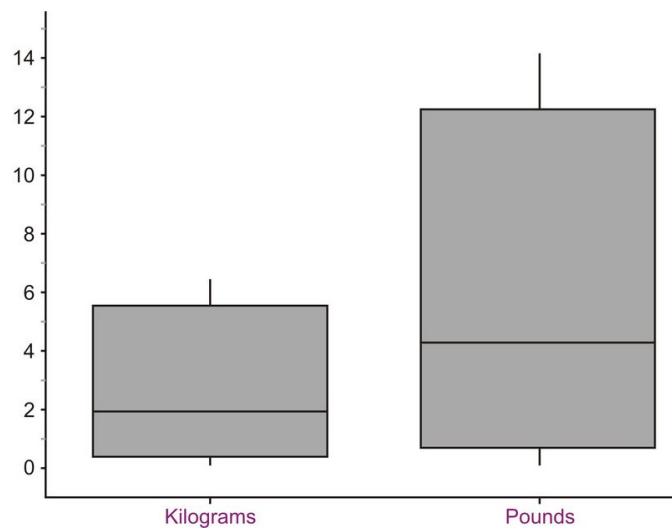
Here is the data set given in pounds. The weight of each in kilograms was multiplied by 2.2.

TABLE 2.19:

Material	Pounds
Plastics	13.7
Lead	3.8
Aluminum	8.4
Iron	12.2
Copper	4.7
Tin	0.6
Zinc	1.3
Nickel	0.5
Barium	0.1
Other elements and chemicals	14.2

When all values are **multiplied** by a factor of 2.2, the calculation of the mean is also multiplied by 2.2, so the center of the distribution would be increased by the same factor. Similarly, calculations of the range, interquartile range, and standard deviation will also be increased by the same factor. In other words, the center and the measures of spread will increase proportionally.

This should result in the graph maintaining the same shape, but being stretched out, or elongated. Here are the side-by-side box plots for both distributions showing the effects of changing units.



We can generalize the effect of **multiplying** all data values by a constant: **The mean and other measures of center will also be multiplied by the constant.** Additionally, even though it was not shown in the example, **the measures of spread, including the standard deviation, will also be multiplied by the constant.**

What if, instead, we **add** a constant value to every data value? Then, the entire distribution is shifted to the right (increased) by the amount of the constant. **The mean, and other measures of center will be increased by the constant.** However, **the standard deviation will not be affected at all** because the distribution was not stretched or compressed; it was simply moved.

Example: Consider IQ scores that have a mean of 100 and a standard deviation of 15. If we were to add 5 points to each IQ score, then an old score of 100 would now have a new value of 105. Thus, the new mean IQ score would now be 105. An old score of 115 (one standard deviation above the mean) would now have a value of 120. The new value of 120 is still one standard deviation above the mean, and the new standard deviation is $(120 - 105) = 15$, which is equal to the original standard deviation. Thus, the standard deviation was unchanged. Addition of a constant does change the value of the mean, but it doesn't affect the value of the standard deviation.

Lesson Summary

The **five-number summary** is a useful collection of statistical measures consisting of the following in ascending order: minimum, lower quartile, median, upper quartile, maximum. A **box-and-whisker plot**, or **boxplot**, is a graphical representation of the five-number summary showing a box bounded by the lower and upper quartiles and the median as a line in the box. The whiskers are line segments extended from the quartiles to the minimum and maximum values. Each whisker and section of the box contains approximately 25% of the data. The width of the box is the interquartile range, or *IQR*, and shows the spread of the middle 50% of the data. Box-and-whisker plots are effective at giving an overall impression of the shape, center, and spread of a data set. While an **outlier** is simply a point that is not typical of the rest of the data, there is an accepted definition of an outlier in the context of a box-and-whisker plot. Any point that is more than 1.5 times the length of the box (*IQR*) from either end of the box is considered to be an outlier.

When we change the units of a distribution by **multiplication**, the center and spread will both be affected. When we change the units of a distribution by **addition**, the center is affected, but the spread is not.

Points to Consider

- What characteristics of a data set make it easier or harder to represent it using dot plots, stem-and-leaf plots, histograms, and box-and-whisker plots?
- Which plots are most useful to interpret the ideas of shape, center, and spread?
- What effects do other transformations of the data have on the shape, center, and spread?

Review Questions

1. Here are the 1998 data on the percentage of capacity of reservoirs in Idaho.

70, 84, 62, 80, 75, 95, 69, 48, 76, 70, 45, 83, 58, 75, 85, 70
62, 64, 39, 68, 67, 35, 55, 93, 51, 67, 86, 58, 49, 47, 42, 75

- a. Find the five-number summary for this data set.
- b. Show all work to determine if there are true outliers according to the $1.5 * IQR$ rule.
- c. Create a box-and-whisker plot showing any outliers.
- d. Describe the shape, center, and spread of the distribution of reservoir capacities in Idaho in 1998.
- e. Based on your answer in part (d), how would you expect the mean to compare to the median? Calculate the mean to verify your expectation.

2. Here are the 1998 data on the percentage of capacity of reservoirs in Utah.

80, 46, 83, 75, 83, 90, 90, 72, 77, 4, 83, 105, 63, 87, 73, 84, 0, 70, 65, 96, 89, 78, 99, 104, 83, 81

- a. Find the five-number summary for this data set.
 - b. Show all work to determine if there are true outliers according to the $1.5 * IQR$ rule.
 - c. Create a box-and-whisker plot showing any outliers.
 - d. Describe the shape, center, and spread of the distribution of reservoir capacities in Utah in 1998.
 - e. Based on your answer in part (d) how would you expect the mean to compare to the median? Calculate the mean to verify your expectation.
3. Graph the box plots for Idaho and Utah on the same axes. Write a few statements comparing the water levels in Idaho and Utah by discussing the shape, center, and spread of the distributions.
4. If the median of a distribution is less than the mean, which of the following statements is the most correct?
- (a) a. The distribution is skewed left.
b. The distribution is skewed right.
c. There are outliers on the left side.
d. There are outliers on the right side.
e. (b) or (d) could be true.
5. A tennis coach weighs all the members of the girls' tennis team; the resulting mean and standard deviation are 130 pounds and 5 pounds, respectively. She later realizes that the scale she used was not calibrated properly, and each girl's weight was 2 pounds heavier than it should have been. Without reweighing each girl, what values should the coach use for the mean and standard deviation of these values?
- A. $\mu = 128$ pounds; $\sigma = 5$ pounds
 - B. $\mu = 128$ pounds; $\sigma = 3$ pounds
 - C. $\mu = 128$ pounds; σ cannot be determined
 - D. $\mu = 132$ pounds; $\sigma = 5$ pounds
 - E. $\mu = 132$ pounds; $\sigma = 3$ pounds

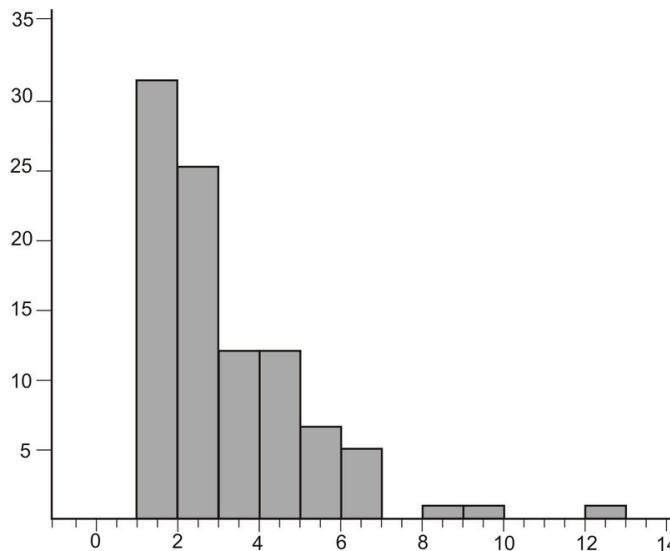
Answers to Review Questions. (1)(a) [35, 53, 67.5, 75.5, 95]. (1)(b) No outliers (1)(c) see detailed answers (1)(d) Shape is roughly symmetrical, slightly skewed left Center 67, Spread 60 (1)(e) Mean should be slightly less than the median (2)(a) [0, 72, 82, 89, 105]. (2)(b) Outliers are 0, 4, and 46. (2)(c) see detailed answers (2)(d) Shape: skewed left. Center: about 82 (median) Spread: range is 105. (2)(e) Mean should be less than median. Mean = 75.4 and Median = 82. (3)(a) see detailed answers. Shape: Idaho symmetric, Utah is highly left-skewed. Center: Idaho's center is lower than Utah's. Spread: Range for Idaho is 60%, range (without outliers) for Utah would be 59%. (4) E (5) A

References

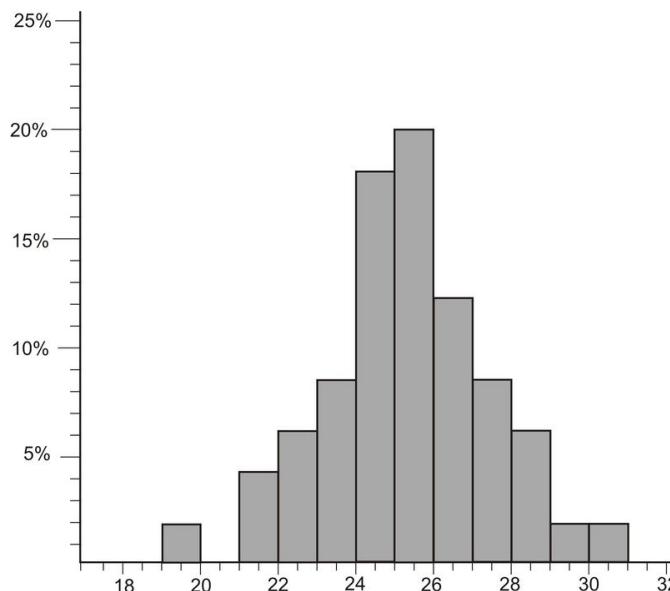
¹ Kunzig, Robert. Drying of the West. National Geographic, February 2008, Vol. 213, No. 2, Page 94.

Part One: Questions

1. Which of the following can be inferred from this histogram?

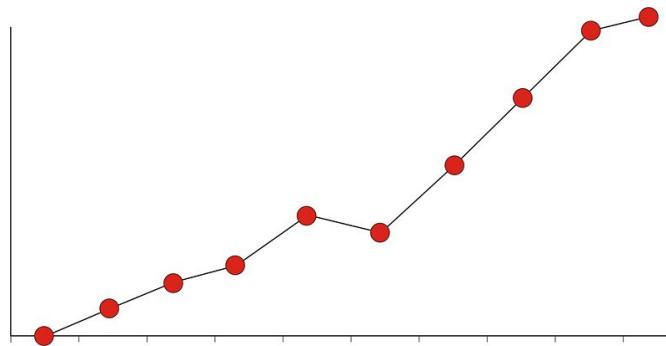


- a. The mode is 1.
 b. mean <median
 c. median <mean
 d. The distribution is skewed left.
 e. None of the above can be inferred from this histogram.
2. Sean was given the following relative frequency histogram to read.



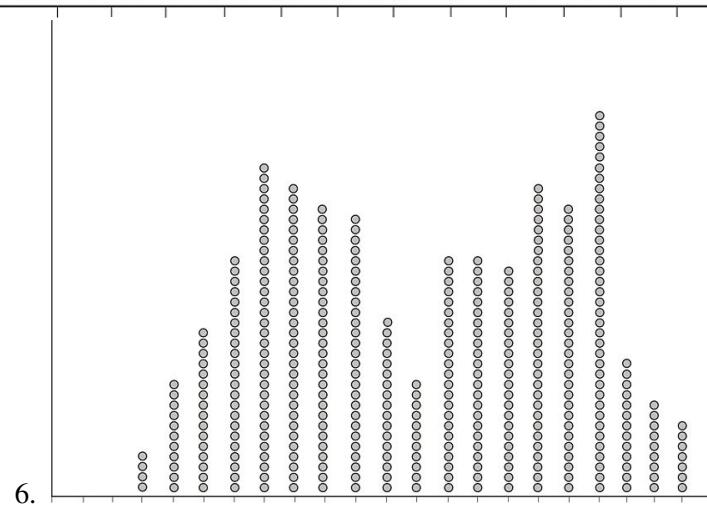
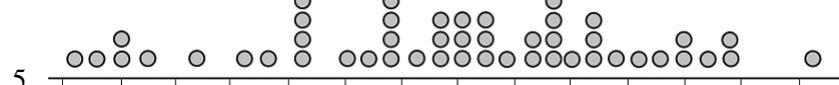
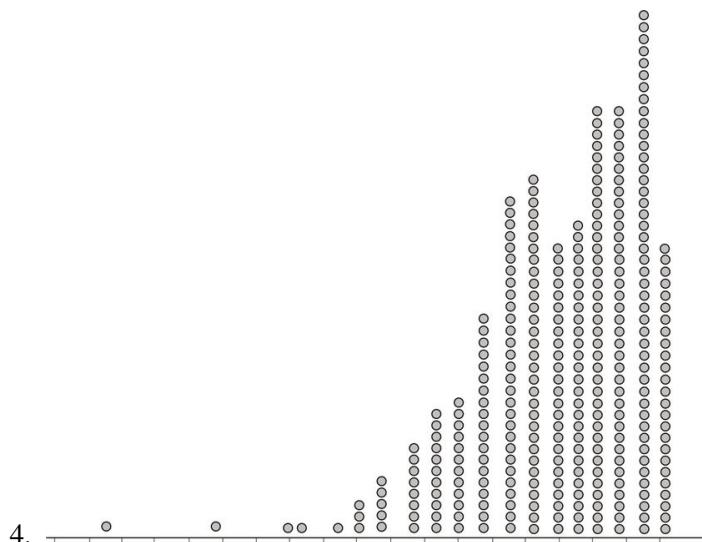
Unfortunately, the copier cut off the bin with the highest frequency. Which of the following could possibly be the relative frequency of the cut-off bin?

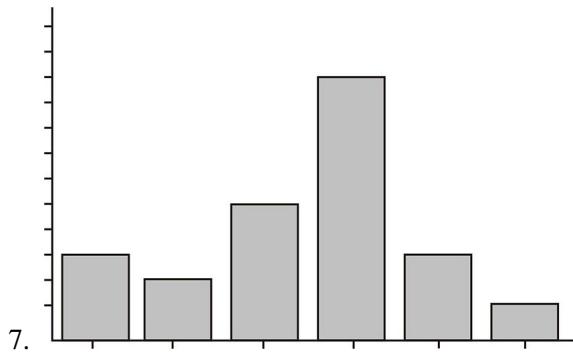
- a. 16
 b. 24
 c. 32
 d. 68
3. Tianna was given a graph for a homework question in her statistics class, but she forgot to label the graph or the axes and couldn't remember if it was a frequency polygon or an ogive plot. Here is her graph:



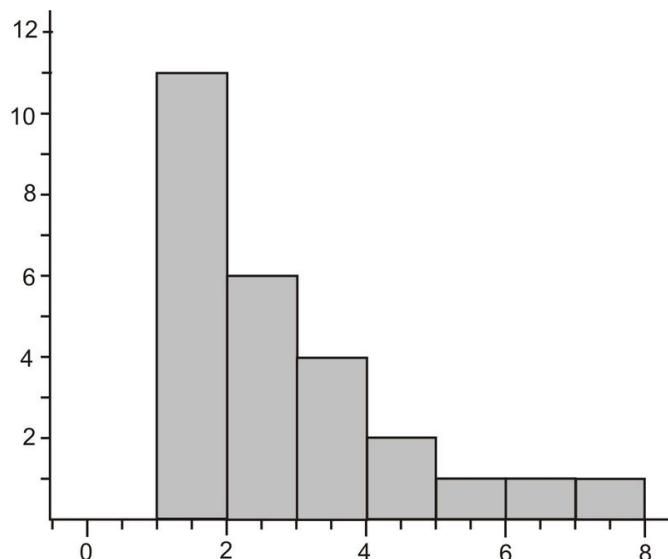
Identify which of the two graphs she has and briefly explain why.

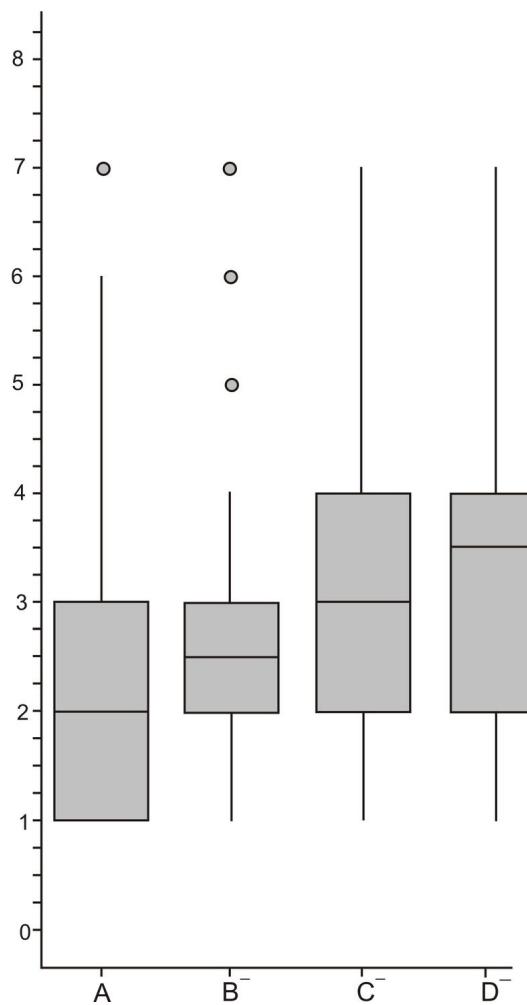
In questions 4-7, match the distribution with the choice of the correct real-world situation that best fits the graph.



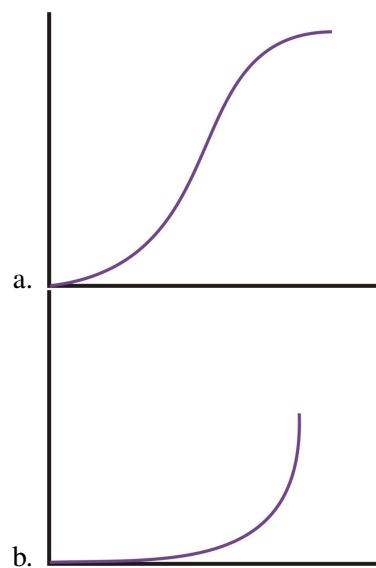


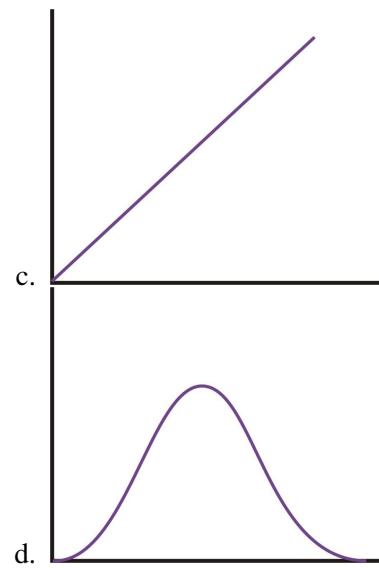
- a. Endy collected and graphed the heights of all the 12th grade students in his high school.
 - b. Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or piggy bank. She created a plot of the dates of the pennies.
 - c. Thamar asked her friends what their favorite movie was this year and graphed the results.
 - d. Jeno bought a large box of doughnut holes at the local pastry shop, weighed each of them, and then plotted their weights to the nearest tenth of a gram.
8. Which of the following box plots matches the histogram?



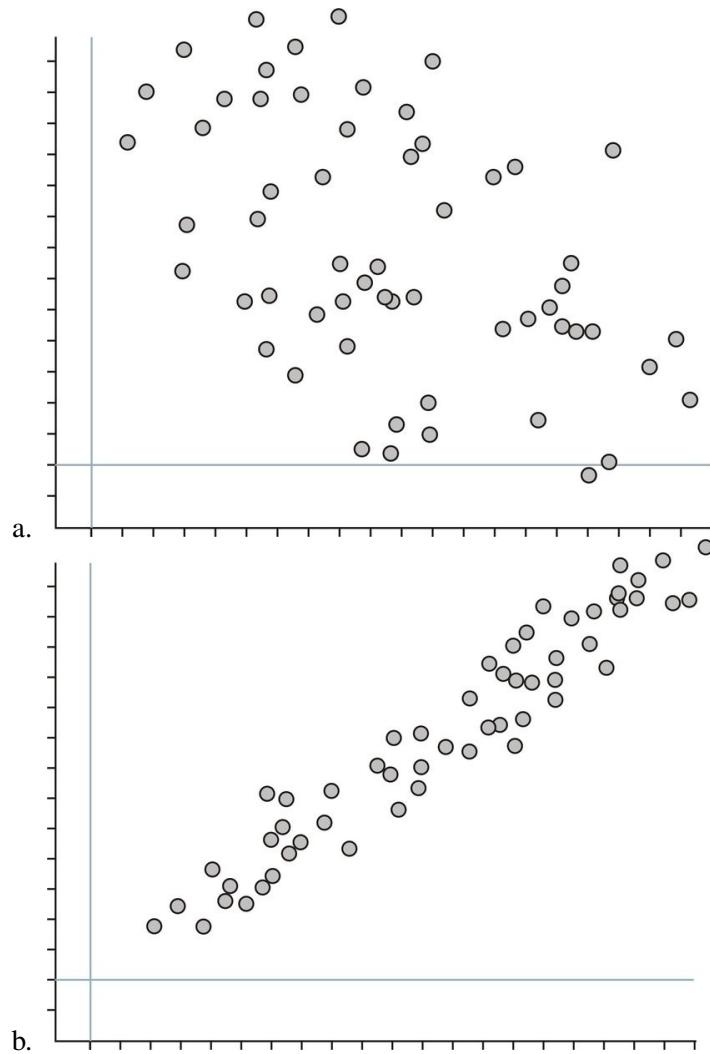


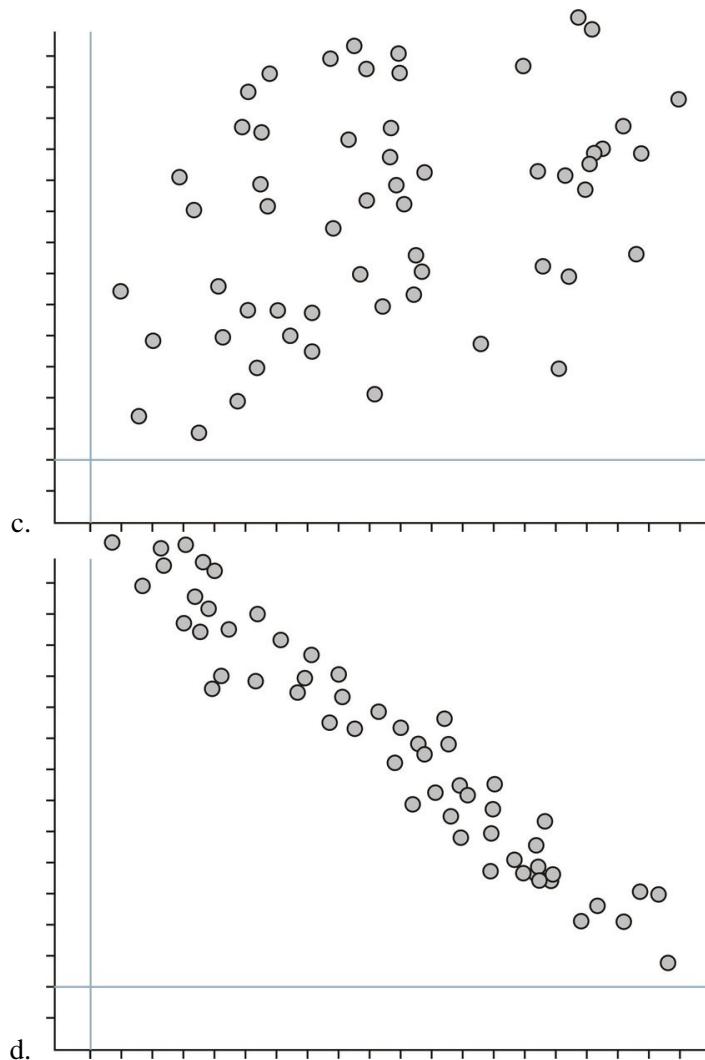
9. If a data set is roughly symmetric with no skewing or outliers, which of the following would be an appropriate sketch of the shape of the corresponding ogive plot?





10. Which of the following scatterplots shows a strong, negative association?





Part Two: Open-Ended Questions

- The Burj Dubai will become the world's tallest building when it is completed. It will be twice the height of the Empire State Building in New York.

TABLE 2.20:

Building	City	Height (ft)
Taipei 101	Tapei	1671
Shanghai World Financial Center	Shanghai	1614
Petronas Tower	Kuala Lumpur	1483
Sears Tower	Chicago	1451
Jin Mao Tower	Shanghai	1380
Two International Finance Center	Hong Kong	1362
CITIC Plaza	Guangzhou	1283
Shun Hing Square	Shenzen	1260
Empire State Building	New York	1250
Central Plaza	Hong Kong	1227
Bank of China Tower	Hong Kong	1205

TABLE 2.20: (continued)

Building	City	Height (ft)
Bank of America Tower	New York	1200
Emirates Office Tower	Dubai	1163
Tuntex Sky Tower	Kaohsiung	1140

The chart lists the 15 tallest buildings in the world (as of 12/2007).

- (a) Complete the table below, and draw an ogive plot of the resulting data.

TABLE 2.21:

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
--------------	------------------	---------------------------	-----------------------------	--------------------------------------

- (b) Use your ogive plot to approximate the median height for this data.

- (c) Use your ogive plot to approximate the upper and lower quartiles.

- (d) Find the 90th percentile for this data (i.e., the height that 90% of the data is less than).

2. Recent reports have called attention to an inexplicable collapse of the Chinook Salmon population in western rivers (see <http://www.nytimes.com/2008/03/17/science/earth/17salmon.html>). The following data tracks the fall salmon population in the Sacramento River from 1971 to 2007.

TABLE 2.22:

Year *	Adults	Jacks
1971-1975	164,947	37,409
1976-1980	154,059	29,117
1981-1985	169,034	45,464
1986-1990	182,815	35,021
1991-1995	158,485	28,639
1996	299,590	40,078
1997	342,876	38,352
1998	238,059	31,701
1999	395,942	37,567
2000	416,789	21,994
2001	546,056	33,439
2002	775,499	46,526
2003	521,636	29,806
2004	283,554	67,660
2005	394,007	18,115
2006	267,908	8,048
	87,966	1,897

Figure: Total Fall Salmon Escapement in the Sacramento River. *Source:* http://www.pccouncil.org/newsreleases/Sac_to_adult_and_jack_escapingethru%202007.pdf

During the years from 1971 to 1995, only 5-year averages are available.

In case you are not up on your salmon facts, there are two terms in this chart that may be unfamiliar. Fish escapement refers to the number of fish who escape the hazards of the open ocean and return to their freshwater streams and

rivers to spawn. A Jack salmon is a fish that returns to spawn before reaching full adulthood.

(a) Create one line graph that shows both the adult and jack populations for these years. The data from 1971 to 1995 represent the five-year averages. Devise an appropriate method for displaying this on your line plot while maintaining consistency.

(b) Write at least two complete sentences that explain what this graph tells you about the change in the salmon population over time.

3. The following data set about Galapagos land area was used in the first chapter.

TABLE 2.23:

Island	Approximate Area (sq. km)
Baltra	8
Darwin	1.1
Españaola	60
Fernandina	642
Floreana	173
Genovesa	14
Isabela	4640
Marchena	130
North Seymour	1.9
Pinta	60
Pinzón	18
Rabida	4.9
San Cristóbal	558
Santa Cruz	986
Santa Fe	24
Santiago	585
South Plaza	0.13
Wolf	1.3

Figure: Land Area of Major Islands in the Galapagos Archipelago. *Source:* http://en.wikipedia.org/wiki/Gal%C3%A1pagos_Islands

(a) Choose two methods for representing this data, one categorical, and one numerical, and draw the plot using your chosen method.

(b) Write a few sentences commenting on the shape, spread, and center of the distribution in the context of the original data. You may use summary statistics to back up your statements.

4. Investigation: The National Weather Service maintains a vast array of data on a variety of topics. Go to: <http://lwf.ncdc.noaa.gov/oa/climate/online/ccd/snowfall.html>. You will find records for the mean snowfall for various cities across the US.
- Create a back-to-back stem-and-leaf plot for all the cities located in each of two geographic regions. (Use the simplistic breakdown found at <http://library.thinkquest.org/4552/> to classify the states by region.)
 - Write a few sentences that compare the two distributions, commenting on the shape, spread, and center in the context of the original data. You may use summary statistics to back up your statements.

Keywords

Back-to-back stem plots

Stem plots can also be a useful tool for comparing two distributions when placed next to each other. These are commonly called *back-to-back stem plots*.

Bar graph

the bars in a bar graph usually are separated slightly. The graph is just a series of disjoint categories.

Bias

The systematic error in sampling is called *bias*.

Bivariate data

The goal of examining *bivariate data* is usually to show some sort of relationship or association between the two variables.

Box-and-whisker plot

A *box-and-whisker plot* is a very convenient and informative way to represent single-variable data.

Cumulative frequency histogram

A *relative cumulative frequency histogram* would be the same, except that the vertical bars would represent the relative cumulative frequencies of the data

Density curves

The most important feature of a density curve is symmetry. The first density curve above is *symmetric* and mound-shaped.

Dot plot

A *dot plot* is one of the simplest ways to represent numerical data.

Explanatory variable

the time in years is considered the *explanatory variable*, or independent variable.

Five-number summary

The *five-number summary* is a numerical description of a data set comprised of the following measures (in order): minimum value, lower quartile, median, upper quartile, maximum value.

Frequency polygon

A *frequency polygon* is similar to a histogram, but instead of using bins, a polygon is created by plotting the frequencies and connecting those points with a series of line segments.

Frequency tables

to create meaningful and useful categories for a frequency table.

Histogram

A histogram is a graphical representation of a frequency table (either actual or relative frequency).

Modified box plot

This box plot will show an outlier as a single, disconnected point and will stop the whisker at the previous point.

Mound-shaped

it has a single large concentration of data that appears like a mountain. A data set that is shaped in this way is typically referred to as *mound-shaped*.

Negative linear association

If the ellipse cloud were trending down in this manner, we would say the data had a *negative linear association*.

Ogive plot

This plot is commonly referred to as an *ogive plot*.

Pie graph

data that can be represented in a bar graph can also be shown using a *pie graph* (also commonly called a circle graph or pie chart).

Positive linear association

Data that are oriented in this manner are said to have a *positive linear association*. That is, as one variable increases, the other variable also increases.

Relative cumulative frequency histogram

it is helpful to know how the data accumulate over the range of the distribution.

Relative cumulative frequency plot

In a *relative cumulative frequency plot*, we use the point on the right side of each bin.

Relative frequency histogram

A *relative frequency histogram* is just like a regular histogram, but instead of labeling the frequencies on the vertical axis, we use the percentage of the total data that is present in that bin.

Response variable

can be identified as having an impact on the value of the other variable, the response (dependent) variable.

Scatterplot

Bivariate data can be represented using a scatterplot to show what, if any, association there is between the two variables.

Skewed left

the left tail of the distribution is stretched out, so this distribution is *skewed left*.

Skewed right

The right side of the data is spread out across a wider area. This type of distribution is referred to as *skewed right*.

Stem-and-leaf plot

A *stem-and-leaf plot* is a similar plot in which it is much easier to read the actual data values.

Symmetric

A data set that is mound shaped can be classified as either symmetric or skewed.

Tail

It is the direction of the long, spread out section of data, called the *tail*.

CHAPTER**3****An Introduction to Probability****Chapter Outline**

-
- 3.1 EVENTS, SAMPLE SPACES, AND PROBABILITY**
 - 3.2 COMPOUND EVENTS**
 - 3.3 THE COMPLEMENT OF AN EVENT**
 - 3.4 CONDITIONAL PROBABILITY**
 - 3.5 ADDITION AND MULTIPLICATION RULES**
 - 3.6 BASIC COUNTING RULES**
 - 3.7 REFERENCES**
-

3.1 Events, Sample Spaces, and Probability

Learning Objectives

- List simple events and sample spaces.

Introduction

The concept of probability plays an important role in our daily lives. Assume you have an opportunity to invest some money in a software company. Suppose you know that the company's records indicate that in the past five years, its profits have been consistently decreasing. Would you still invest your money in it? Do you think the chances are good for the company in the future?

Here is another illustration. Suppose that you are playing a game that involves tossing a single die. Assume that you have already tossed it 10 times, and every time the outcome was the same, a 2. What is your prediction of the eleventh toss? Would you be willing to bet \$100 that you will not get a 2 on the next toss? Do you think the die is loaded?

Notice that the decision concerning a successful investment in the software company and the decision of whether or not to bet \$100 on the next outcome of the die are both based on probabilities of certain sample results. Namely, the software company's profits have been declining for the past five years, and the outcome of rolling a 2 ten times in a row seems strange. From these sample results, we might conclude that we are not going to invest our money in the software company or bet on this die. In this lesson, you will learn mathematical ideas and tools that can help you understand such situations.

Events, Sample Spaces, and Probability

An *event* is something that occurs, or happens. For example, flipping a coin is an event, and so is walking in the park and passing by a bench. Anything that could possibly happen is an event.

Every event has one or more possible outcomes. While tossing a coin is an event, getting tails is the outcome of that event. Likewise, while walking in the park is an event, finding your friend sitting on the bench is an outcome of that event.

Suppose a coin is tossed once. There are two possible outcomes, either heads, H , or tails, T . Notice that if the experiment is conducted only once, you will observe only one of the two possible outcomes. An *experiment* is the process of taking a measurement or making an observation. These individual outcomes for an experiment are each called *simple events*.

Example: A die has six possible outcomes: 1, 2, 3, 4, 5, or 6. When we toss it once, only one of the six outcomes of this experiment will occur. The one that does occur is called a simple event.

Example: Suppose that two pennies are tossed simultaneously. We could have both pennies land heads up (which we write as HH), or the first penny could land heads up and the second one tails up (which we write as HT), etc. We will see that there are four possible outcomes for each toss, which are HH, HT, TH , and TT . The table below shows all the possible outcomes.

	<i>H</i>		<i>T</i>
<i>H</i>	<i>HH</i>		<i>HT</i>
<i>T</i>	<i>TH</i>		<i>TT</i>

Figure: The possible outcomes of flipping two coins.

What we have accomplished so far is a listing of all the possible simple events of an experiment. This collection is called the ***sample space*** of the experiment.

The sample space is the set of all possible outcomes of an experiment, or the collection of all the possible simple events of an experiment. We will denote a sample space by S .

Example: We want to determine the sample space of throwing a die and the sample space of tossing a coin.

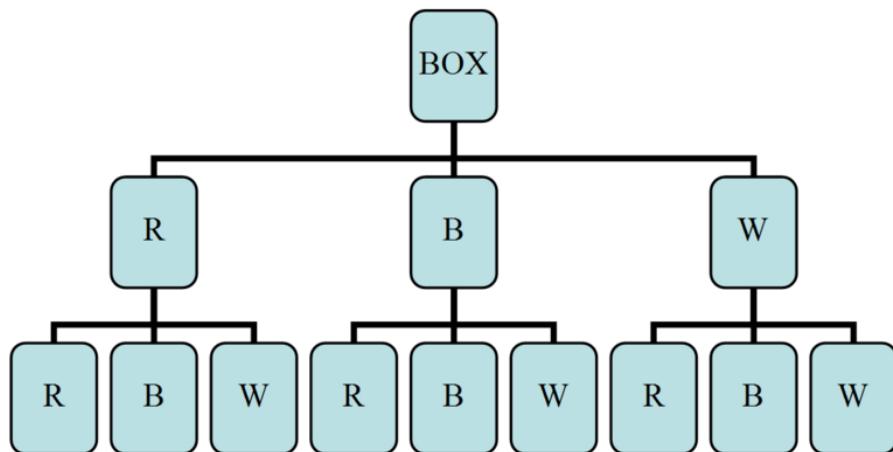
Solution: As we know, there are 6 possible outcomes for throwing a die. We may get 1, 2, 3, 4, 5, or 6, so we write the sample space as the set of all possible outcomes:

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

Similarly, the sample space of tossing a coin is either heads, H , or tails, T , so we write $S = \{H, T\}$.

Example: Suppose a box contains three balls, one red, one blue, and one white. One ball is selected, its color is observed, and then the ball is placed back in the box. The balls are scrambled, and again, a ball is selected and its color is observed. What is the sample space of the experiment?

It is probably best if we draw a *tree diagram* to illustrate all the possible selections.



As you can see from the tree diagram, it is possible that you will get the red ball, R , on the first drawing and then another red one on the second, RR . You can also get a red one on the first and a blue on the second, and so on. From the tree diagram above, we can see that the sample space is as follows:

$$S = \{RR, RB, RW, BR, BB, BW, WR, WB, WW\}$$

Each pair in the set above gives the first and second drawings, respectively. That is, RW is different from WR .

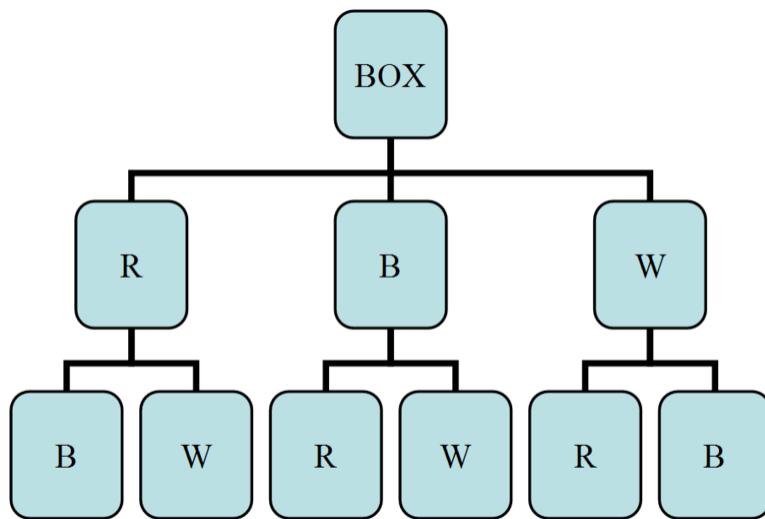
We can also represent all the possible drawings by a table or a matrix:

	<i>R</i>	<i>B</i>	<i>W</i>
<i>R</i>	<i>RR</i>	<i>RB</i>	<i>RW</i>
<i>B</i>	<i>BR</i>	<i>BB</i>	<i>BW</i>
<i>W</i>	<i>WR</i>	<i>WB</i>	<i>WW</i>

Figure: Table representing the possible outcomes diagrammed in the previous figure. The first column represents the first drawing, and the first row represents the second drawing.

Example: Consider the same experiment as in the last example. This time we will draw one ball and record its color, but we will not place it back into the box. We will then select another ball from the box and record its color. What is the sample space in this case?

Solution: The tree diagram below illustrates this case:



You can clearly see that when we draw, say, a red ball, the blue and white balls will remain. So on the second selection, we will either get a blue or a white ball. The sample space in this case is as shown:

$$S = \{RB, RW, BR, BW, WR, WB\}$$

Now let us return to the concept of probability and relate it to the concepts of sample space and simple events. If you toss a fair coin, the chance of getting tails, T , is the same as the chance of getting heads, H . Thus, we say that the probability of observing heads is 0.5, and the probability of observing tails is also 0.5. The probability, P , of an outcome, A , always falls somewhere between two extremes: 0, which means the outcome is an impossible event, and 1, which means the outcome is guaranteed to happen. Most outcomes have probabilities somewhere in-between.

Property 1: $0 \leq P(A) \leq 1$, for any event, A .

The probability of an event, A , ranges from 0 (impossible) to 1 (certain).

In addition, the probabilities of all possible simple outcomes of an event must add up to 1. This 1 represents certainty that one of the outcomes must happen. For example, tossing a coin will produce either heads or tails. Each of these two outcomes has a probability of 0.5. This means that the total probability of the coin landing either heads or tails is $0.5 + 0.5 = 1$. That is, we know that if we toss a coin, we are certain to get heads or tails.

Property 2: $\sum P(A) = 1$ when summed over all possible simple outcomes.

The sum of the probabilities of all possible outcomes must add up to 1. (Note that the symbol Σ means to "sum up" or add all the probabilities.)

Notice that tossing a coin or throwing a die results in outcomes that are all equally probable. That is, each outcome has the same probability as all the other outcomes in the same sample space. Getting heads or tails when tossing a coin produces an equal probability for each outcome, 0.5. Throwing a die has 6 possible outcomes, each also having the same probability, $\frac{1}{6}$. We refer to this kind of probability as classical probability. *Classical probability* is defined to be the ratio of the number of cases favorable to an event to the number of all outcomes possible, where each of the outcomes is equally likely.

Probability is usually denoted by P , and the respective elements of the sample space (the outcomes) are denoted by A, B, C , etc. The mathematical notation that indicates the probability that an outcome, A , happens is $P(A)$. We use the following formula to calculate the probability of an outcome occurring:

$$P(A) = \frac{\text{The number of outcomes for } A \text{ to occur}}{\text{The size of the sample space}}$$

Example: When tossing two coins, what is the probability of getting a head on both coins, HH ? Is the probability classical?

Since there are 4 elements (outcomes) in the sample space set, $\{HH, HT, TH, TT\}$, its size is 4. Furthermore, there is only 1 HH outcome that can occur. Therefore, using the formula above, we can calculate the probability as shown:

$$P(A) = \frac{\text{The number of outcomes for } HH \text{ to occur}}{\text{The size of the sample space}} = \frac{1}{4} = 25\%$$

Notice that each of the 4 possible outcomes is equally likely. The probability of each is 0.25. Also notice that the total probability of all possible outcomes in the sample space is 1.

Example: What is the probability of throwing a die and getting $A = 2, 3$, or 4 ?

There are 6 possible outcomes when you toss a die. Thus, the total number of outcomes in the sample space is 6. The event we are interested in is getting a 2, 3, or 4, and there are three ways for this event to occur.

$$P(A) = \frac{\text{The number of outcomes for } 2, 3, \text{ or } 4 \text{ to occur}}{\text{The size of the sample space}} = \frac{3}{6} = \frac{1}{2} = 50\%$$

Therefore, there is a probability of 0.5 that we will get 2, 3, or 4.

Example: Consider tossing two coins. **Assume the coins are not balanced.** The design of the coins is such that they produce the probabilities shown in the table below:

TABLE 3.1:

Outcome	Probability
HH	$\frac{4}{9}$
HT	$\frac{2}{9}$
TH	$\frac{2}{9}$
TT	$\frac{1}{9}$

Figure: Probability table for flipping two **weighted** coins.

What is the probability of observing exactly one head, and what is the probability of observing at least one head?

Notice that the simple events HT and TH each contain only one head. Thus, we can easily calculate the probability of observing exactly one head by simply adding the probabilities of the two simple events:

$$\begin{aligned} P &= P(HT) + P(TH) \\ &= \frac{2}{9} + \frac{2}{9} \\ &= \frac{4}{9} \end{aligned}$$

Similarly, the probability of observing at least one head is:

$$\begin{aligned} P &= P(HH) + P(HT) + P(TH) \\ &= \frac{4}{9} + \frac{2}{9} + \frac{2}{9} = \frac{8}{9} \end{aligned}$$

Lesson Summary

An **event** is something that occurs, or happens, with one or more possible outcomes.

An **experiment** is the process of taking a measurement or making an observation.

A **simple event** is the simplest outcome of an experiment.

The **sample space** is the set of all possible outcomes of an experiment, typically denoted by S .

Review Questions

1. Consider an experiment composed of throwing a die followed by throwing a coin.
 - a. List the simple events and assign a probability for each simple event.
 - b. What are the probabilities of observing the following events?
 - (i) A 2 on the die and H on the coin
 - (ii) An even number on the die and T on the coin
 - (iii) An even number on the die
 - (iv) T on the coin
 2. A box contains two blue marbles and three red ones. Two marbles are drawn randomly **without replacement**. Refer to the blue marbles as $B1$ and $B2$ and the red ones as $R1$, $R2$, and $R3$.
 - a. List the outcomes in the sample space.
 - b. Determine the probability of observing each of the following events:
 - (i) Drawing 2 blue marbles
 - (ii) Drawing 1 red marble and 1 blue marble
 - (iii) Drawing 2 red marbles
- Answers:** (1.) (a) $1/12$ (1.) (b) (i) $1/12$ (1.) (b) (ii) $\frac{1}{4}$ (1.) (b) (iii) $\frac{1}{2}$ (1.) (b) (iv) $\frac{1}{2}$ (2.) (a) See detailed answers (2.) (b) (i) $1/10$ (2.) (b) (ii) $3/5$ (2.) (b) (iii) $3/10$

3.2 Compound Events

Learning Objectives

- Know basic operations of unions and intersections.
- Calculate the probability of occurrence of two (or more) simultaneous events.
- Calculate the probability of occurrence of either of two (or more) events.

Introduction

In this lesson, you will learn how to combine two or more events by finding the union of the two events or the intersection of the two events. You will also learn how to calculate probabilities related to unions and intersections.

Union and Intersection

Sometimes we need to combine two or more events into one *compound event*. This compound event can be formed in two ways.

The *union of events A and B* occurs if either event A, event B, or both occur in a single performance of an experiment. We denote the union of the two events by the symbol $A \cup B$. You read this as either “A union B” or “A or B.” $A \cup B$ means everything that is in set A or in set B or in both sets.

The *intersection of events A and B* occurs if both event A and event B occur in a single performance of an experiment. It is where the two events overlap. We denote the intersection of two events by the symbol $A \cap B$. You read this as either “A intersection B” or “A and B.” $A \cap B$ means everything that is in set A and in set B. That is, when looking at the intersection of two sets, we are looking for where the sets overlap.

Example: Consider the throw of a die experiment. Assume we define the following events:

A : observe an even number

B : observe a number less than or equal to 3

- a. Describe $A \cup B$ for this experiment.
- b. Describe $A \cap B$ for this experiment.
- c. Calculate $P(A \cup B)$ and $P(A \cap B)$, assuming the die is fair.

The sample space of a fair die is $S = \{1, 2, 3, 4, 5, 6\}$, and the sample spaces of the events A and B above are $A = \{2, 4, 6\}$ and $B = \{1, 2, 3\}$.

1. An observation on a single toss of the die is an element of the union of A and B if it is either an even number, a number that is less than or equal to 3, or a number that is both even and less than or equal to 3. In other words, the simple events of $A \cup B$ are those for which A occurs, B occurs, or both occur:

$$A \cup B = \{2, 4, 6\} \cup \{1, 2, 3\} = \{1, 2, 3, 4, 6\}$$

2. An observation on a single toss of the die is an element of the intersection of A and B if it is a number that is both even and less than 3. In other words, the simple events of $A \cap B$ are those for which both A and B occur:

$$A \cap B = \{2, 4, 6\} \cap \{1, 2, 3\} = \{2\}$$

3. Remember, the probability of an event is the sum of the probabilities of its simple events. This is shown for $A \cup B$ as follows:

$$\begin{aligned} P(A \cup B) &= P(1) + P(2) + P(3) + P(4) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{5}{6} \end{aligned}$$

Similarly, this can also be shown for $A \cap B$:

$$P(A \cap B) = P(2) = \frac{1}{6}$$

Intersections and unions can also be defined for more than two events. For example, $A \cup B \cup C$ represents the union of three events.

Example: Refer to the above example and answer the following questions based on the definitions of the new events C and D .

C : observe a number that is greater than 5

D : observe a number that is exactly 5

- a. Find the simple events in $A \cup B \cup C$.
- b. Find the simple events in $A \cap D$.
- c. Find the simple events in $A \cap B \cap C$.

1. Since $C = \{6\}$, $A \cup B \cup C = \{2, 4, 6\} \cup \{1, 2, 3\} \cup \{6\} = \{1, 2, 3, 4, 6\}$.

2. Since $D = \{5\}$, $A \cap D = \{2, 3, 6\} \cap \{5\} = \emptyset$,

where \emptyset is the empty set. This means that there are no elements in the set $A \cap D$.

3. Here, we need to be a little careful. We need to find the intersection of the three sets. To do so, it is a good idea to use the associative property by first finding the intersection of sets A and B and then intersecting the resulting set with C .

Again, we get the empty set.

Lesson Summary

The **union** of the two events A and B , written $A \cup B$, occurs if either event A , event B , or both occur on a single performance of an experiment. A union is an 'or' relationship.

The **intersection** of the two events A and B , written $A \cap B$, occurs only if both event A and event B occur on a single performance of an experiment. An intersection is an 'and' relationship. Intersections and unions can be used to combine more than two events.

Review Questions

1. Consider 3 sets, A , B , and C . $A = \{1, 2, 5, 7, 9\}$, $B = \{4\}$, and $C = \{1, 2, 4, 8\}$.

Find $A \cap B$, $A \cup B$, $A \cup C$, $(B \cap C) \cup A$, $(A \cup B) \cup C$, and $A \cap B \cap C$.

Answers: 1. $A \cap B = \emptyset$. $A \cup B = \{1, 2, 4, 5, 7, 9\}$ $A \cup C = \{1, 2, 4, 5, 7, 8, 9\}$. $(B \cap C) \cup A = \{1, 2, 4, 5, 7, 9\}$. $(A \cup B) \cup C = \{1, 2, 4, 5, 7, 8, 9\}$. $A \cap B \cap C = \emptyset$.

3.3 The Complement of an Event

Learning Objectives

- Know the definition of the complement of an event.
- Use the complement of an event to calculate the probability of an event.
- Understand the Complement Rule.

Introduction

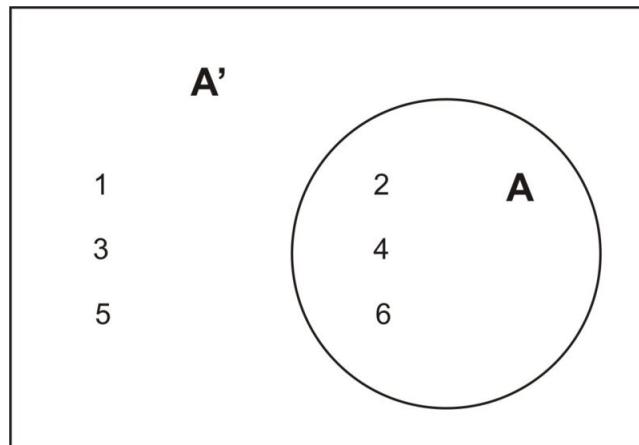
In this lesson, you will learn what is meant by the complement of an event, and you will be introduced to the Complement Rule. You will also learn how to calculate probabilities when the complement of an event is involved.

The Complement of an Event

The *complement* A' of the event A consists of all elements of the sample space that are not in A .

Example: Let us refer back to the experiment of throwing one die. As you know, the sample space of a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. If we define the event A as observing an odd number, then $A = \{1, 3, 5\}$. The complement of A will be all the elements of the sample space that are not in A . Thus, $A' = \{2, 4, 6\}$

A *Venn diagram* that illustrates the relationship between A and A' is shown below:



This leads us to say that the sum of the possible outcomes for event A and the possible outcomes for its complement, A' , is all the possible outcomes in the sample space of the experiment. Therefore, the probabilities of an event and its complement must sum to 1.

The Complement Rule

The *Complement Rule* states that the sum of the probabilities of an event and its complement must equal 1.

$$P(A) + P(A') = 1$$

As you will see in the following examples, it is sometimes easier to calculate the probability of the complement of an event than it is to calculate the probability of the event itself. Once this is done, the probability of the event, $P(A)$, is calculated using the relationship $P(A) = 1 - P(A')$.

Example: Suppose you know that the probability of getting the flu this winter is 0.43. What is the probability that you will not get the flu?

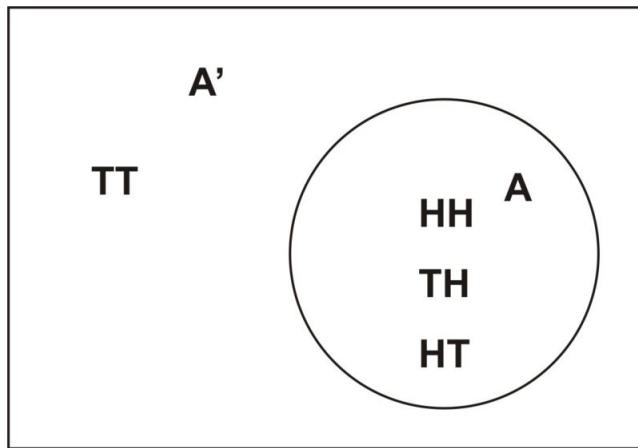
Let the event A be getting the flu this winter. We are given $P(A) = 0.43$. The event not getting the flu is A' . Thus, $P(A') = 1 - P(A) = 1 - 0.43 = 0.57$.

Example: Two coins are tossed simultaneously. Let the event A be observing at least one head.

What is the complement of A , and how would you calculate the probability of A by using the Complement Rule?

Since the sample space of event $A = \{HT, TH, HH\}$, the complement of A will be all events in the sample space that are not in A . In other words, the complement will be all the events in the sample space that do not involve heads. That is, $A' = \{TT\}$.

We can draw a simple Venn diagram that shows A and A' when tossing two coins as follows:



The second part of the problem is to calculate the probability of A using the Complement Rule. Recall that $P(A) = 1 - P(A')$. This means that by calculating $P(A')$, we can easily calculate $P(A)$ by subtracting $P(A')$ from 1.

$$\begin{aligned}P(A') &= P(TT) = \frac{1}{4} \\P(A) &= 1 - P(A') = 1 - \frac{1}{4} = \frac{3}{4}\end{aligned}$$

Obviously, we would have gotten the same result if we had calculated the probability of event A occurring directly. The next example, however, will show you that sometimes it is much easier to use the Complement Rule to find the answer that we are seeking.

Example: Consider the experiment of tossing a coin ten times. What is the probability that we will observe at least one head?

What are the simple events of this experiment? As you can imagine, there are many simple events, and it would take a very long time to list them. One simple event may be $HTHTHTHHTH$, another may be $THTHHHTHTH$, and so on. There are, in fact, $2^{10} = 1024$ ways to observe at least one head in ten tosses of a coin.

To calculate the probability, it's necessary to keep in mind that each time we toss the coin, the chance is the same for heads as it is for tails. Therefore, we can say that each simple event among the 1024 possible events is equally likely to occur. Thus, the probability of any one of these events is $\frac{1}{1024}$.

We are being asked to calculate the probability that we will observe at least one head. You will probably find it difficult to calculate, since heads will almost always occur at least once during 10 consecutive tosses. However, if we determine the probability of the complement of A (i.e., the probability that no heads will be observed), our answer will become a lot easier to calculate. The complement of A contains only one event: $A' = \{TTTTTTTTT\}$. This is the only event in which no heads appear, and since all simple events are equally likely, $P(A') = \frac{1}{1024}$.

Using the Complement Rule, $P(A) = 1 - P(A') = 1 - \frac{1}{1024} = \frac{1023}{1024} = 0.999$.

That is a very high percentage chance of observing at least one head in ten tosses of a coin.

Lesson Summary

The **complement** A' of the event A consists of all outcomes in the sample space that are not in event A .

The **Complement Rule** states that the sum of the probabilities of an event and its complement must equal 1, or for the event A , $P(A) + P(A') = 1$.

Review Questions

1. A fair coin is tossed three times. Two events are defined as follows:

A : at least one head is observed

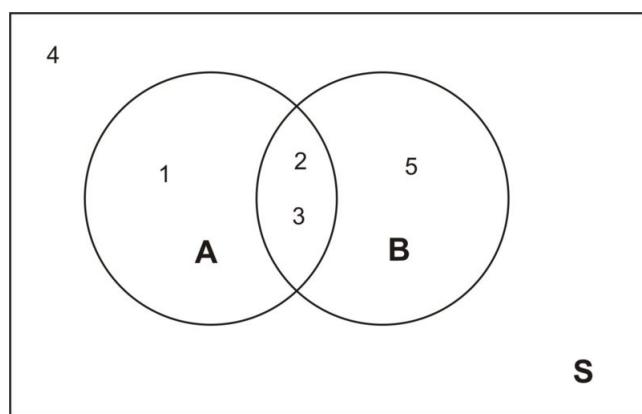
B : an odd number of heads is observed

- a. List the sample space for tossing the coin three times.
- b. List the outcomes of A .
- c. List the outcomes of B .
- d. List the outcomes of the following events: $A \cup B, A', A \cap B$.
- e. Find each of the following: $P(A), P(B), P(A \cup B), P(A'), P(A \cap B)$.

2. The Venn diagram below shows an experiment with five simple events. The two events A and B are shown.

The probabilities of the simple events are as follows: $P(1) = \frac{1}{10}, P(2) = \frac{2}{10}, P(3) = \frac{3}{10}, P(4) = \frac{1}{10}, P(5) = \frac{3}{10}$.

Find each of the following: $P(A'), P(B'), P(A' \cap B), P(A \cap B), P(A \cup B'), P(A \cup B), P(A \cap B'), P[(A \cup B)']$.



Answers to Review Questions

(1)(a). (TTT), (HTT), (THT), (TTH), (THH), (HTH), (HHT), (HHH).

(1) (b) (HTT), (THT), (TTH), (THH), (HTH), (HHT), (HHH).

(1)(c) (HTT), (THT), (TTH), (THH), (HTH), (HHT), (HHH).

(1)(d) $A \cup B = \{(HTT), (THT), (TTH), (THH), (HTH), (HHT), (HHH)\}$. $A' = \{(TTT)\}$. $A \cap B = \{(HTT), (THT), (TTH), (HHH)\}$.

(1)(e) $P(A) = 7/8$ $P(B) = \frac{1}{2}$ $P(A \cup B) = 7/8$ $P(A') = 1/8$ $P(A \cap B) = \frac{1}{2}$

2. $P(A') = 2/5$

$$P(B') = 1/5$$

$$P(A' \cap B) = 3/10$$

$$P(A \cap B) = \frac{1}{2}$$

$$P(A \cup B') = 7/10$$

$$P(A \cup B) = 9/10$$

$$P(A \cap B') = 1/10$$

$$P[(A \cup B)'] = 1/10$$

3.4 Conditional Probability

Learning Objective

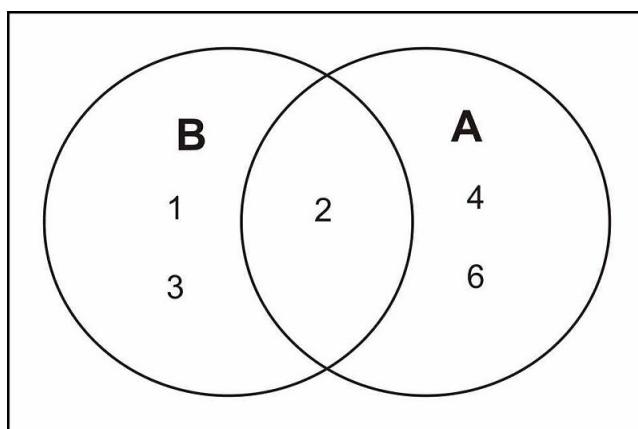
- Calculate the conditional probability that event A occurs, given that event B has occurred, using the conditional probability formula and the natural frequencies approach.

Introduction

In this lesson, you will learn about the concept of conditional probability and be presented with some examples of how conditional probability is used in the real world. You will also learn the appropriate notation associated with conditional probability.

Notation

We know that the probability of observing an even number on a throw of a die is 0.5. Let the event of observing an even number be event A . Now suppose that we throw the die, and we know that the result is a number that is 3 or less. Call this event B . Would the probability of observing an even number on that particular throw still be 0.5? The answer is no, because with the introduction of event B , we have reduced our sample space from 6 simple events to 3 simple events. In other words, since we have a number that is 3 or less, we now know that we have a 1, 2 or 3. This becomes, in effect, our sample space. Now the probability of observing a 2 is $\frac{1}{3}$. With the introduction of a particular condition (event B), we have changed the probability of a particular outcome. The Venn diagram below shows the reduced sample space for this experiment, given that event B has occurred:



The only even number in the sample space for B is the number 2. We conclude that the probability that A occurs, given that B has occurred, is $1:3$, or $\frac{1}{3}$. We write this with the notation $P(A|B)$, which reads “the probability of A , given B .” So for the die toss experiment, we would write $P(A|B) = \frac{1}{3}$.

Conditional Probability of Two Events

If A and B are two events, then the probability of event A occurring, given that event B has occurred, is called conditional probability. We write it with the notation $P(A|B)$, which reads “the probability of A , given B .”

To calculate the conditional probability that event A occurs, given that event B has occurred, take the ratio of the probability that both A and B occur to the probability that B occurs. That is:

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

For our example above, the die toss experiment, we proceed as is shown below:

A : observe an even number

B : observe a number less than or equal to 3

To find the conditional probability, we use the formula as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(2)}{P(1) + P(2) + P(3)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Example: A medical research center is conducting experiments to examine the relationship between cigarette smoking and cancer in a particular city in the USA. Let A represent an individual who smokes, and let C represent an individual who develops cancer. This means that AC represents an individual who smokes and develops cancer, AC' represents an individual who smokes but does not develop cancer, and so on. We have four different possibilities, or simple events, and they are shown in the table below, along with their associated probabilities.

TABLE 3.2:

Simple Events	Probabilities
AC	0.10
AC'	0.30
$A'C$	0.05
$A'C'$	0.55

Figure: A table of probabilities for combinations of smoking, A , and developing cancer, C .

These simple events can be studied, along with their associated probabilities, to examine the relationship between smoking and cancer.

We have:

A : individual smokes

C : individual develops cancer

A' : individual does not smoke

C' : individual does not develop cancer

A very powerful way of examining the relationship between cigarette smoking and cancer is to compare the conditional probability that an individual gets cancer, given that he/she smokes, with the conditional probability that an individual gets cancer, given that he/she does not smoke. In other words, we want to compare $P(C|A)$ with $P(C|A')$.

Recall that $P(C|A) = \frac{P(C \cap A)}{P(A)}$.

Before we can use this relationship, we need to calculate the value of the denominator. $P(A)$ is the probability of an individual being a smoker in the city under consideration. To calculate it, remember that the probability of an event is the sum of the probabilities of all its simple events. A person can smoke and have cancer, or a person can smoke and not have cancer. That is:

$$P(A) = P(AC) + P(AC') = 0.10 + 0.30 = 0.4$$

This tells us that according to this study, the probability of finding a smoker selected at random from the sample space (the city) is 40%. We can continue on with our calculations as follows:

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(AC)}{P(A)} = \frac{0.10}{0.40} = 0.25 = 25\%$$

Similarly, we can calculate the conditional probability of a nonsmoker developing cancer:

$$P(C|A') = \frac{P(A' \cap C)}{P(A')} = \frac{P(A'C)}{P(A')} = \frac{0.05}{0.60} = 0.08 = 8\%$$

In this calculation, $P(A') = P(A'C) + P(A'C') = 0.05 + 0.55 = 0.60$. $P(A')$ can also be found by using the Complement Rule as shown: $P(A') = 1 - P(A) = 1 - 0.40 = 0.60$.

From these calculations, we can clearly see that a relationship exists between smoking and cancer. The probability that a smoker develops cancer is 25%, and the probability that a nonsmoker develops cancer is only 8%. The ratio between the two probabilities is $\frac{0.25}{0.08} = 3.125$, which means a smoker is more than three times more likely to develop cancer than a nonsmoker. Keep in mind, though, that it would not be accurate to say that smoking causes cancer. However, our findings do suggest a strong link between smoking and cancer.

The Natural Frequencies Approach

Here is an alternate way of conceptualizing the problem above, called the **natural frequencies approach**, and it doesn't require the use of a confusing formula. We will use the probability information given above to demonstrate this approach. Suppose you have 1000 people. Of these 1000 people, 100 smoke and have cancer, and 300 smoke and don't have cancer. Therefore, of the 400 people who smoke, 100 have cancer. The probability of having cancer, given that you smoke, is $\frac{100}{400} = 0.25$.

Of these 1000 people, 50 don't smoke and have cancer, and 550 don't smoke and don't have cancer. Thus, of the 600 people who don't smoke, 50 have cancer. Therefore, the probability of having cancer, given that you don't smoke, is $\frac{50}{600} = 0.08$.

Lesson Summary

If A and B are two events, then the probability of event A occurring, given that event B has occurred, is called **conditional probability**. We write it with the notation $P(A|B)$, which reads “the probability of A , given B .”

Conditional probability can be found with the **equation** $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Another way to determine a conditional probability is to use the **natural frequencies approach**.

Review Questions

1. If $P(A) = 0.3$, $P(B) = 0.7$, and $P(A \cap B) = 0.15$, find $P(A|B)$ and $P(B|A)$.
2. Two fair coins are tossed.
 - a. List all the possible outcomes in the sample space.
 - b. Suppose two events are defined as follows:

A : At least one head appears

B : Only one head appears

Find $P(A)$, $P(B)$, $P(A \cap B)$, $P(A|B)$, and $P(B|A)$.

3. Let a single balanced 6-sided die be rolled. Using the Venn diagram below, let A be the event that a number less than or equal to 5 is rolled. Let B be the event that a number 4 or larger is obtained. Find $P(A)$, $P(B)$, $P(A \cap B)$, $P(A|B)$, and $P(B|A)$.

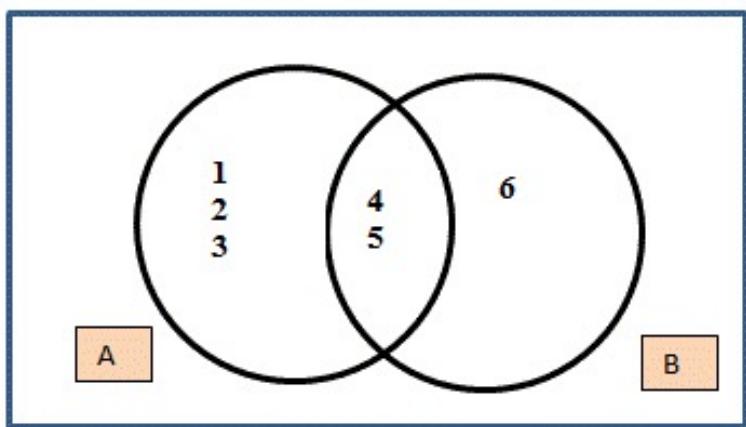


FIGURE 3.1

Question 4 is based on the reading “Simple, Joint, Marginal and Conditional Probabilities.”

4. Students at a beach town in Florida are asked their favorite summer sport. Their responses are summarized in the table. (Use the natural frequencies approach to solve answer these questions.)

TABLE 3.3: Summer Sports

	Tennis	Beach Volleyball	Surfing	Totals
Male	22	16	18	56
Female	26	12	6	44
Totals	48	28	24	100

- Find the *marginal* probability that a student prefers surfing.
 - Find the *joint* probability of females who prefer beach volleyball.
 - Given that a student prefers surfing, what is the probability that the student is male?
 - Given that a student is female, what is the probability that she prefers tennis?
 - Given that a student prefers tennis, what is the probability that the student is female?
-

Answers to Review Exercises

1. $P(A|B) = .214$ $P(B|A) = 0.50$.

2.a. (TT), (HT), (TH), (HH)

2.b. Event A is (HT), (TH), (HH) and Event B is (HT), (TH).

$$P(A) = 3/4$$

$$P(B) = 1/2$$

$$P(A \cap B) = 1/2$$

$$P(A|B) = 1$$

$$P(B|A) = 2/3$$

3. $P(A) = 5/6$

$$P(B) = 1/2$$

$$P(A \cap B) = 1/3$$

$$P(A|B) = 2/3$$

$$P(B|A) = 2/5$$

4. a. $P(\text{Surfing}) = 6/25$

b. $P(\text{Female} \cap \text{Volleyball}) = 3/25$

c. $P(\text{Male} | \text{Surfing}) = 3/4$

d. $P(\text{Tennis} | \text{Female}) = 13/22$

e. $P(\text{Female} | \text{Tennis}) = 13/24$

3.5 Addition and Multiplication Rules

Learning Objectives

- Calculate probabilities using the Addition Rule for mutually exclusive and nonmutually exclusive events.
- Calculate probabilities using the Multiplication Rule for independent and nonindependent events.

Introduction

In this lesson, you will learn how to combine probabilities with the Addition Rule and the Multiplication Rule. Through the examples in this lesson, it will become clear when to use which rule. You will also be presented with information about mutually exclusive events and independent events.

Venn Diagrams

When the probabilities of certain events are known, we can use these probabilities to calculate the probabilities of their respective unions and intersections. We use two rules, the Addition Rule and the Multiplication Rule, to find these probabilities. The examples that follow will illustrate how we can do this.

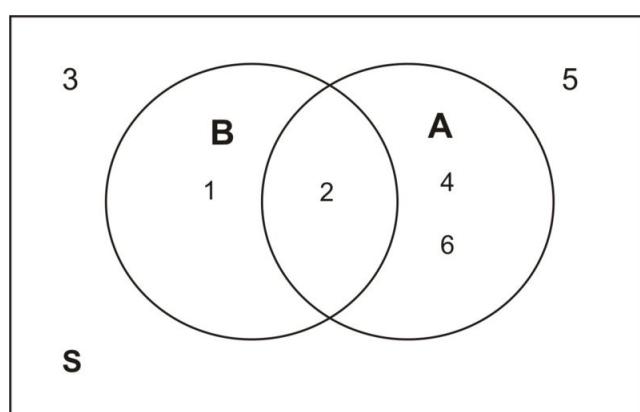
Example: Suppose we have a loaded (unfair) die, and we toss it several times and record the outcomes. We will define the following events:

A : observe an even number

B : observe a number less than 3

Let us suppose that we have $P(A) = 0.4$, $P(B) = 0.3$, and $P(A \cap B) = 0.1$. We want to find $P(A \cup B)$.

It is probably best to draw a Venn diagram to illustrate this situation. As you can see, the probability of events A or B occurring is the union of the individual probabilities of each event.



Therefore, adding the probabilities together, we get the following:

$$P(A \cup B) = P(1) + P(2) + P(4) + P(6)$$

We have also previously determined the probabilities below:

$$P(A) = P(2) + P(4) + P(6) = 0.4$$

$$P(B) = P(1) + P(2) = 0.3$$

$$P(A \cap B) = P(2) = 0.1$$

If we add the probabilities $P(A)$ and $P(B)$, we get:

$$P(A) + P(B) = P(2) + P(4) + P(6) + P(1) + P(2)$$

Note that $P(2)$ is included twice. We need to be sure not to double-count this probability. Also note that 2 is in the intersection of A and B . It is where the two sets overlap. This leads us to the following:

$$P(A \cup B) = P(1) + P(2) + P(4) + P(6)$$

$$P(A) = P(2) + P(4) + P(6)$$

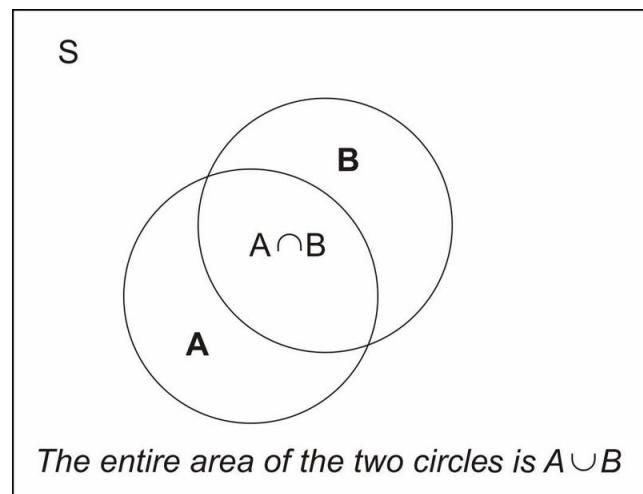
$$P(B) = P(1) + P(2)$$

$$P(A \cap B) = P(2)$$

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

This is the *Addition Rule of Probability*, which is demonstrated below:

$$P(A \cup B) = 0.4 + 0.3 - 0.1 = 0.6$$



What we have shown is that the probability of the union of two events, A and B , can be obtained by adding the individual probabilities, $P(A)$ and $P(B)$, and subtracting the probability of their intersection (or overlap), $P(A \cap B)$. The Venn diagram above illustrates this union.

Addition Rule of Probability

The probability of the union of two events can be obtained by adding the individual probabilities and subtracting the probability of their intersection: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

We can rephrase the definition as follows: The probability that either event A or event B occurs is equal to the probability that event A occurs plus the probability that event B occurs minus the probability that both occur.

Example: Consider the experiment of randomly selecting a card from a deck of 52 playing cards. What is the probability that the card selected is either a spade or a face card?

Our event is defined as follows:

$$E : \text{card selected is either a spade or a face card}$$

There are 13 spades and 12 face cards, and of the 12 face cards, 3 are spades. Therefore, the number of cards that are either a spade or a face card or both is $13 + 9 = 22$. That is, event E occurs when 1 of 22 cards is selected, the 22 cards being the 13 spade cards and the 9 face cards that are not spade. To find $P(E)$, we use the Addition Rule of Probability. First, define two events as follows:

$$C : \text{card selected is a spade}$$

$$D : \text{card selected is a face card}$$

Note that $P(E) = P(C \cup D) = P(C) + P(D) - P(C \cap D)$. Remember, with event C , 1 of 13 cards that are spades can be selected, and with event D , 1 of 12 face cards can be selected. Event $C \cap D$ occurs when 1 of the 3 face card spades is selected. These cards are the king, jack, and queen of spades. Using the Addition Rule of Probability formula:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= 0.250 + 0.231 - 0.058 \\ &= 0.423 \\ &= 42.3\% \end{aligned}$$

Recall that we are subtracting 0.058 because we do not want to double-count the cards that are at the same time spades and face cards.

Example: If you know that 84.2% of the people arrested in the mid 1990's were males, 18.3% of those arrested were under the age of 18, and 14.1% were males under the age of 18, what is the probability that a person selected at random from all those arrested is either male or under the age of 18?

First, define the events:

$$A : \text{person selected is male}$$

B : person selected in under 18

Also, keep in mind that the following probabilities have been given to us:

$$P(A) = 0.842$$

$$P(B) = 0.183$$

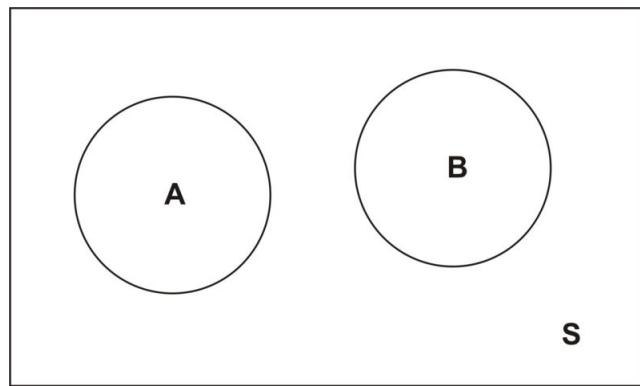
$$P(A \cap B) = 0.141$$

Therefore, the probability of the person selected being male or under 18 is $P(A \cup B)$ and is calculated as follows:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.842 + 0.183 - 0.141 \\ &= 0.884 \\ &= 88.4\% \end{aligned}$$

This means that 88.4% of the people arrested in the mid 1990's were either male or under 18. If $A \cap B$ is empty ($A \cap B = \emptyset$), or, in other words, if there is not overlap between the two sets, we say that A and B are *mutually exclusive*.

The figure below is a Venn diagram of mutually exclusive events. For example, set A might represent all the outcomes of drawing a card, and set B might represent all the outcomes of tossing three coins. These two sets have no elements in common.



If the events A and B are mutually exclusive, then the probability of the union of A and B is the sum of the probabilities of A and B : $P(A \cup B) = P(A) + P(B)$.

Note that since the two events are mutually exclusive, there is no double-counting.

Example: If two coins are tossed, what is the probability of observing at least one head?

First, define the events as follows:

A : observe only one head

B : observe two heads

Now the probability of observing at least one head can be calculated as shown:

$$P(A \cup B) = P(A) + P(B) = 0.5 + 0.25 = 0.75 = 75\%$$

Multiplication Rule of Probability

Recall from the previous section that conditional probability is used to compute the probability of an event, given that another event has already occurred:

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

This can be rewritten as $P(A \cap B) = P(A|B) \bullet P(B)$ and is known as the **Multiplication Rule of Probability**.

The Multiplication Rule of Probability says that the probability that both A and B occur equals the probability that B occurs times the conditional probability that A occurs, given that B has occurred.

Example: In a certain city in the USA some time ago, 30.7% of all employed female workers were white-collar workers. If 10.3% of all workers employed at the city government were female, what is the probability that a randomly selected employed worker would have been a female white-collar worker?

We first define the following events:

F : randomly selected worker who is female

W : randomly selected white-collar worker

We are trying to find the probability of randomly selecting a female worker who is also a white-collar worker. This can be expressed as $P(F \cap W)$.

According to the given data, we have:

$$P(F) = 10.3\% = 0.103$$

$$P(W|F) = 30.7\% = 0.307$$

Now, using the Multiplication Rule of Probability, we get:

$$P(F \cap W) = P(F)P(W|F) = (0.103)(0.307) = 0.0316 = 3.16\%$$

Thus, 3.16% of all employed workers were white-collar female workers.

Example: A college class has 42 students of which 17 are male and 25 are female. Suppose the teacher selects two students at random from the class. Assume that the first student who is selected is not returned to the class population. What is the probability that the first student selected is female and the second is male?

Here we can define two events:

$F1$: first student selected is female

$M2$: second student selected is male

In this problem, we have a conditional probability situation. We want to determine the probability that the first student selected is female and the second student selected is male. To do so, we apply the Multiplication Rule:

$$P(F1 \cap M2) = P(F1)P(M2|F1)$$

Before we use this formula, we need to calculate the probability of randomly selecting a female student from the population. This can be done as follows:

$$P(F1) = \frac{25}{42} = 0.595$$

Now, given that the first student selected is not returned back to the population, the remaining number of students is 41, of which 24 are female and 17 are male.

Thus, the conditional probability that a male student is selected, given that the first student selected was a female, can be calculated as shown below:

$$P(M2 | F1) = 17/41 = 0.415$$

Substituting these values into our equation, we get:

$$P(F1 \cap M2) = P(F1)P(M2|F1) = (0.595)(0.415) = 0.247 = 24.7\%$$

We conclude that there is a probability of 24.7% that the first student selected is female and the second student selected is male.

Example: Suppose a coin was tossed twice, and the observed face was recorded on each toss. The following events are defined:

A : first toss is a head

B : second toss is a head

Does knowing that event A has occurred affect the probability of the occurrence of B ?

The sample space of this experiment is $S = \{HH, HT, TH, TT\}$, and each of these simple events has a probability of 0.25. So far we know the following information:

$$\begin{aligned}
 P(A) &= P(HT) + P(HH) = \frac{1}{4} + \frac{1}{4} = 0.5 \\
 P(B) &= P(TH) + P(HH) = \frac{1}{4} + \frac{1}{4} = 0.5 \\
 A \cap B &= \{\text{HH}\} \\
 P(A \cap B) &= 0.25
 \end{aligned}$$

Now, what is the conditional probability? It is as follows:

$$\begin{aligned}
 P(B|A) &= \frac{P(A \cap B)}{P(A)} \\
 &= \frac{\frac{1}{4}}{\frac{1}{2}} \\
 &= \frac{1}{2}
 \end{aligned}$$

What does this tell us? It tells us that $P(B) = \frac{1}{2}$ and also that $P(B|A) = \frac{1}{2}$. This means knowing that the first toss resulted in heads does not affect the probability of the second toss being heads. In other words, $P(B|A) = P(B)$.

When this occurs, we say that events A and B are *independent events*.

Independence

If event B is independent of event A , then the occurrence of event A does not affect the probability of the occurrence of event B . Therefore, we can write $P(B) = P(B|A)$.

Recall that $P(B|A) = \frac{P(B \cap A)}{P(A)}$. Therefore, if B and A are independent, the following must be true:

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

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

That is, if two events are independent, $P(A \cap B) = P(A) \bullet P(B)$.

Example: The table below gives the number of physicists (in thousands) in the US cross-classified by specialty ($P1, P2, P3, P4$) and base of practice ($B1, B2, B3$). (Remark: The numbers are absolutely hypothetical and do not reflect the actual numbers in the three bases.) Suppose a physicist is selected at random. Is the event that the physicist selected is based in academia independent of the event that the physicist selected is a nuclear physicist? In other words, is event $B1$ independent of event $P3$?

TABLE 3.4:

		Academia (B1)	Industry (B2)	Government (B3)	Total
General Physics (P1)	10.3	72.3	11.2	93.8	
Semiconductors (P2)	11.4	0.82	5.2	17.42	

TABLE 3.4: (continued)

		Academia (B1)	Industry (B2)	Government (B3)	Total
Nuclear Physics (P3)		1.25	0.32	34.3	35.87
Astrophysics (P4)		0.42	31.1	35.2	66.72
Total		23.37	104.54	85.9	213.81

Figure: A table showing the number of physicists in each specialty (thousands). These data are hypothetical.

We need to calculate $P(B1|P3)$ and $P(B1)$. If these two probabilities are equal, then the two events $B1$ and $P3$ are indeed independent. From the table, we find the following:

$$P(B1) = \frac{23.37}{213.81} = 0.109$$

and

$$P(B1|P3) = \frac{P(B1 \cap P3)}{P(P3)} = \frac{1.25}{35.87} = 0.035$$

Thus, $P(B1|P3) \neq P(B1)$, and so events $B1$ and $P3$ are not independent.

Caution! If two events are mutually exclusive (they have no overlap), they are not independent. If you know that events A and B do not overlap, then knowing that B has occurred gives you information about A (specifically that A has not occurred, since there is no overlap between the two events). Therefore, $P(A|B) \neq P(A)$.

Lesson Summary

The Addition Rule of Probability states that the union of two events can be found by adding the probabilities of each event and subtracting the intersection of the two events, or $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If $A \cap B$ contains no simple events, then A and B are mutually exclusive. Mathematically, this means $P(A \cup B) = P(A) + P(B)$.

The Multiplication Rule of Probability states $P(A \cap B) = P(B) \bullet P(A|B)$.

If event B is independent of event A , then the occurrence of event A does not affect the probability of the occurrence of event B . Mathematically, this means $P(B) = P(B|A)$. Another formulation of independence is that if the two events A and B are independent, then $P(A \cap B) = P(A) \bullet P(B)$.

Review Questions

- Two fair dice are tossed, and the following events are identified:

A : sum of the numbers is odd

B : sum of the numbers is 9, 11, or 12

- Are events A and B independent? Why or why not?
- Are events A and B mutually exclusive? Why or why not?

The probability that a certain brand of television fails when first used is 0.1. If it does not fail immediately, the probability that it will work properly for 1 year is 0.99. What is the probability that a new television of the same brand will last 1 year?

3. Assume that the births of babies at a hospital are independent events, and that the probability of a boy baby is equal to the probability of a girl baby, and that each probability is 0.5.

a. What is the probability that the next 3 births at the hospital will be all boys?

b. What is the probability that the next 3 births at the hospital will be all girls?

c. Using your answers from parts (a) and (b) above, what is the probability that the next 3 births will be babies of the same gender?

4. In a certain neighborhood, 100 residents are asked whether they own a cell phone and whether they own an iPad. 60 of the residents own a cell phone, 40 own an iPad, and 20 own both a cell phone and an iPad.

a. Create a Venn diagram to summarize the data in this survey. Designate event A as {the resident owns a cell phone} and event B as {the resident owns an iPad}. Note that there will be some residents who own neither a cell phone nor an iPad. Include this in your Venn diagram.

b. Calculate the following probabilities: $P(A)$, $P(B)$, $P(A|B)$, $P(B|A)$, $P(A \cap B)$, $P(A \cup B)$, $P(A')$, $P[(A \cup B)']$, $P[(A \cap B)']$.

5. A standard deck of playing cards is used. The deck is shuffled, and a single card is randomly drawn from the deck.

a. What is the probability of drawing a red card?

b. What is the probability of drawing a spade?

c. What is the probability of drawing a face card?

d. What is the probability of drawing a single card that is both red **and** a face card?

e. What is the probability of drawing a single card that is both a heart **and** a queen?

f. What is the probability of drawing a heart **or** a queen?

g. Are the events {the card is a club} and {the card is a jack} mutually exclusive?

h. Are the events {the card is a club} and {the card is a jack} independent?

I. Find the conditional probability $P(\text{King} | \text{Spade})$. Compare this to $P(\text{King})$. What do you conclude?

6. Event A occurs with probability 0.3. If event A and B are mutually exclusive, then

A. $P(B) \geq 0.3$

B. $P(B) \leq 0.7$

C. $P(B) \geq 0.7$

D. $P(B) \leq 0.3$

7. Event A occurs with probability 0.3 and event B occurs with probability 0.4. If A and B are independent, we may conclude

A. $P(A \cap B) = 0.12$

B. $P(A|B) = 0.3$

C. $P(B|A) = 0.4$

D. all of the above

Answers: (1)(a) Not independent (1)(b) Not mutually exclusive (2) 0.891 (3)(a) 0.125 (3)(b) 0.125 (3)(c) 0.25 (4)(a) see detailed answers (4)(b) $P(A) = 0.6$ $P(B) = 0.4$. $P(A|B) = 0.5$. $P(B|A) = 0.33$. $P(A \cap B) = 0.20$. $P(A \cup B) = 0.8$. $P(A') = 0.4$. $P[(A \cup B)'] = 0.20$. $P[(A \cap B)'] = 0.80$.

(5) (a) $P(\text{Red Card}) = 0.5$. (5)(b) $P(\text{Spade}) = 13/52 = 0.25$ (5)(c) $P(\text{Face Card}) = 3/13$. (5)(d) $P(\text{Red and Face Card}) = (26/52) \times (6/52) \approx 0.058$. (5)(e) $P(\text{Heart and Queen}) = P(\text{Heart} \cap \text{Queen}) = 1/52$. (5)(f) $P(\text{Heart or Queen}) = 4/13$. (5)(g) not mutually exclusive (5)(h) independent. (5)(i) independent (6) B (7) D

3.6 Basic Counting Rules

Learning Objectives

- 2• Calculate different ways of ordering and combining items, using multiplication, combinations, and permutations.

Introduction

Inferential Statistics is a method of statistics that consists of drawing conclusions about a population based on information obtained from samples. Samples are used because it can be quite costly in time and money to study an entire population. In addition, because of the inability to actually reach everyone in a census, a sample can be more accurate than a census.

The most important characteristic of any sample is that it must be a very good representation of the population. It would not make sense to use the average height of basketball players to make an inference about the average height of the entire US population. Likewise, it would not be reasonable to estimate the average income of the entire state of California by sampling the average income of the wealthy residents of Beverly Hills. The goal of sampling is to obtain a representative sample. There are a number of different methods for taking representative samples, and in this lesson, you will be presented with the various counting rules used to assist researchers in choosing a good sample and in calculating probabilities.

The Multiplicative Rule of Counting

The *Multiplicative Rule of Counting* states the following:

(I) If there are n possible outcomes for event A and m possible outcomes for event B , then there are a total of nm possible outcomes for event A followed by event B .

This rule says that you simply multiply the number of items in each category to get the total possible number of possible outcomes.

Example: A restaurant offers a special dinner menu every day. There are three entrées, five appetizers, and four desserts to choose from. A customer can only select one item from each category. How many different meals can be ordered from the special dinner menu?

Let's summarize what we have.

Entrees: 3

Appetizer: 5

Dessert: 4

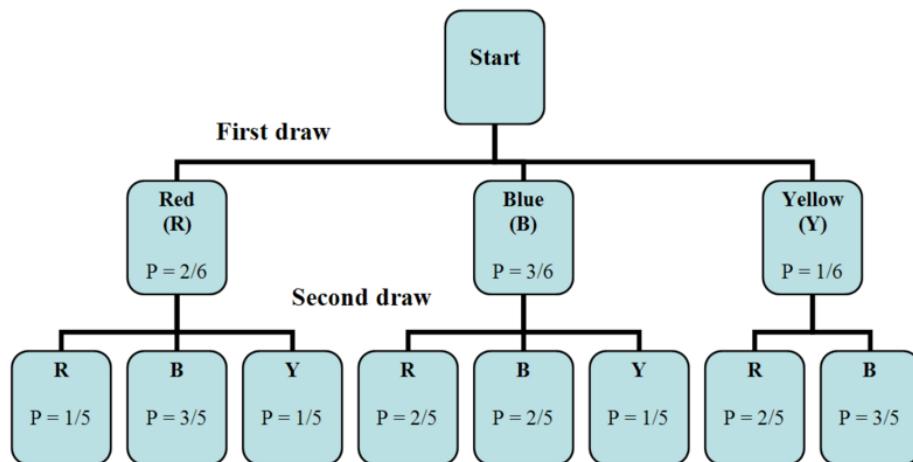
We use the Multiplicative Rule of Counting to calculate the number of different dinners that can be selected. We simply multiply each of the numbers of choices per item together: $(3)(5)(4) = 60$. Thus, there are 60 different dinners that can be ordered by the customers.

In this next example, you learn how to use a tree diagram to assist with determining the number of items in a sample space and how to calculate the associated probabilities for each of the possible outcomes.

Example: Suppose there are six balls in a box. They are identical, except in color. Two balls are red, three are blue,

and one is yellow. We will draw one ball, record its color, and then set it aside. Next, we will draw another ball and record its color. With the aid of a tree diagram, we can calculate the probability of each of the possible outcomes of the experiment.

We first draw a tree diagram to aid us in seeing all the possible outcomes of this experiment.



The tree diagram shows us the two stages of drawing two balls without putting the first one back into the box. In the first stage, we pick a ball blindly. Since there are 2 red balls, 3 blue balls, and 1 yellow ball, the probability of getting a red ball is $\frac{2}{6}$, the probability of getting a blue ball is $\frac{3}{6}$, and the probability of getting a yellow ball is $\frac{1}{6}$.

Remember that the probability associated with the second ball depends on the color of the first ball. Therefore, the two stages are not independent. To calculate the probabilities when selecting the second ball, we can look back at the tree diagram.

We follow each path in the tree diagram and we find that there are eight possible outcomes for the experiment:

RR: red on the 1st and red on the 2nd

RB: red on the 1st and blue on the 2nd

RY: red on the 1st and yellow on the 2nd

BR: blue on the 1st and red on the 2nd

BB: blue on the 1st and blue on the 2nd

BY: blue on the 1st and yellow on the 2nd

YR: yellow on the 1st and red on the 2nd

YB: yellow on the 1st and blue on the 2nd

We want to calculate the probability of each of these outcomes. This is done as is shown below.

$$\begin{aligned}
 P(RR) &= \frac{2}{6} \bullet \frac{1}{5} = \frac{2}{30} \\
 P(RB) &= \frac{2}{6} \bullet \frac{3}{5} = \frac{6}{30} \\
 P(RY) &= \frac{2}{6} \bullet \frac{1}{5} = \frac{2}{30} \\
 P(BR) &= \frac{3}{6} \bullet \frac{2}{5} = \frac{6}{30} \\
 P(YB) &= \frac{3}{6} \bullet \frac{2}{5} = \frac{6}{30} \\
 P(YB) &= \frac{3}{6} \bullet \frac{1}{5} = \frac{3}{30} \\
 P(YB) &= \frac{1}{6} \bullet \frac{2}{5} = \frac{2}{30} \\
 P(YB) &= \frac{1}{6} \bullet \frac{3}{5} = \frac{3}{30}
 \end{aligned}$$

Notice that all of the probabilities add up to 1, as they should.

When using a tree diagram to compute probabilities, you multiply the probabilities as you move along a branch. In the above example, if we are interested in the outcome RR , we note that the probability of picking a red ball on the first draw is $\frac{2}{6}$. We then go to the second branch, choosing a red ball on the second draw, the probability of which is $\frac{1}{5}$. Therefore, the probability of choosing RR is $(\frac{2}{6})(\frac{1}{5})$. The method used to solve the example above can be generalized to any number of stages.

We continue with the Multiplicative Rule of Counting to learn other types of problems that can be solved using this technique.

Example: In how many different ways can you seat 8 people at a dinner table?

For the first person, there are eight seat choices. For the second, there are seven remaining choices, since one person has already been seated. For the third seat, there are 6 choices, since two people are already seated. By the time we get to the last person, there is only one seat left. Therefore, using the Multiplicative Rule above, we get $(8)(7)(6)(5)(4)(3)(2)(1) = 40,320$.

The multiplication pattern above appears so often in statistics that it has its own name, which is *factorial*, and its own symbol, which is '!'. When describing it, we say, "Eight factorial," and we write, "8!"

Factorial Notation

$$n! = n(n-1)(n-2)(n-3)\dots(3)(2)(1)$$

The arrangement of elements in a distinct order, as the example above shows, is called a **permutation**. Thus, from the example above, there are 40,320 possible permutations of 8 people being seated at the dinner table.

Make a note that $0! = 1$.

Counting Rule for Permutations

The *Counting Rule for Permutations* states the following:

The number of ways to arrange n different objects in order within r positions is $P_r^n = \frac{n!}{(n-r)!}$.

Example: Let's revisit the dinner table seating arrangement. In how many different ways can you seat 8 people at a dinner table?

Now let's use the Counting Rule for Permutations to calculate the number of ways of seating the 8 dinner guests. We want to seat each of 8 people ($n = 8$) into 8 positions ($r = 8$).

$$P_8^8 = \frac{8!}{(8-8)!} = \frac{(8)(7)(6)(5)(4)(3)(2)(1)}{1} = 40,320$$

This gives us the identical answer that we obtained earlier by using the Multiplicative Rule for Counting.

Example: Let's compute the number of ordered seating arrangements we have with 8 people and only 5 seats.

In this case, we are considering a total of $n = 8$ people, and we wish to arrange $r = 5$ of these people to be seated. Substituting into the permutation equation, we get the following:

$$\begin{aligned} P_r^n &= \frac{n!}{(n-r)!} \\ &= \frac{8!}{(8-5)!} \\ &= \frac{8!}{3!} \\ &= \frac{40,320}{6} \\ &= 6,720 \end{aligned}$$

Another way of solving this problem is to use the Multiplicative Rule of Counting. Since there are only 5 seats available for 8 people, for the first seat, there are 8 people available. For the second seat, there are 7 remaining people available, since one person has already been seated. For the third seat, there are 6 people available, since two people have already been seated. For the fourth seat, there are 5 people available, and for the fifth seat, there are 4 people available. After that, we run out of seats. Thus, $(8)(7)(6)(5)(4) = 6,720$.

Example: The board of directors at The Orion Foundation has 13 members. Three officers will be elected from the 13 members to hold the positions of a provost, a general director, and a treasurer. How many different slates of three candidates are there if each candidate must specify which office he or she wishes to run for?

Each slate is a list of one person for each of three positions: the provost, the general director, and the treasurer. If, for example, Mr. Smith, Mr. Hale, and Ms. Osborn wish to be on the slate together, there are several different slates possible, depending on which one will run for provost, which one will run for general director, and which one will run for treasurer. This means that we are not just asking for the number of different groups of three names on the slate, but we are also asking for a specific order, since it makes a difference which name is listed in which position.

When computing the answer, $n = 13$ and $r = 3$.

Using the permutation formula, we get the following:

$$\begin{aligned} P_r^n &= \frac{n!}{(n-r)!} \\ &= \frac{13!}{(13-3)!} \\ &= \frac{(13)(12)(11)(10!)}{10!} \\ &= (13)(12)(11) \\ &= 1,716 \end{aligned}$$

Thus, there are 1,716 different slates of officers possible.

Notice that in our previous examples, the order of people or objects was taken into account. What if the order is not important? For example, in the previous example for electing three officers, what if we wish to choose 3 members of the 13 member board to attend a convention. Here, we are more interested in the group of three, but we are not interested in their order. In other words, we are only concerned with different **combinations** of 13 people taken 3 at a time. The permutation formula will not work here, since, in this situation, order is not important. However, we have a new formula that will compute different combinations.

Counting Rule for Combinations

The *Counting Rule for Combinations* states the following:

The number of combinations of n objects taken r at a time is $C_r^n = \frac{n!}{r!(n-r)!}$.

It is important to notice the difference between permutations and combinations. **When we consider grouping and order, we use permutations, but when we consider grouping with no particular order, we use combinations.**

Example: How many different groups of 3 are possible when taken out of 13 people?

Here, we are interested in combinations of 13 people taken 3 at a time. To find the answer, we can use the combination formula: $C_r^n = \frac{n!}{r!(n-r)!}$.

$$C_3^{13} = \frac{13!}{3!(13-3)!} = 286$$

This means that there are 286 different groups of 3 people to go to the convention.

In the above computation, you can see that the difference between the formulas for $_nC_r$ and $_nP_r$ is the factor $r!$ in the denominator of the fraction. Since $r!$ is the number of different orders of r objects, and combinations ignore order, we divide by the number of different orders.

Example: You are taking a philosophy course that requires you to read 5 books out of a list of 10 books. You are free to select any 5 books and read them in whichever order that pleases you. How many different combinations of 5 books are available from a list of 10?

Since consideration of the order in which the books are selected is not important, we compute the number of combinations of 10 books taken 5 at a time. We use the combination formula as is shown below:

$$C_r^n = \frac{n!}{r!(n-r)!}$$

$$C_5^{10} = \frac{10!}{5!(10-5)!} = 252$$

This means that there are 252 different groups of 5 books that can be selected from a list of 10 books.

Lesson Summary

Inferential Statistics is a method of statistics that consists of drawing conclusions about a population based on information obtained from a subset or sample of the population.

The **Multiplicative Rule of Counting** states that if there are n possible outcomes for event A and m possible outcomes for event B , then there are a total of nm possible outcomes for the series of events A followed by B .

The factorial sign, or ' $!$ ', is defined as $n! = n(n-1)(n-2)(n-3)\dots(3)(2)(1)$.

The number of **permutations (ordered arrangements)** of n different objects within r positions is $P_r^n = \frac{n!}{(n-r)!}$.

The number of **combinations (unordered arrangements)** of n objects taken r at a time is $C_r^n = \frac{n!}{r!(n-r)!}$.

Review Questions

1. Flying into Los Angeles from Washington DC, you can choose one of three airlines and can also choose either first class, business class, or economy class. How many travel options do you have?
2. How many different 5-card hands can be chosen from a 52-card deck?
3. Suppose an automobile license plate is designed to show a letter of the English alphabet, followed by a five-digit number. How many different license plates can be issued?
4. There are 30 candidates that are competing for three executive positions of President, Vice President, and Plant Manager. How many different ways can you fill the three positions?
5. In the same company, there are 30 factory workers, and 3 of them are to be chosen to receive training on a new machine. How many different groups of 3 can be chosen?
6. Determine the number of simple events when you toss a coin the following number of times. (Hint: As the numbers get higher, you will need to develop a systematic method of counting all the outcomes.)
 - a. Twice
 - b. Three times
 - c. Five times
 - d. n times (Look for a pattern in the results of a) through c).)

Answers: (1) 9 (2) 2,598,960 (3) 2,600,000 (4) 24,360 (5) 4,060 (6)(a) 2 (6)(b) 8 (6)(c) $2^5 = 32$ simple events (6)(d) 2^n

Keywords for Unit 3

Addition Rule of Probability

The Additive Rule of Probability states that the union of two events can be found by adding the probabilities of each event and subtracting the intersection of the two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Classical probability

Classical probability is defined to be the ratio of the number of cases favorable to an event to the number of all outcomes possible, where each of the outcomes is equally likely.

$$P(A) = \frac{\text{The number of outcomes for } A \text{ to occur}}{\text{The size of the sample space}}$$

Combinations

A combination is selection of objects without regard to order.

Complement

The *complement* A' of the event A consists of all elements of the sample space that are not in A .

Complement Rule

The *Complement Rule* states that the sum of the probabilities of an event and its complement must equal 1. $P(A) + P(A') = 1$

Compound event

we need to combine two or more events into one *compound event*. This compound event can be formed in two ways.

Conditional probability

If A and B are two events, then the probability of event A occurring, given that event B has occurred, is called *conditional probability*.

Counting Rule for Combinations

The number of combinations of n objects taken r at a time is $C_r^n = \frac{n!}{r!(n-r)!}$.

Counting Rule for Permutations

The number of ways to arrange n different objects in order within r positions is $P_r^n = \frac{n!}{(n-r)!}$.

Event

An event is something that occurs, or happens, with one or more possible outcomes.

Experiment

An *experiment* is the any physical action or procedure that is observed and the outcome noted

Factorial

The factorial sign, or '!', is defined as $n! = n(n-1)(n-2)(n-3)\dots(3)(2)(1)$. Remember that $0! = 1$.

Independent events

Two events are considered independent if the occurrence or nonoccurrence of one event has no influence on the occurrence or nonoccurrence of the other event.

Intersection of events

The *intersection of events* A and B occurs if both event A and event B occur in a single performance of an experiment.

Multiplicative Rule of Counting

If there are n possible outcomes for event A and m possible outcomes for event B, then there are a total of nm possible outcomes for event A followed by event B.

Multiplication Rule of Probability

$$P(A \cap B) = P(A|B) \bullet P(B)$$

Mutually exclusive

if there is no overlap between the two sets, we say that A and B are *mutually exclusive*.

Natural frequencies approach

Another way to determine a conditional probability is to use the natural frequencies approach.

Permutation

A Permutation is an arrangement of objects in a definite order

Sample space

The sample space is the set of all possible outcomes of an experiment, or the collection of all the possible simple events of an experiment.

Simple events

An event that has exactly one outcome.

Tree diagram

Tree diagrams are useful for showing all the possible outcomes when there is a series of events.

Union of events

The *union of events A and B* occurs if either event A, event B, or both occur in a single performance of an experiment.

Venn diagram

are diagrams that show all possible logical relations between a finite collection of sets.

3.7 References

1. . . CC BY-NC-SA

CHAPTER

4

Discrete Probability Distributions

Chapter Outline

- 4.1 TWO TYPES OF RANDOM VARIABLES**
 - 4.2 PROBABILITY DISTRIBUTION FOR A DISCRETE RANDOM VARIABLE**
 - 4.3 MEAN AND STANDARD DEVIATION OF DISCRETE RANDOM VARIABLES**
 - 4.4 SUMS AND DIFFERENCES OF INDEPENDENT RANDOM VARIABLES**
 - 4.5 THE BINOMIAL PROBABILITY DISTRIBUTION**
 - 4.6 REFERENCES**
-

4.1 Two Types of Random Variables

Learning Objective

- Learn to distinguish between the two types of random variables: continuous and discrete.

Introduction

The word discrete means countable. For example, the number of students in a class is countable, or discrete. The value could be 2, 24, 34, or 135 students, but it cannot be $\frac{232}{2}$ or 12.23 students. The cost of a loaf of bread is also discrete; it could be \$3.17, for example, where we are counting dollars and cents, but it cannot include fractions of a cent.

On the other hand, if we are measuring the tire pressure in an automobile, we are dealing with a continuous random variable. The air pressure can take values from 0 psi to some large amount that would cause the tire to burst. Another example is the height of your fellow students in your classroom. The values could be anywhere from, say, 4.5 feet to 7.2 feet. In general, quantities such as pressure, height, mass, weight, density, volume, temperature, and distance are examples of continuous random variables. Discrete random variables would usually come from counting, say, the number of chickens in a coop, the number of passing scores on an exam, or the number of voters who showed up to the polls.

Between any two values of a continuous random variable, there are an infinite number of other valid values. This is not the case for discrete random variables, because between any two discrete values, there is an integer number (0, 1, 2, ...) of valid values. Discrete random variables are considered countable values, since you could count a whole number of them. In this chapter, we will only describe and discuss discrete random variables and the aspects that make them important for the study of statistics.

Discrete Random Variables and Continuous Random Variables

In real life, most of our observations are in the form of numerical data that are the observed values of what are called *random variables*. In this chapter, we will study random variables and learn how to find probabilities of specific numerical outcomes.

The number of cars in a parking lot, the average daily rainfall in inches, the number of defective tires in a production line, and the weight in kilograms of an African elephant cub are all examples of *quantitative variables*.

If we let X represent a quantitative variable that can be measured or observed, then we will be interested in finding the numerical value of this quantitative variable. A random variable is a function that maps the elements of the sample space to a set of numbers.

Example: Three voters are asked whether they are in favor of building a charter school in a certain district. Each voter's response is recorded as 'Yes (Y)' or 'No (N)'. What are the random variables that could be of interest in this experiment?

As you may notice, the simple events in this experiment are not numerical in nature, since each outcome is either a 'Yes' or a 'No'. However, one random variable of interest is the number of voters who are in favor of building the school.

The table below shows all the possible outcomes from a sample of three voters. Notice that we assigned 3 to the first simple event (3 'Yes' votes), 2 to the second (2 'Yes' votes), 1 to the third (1 'Yes' vote), and 0 to the fourth (0 'Yes'

votes).

TABLE 4.1:

	Voter #1	Voter #2	Voter #3	Value of Random Variable (number of Yes votes)
1	Y	Y	Y	3
2	Y	Y	N	2
3	Y	N	Y	2
4	N	Y	Y	2
5	Y	N	N	1
6	N	Y	N	1
7	N	N	Y	1
8	N	N	N	0

Figure: Possible outcomes of the random variable in this example from three voters.

In the light of this example, what do we mean by random variable? The adjective 'random' means that the experiment may result in one of several possible values of the variable. For example, if the experiment is to count the number of customers who use the drive-up window in a fast-food restaurant between the hours of 8 AM and 11 AM, the random variable here is the number of customers who drive up within this time interval. This number varies from day to day, depending on random phenomena, such as today's weather, among other things. Thus, we say that the possible values of this random variable range from 0 to the maximum number that the restaurant can handle.

There are two types of random variables—discrete and continuous. Random variables that can assume only a countable number of values are called *discrete*. Random variables that can take on any of the countless number of values in an interval are called *continuous*.

Example: The following are examples of *discrete random variables*:

- The number of cars sold by a car dealer in one month
- The number of students who were protesting the tuition increase last semester
- The number of applicants who have applied for a vacant position at a company
- The number of typographical errors in a rough draft of a book

For each of these, if the variable is X , then $x = 0, 1, 2, 3, \dots$. Note that X can become very large. (In statistics, when we are talking about the random variable itself, we write the variable in uppercase, and when we are talking about the values of the random variable, we write the variable in lowercase.)

Example: The following are examples of *continuous random variables*.

- The length of time it takes a truck driver to go from New York City to Miami
- The depth of drilling to find oil
- The weight of a truck in a truck-weighing station
- The amount of water in a 12-ounce bottle

For each of these, if the variable is X , then $x > 0$ and less than some maximum value possible, but it can take on any value within this range.

Lesson Summary

A random variable represents the numerical value of a simple event of an experiment.

Random variables that can assume only a countable number of values are called discrete.

Random variables that can take on any of the countless number of values in an interval are called continuous.

Review Questions

- Identify whether each is a continuous random variable or a discrete random variable:
 - The weight of a newborn baby
 - The number of tomatoes on a tomato plant
 - The age of a patient at a nursing home
 - The number of elephants at the Brevard Zoo on June 10, 2013. (be careful!)
 - The length of a person's foot
 - A person's shoe size
- Here is the list of all possible outcomes of flipping a coin 3 times. Let the random variable X be defined as the number of heads obtained in each of the 8 outcomes. Fill in the chart with the value of the random variable for each outcome.

TABLE 4.2:

Outcome	Value of Random Variable X
(TTT)	
(HTT)	
(THT)	
(TTH)	
(THH)	
(HTH)	
(HHT)	
(HHH)	

Answers: (1)(a) Continuous (1)(b) Discrete (1)(c) Continuous (1)(d) Neither. This is a constant, not a random variable. (1)(e) Continuous (1)(f) Discrete (2) 0, 1, 1, 1, 2, 2, 2, 3

4.2 Probability Distribution for a Discrete Random Variable

Learning Objectives

- Construct the probability distribution of a discrete random variable.
- Use the probability distribution of a discrete random variable to calculate probabilities of outcomes.

Introduction

In this lesson, you will learn how to construct a probability distribution for a discrete random variable and represent this probability distribution with a graph, a table, or a formula. You will also learn the two conditions that all probability distributions must satisfy.

Probability Distribution for a Discrete Random Variable

The example below illustrates how to specify the possible values that a discrete random variable can assume.

Example: Suppose you simultaneously toss two fair coins. Let X be the number of heads observed. Find the probability associated with each value of the random variable X .

Since there are two coins, and each coin can be either heads or tails, there are four possible outcomes (HH, HT, TH, TT), each with a probability of $\frac{1}{4}$. Since X is the number of heads observed, $x = 0, 1, 2$.

We can identify the probabilities of the simple events associated with each value of X as follows:

$$\begin{aligned} P(x=0) &= P(TT) = \frac{1}{4} \\ P(x=1) &= P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\ P(x=2) &= P(HH) = \frac{1}{4} \end{aligned}$$

This is a complete description of all the possible values of the random variable, along with their associated probabilities. We refer to this as a *probability distribution*. This probability distribution can be represented in different ways. Sometimes it is represented in tabular form and sometimes in graphical form. Both forms are shown below.

In tabular form:

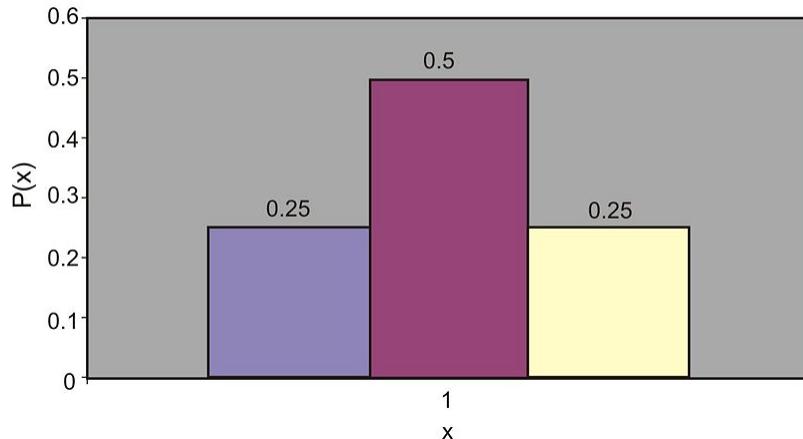
TABLE 4.3:

x	$P(x)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

Figure: The tabular form of the probability distribution for the random variable in the first example.

As a graph:

Probability Distribution for 2-Coin Toss



A probability distribution of a random variable specifies the values the random variable can assume, along with the probability of it assuming each of these values. All probability distributions must satisfy the following two conditions:

$$P(x) \geq 0, \text{ for all values of } X$$

$$\sum P(x) = 1, \text{ for all values of } X$$

Example: What is the probability distribution for the number of yes votes for three voters? (See the first example in section 4.1.)

Since each of the 8 outcomes is equally likely, the following table gives the probability of each value of the random variable.

TABLE 4.4:

Value of Random Variable (Number of Yes Votes)	Probability
3 corresponds to the one possible outcome (YYY)	$\frac{1}{8} = 0.125$
2 corresponds to the 3 possible outcomes (YYN) or (YNY) or (NYY)	$\frac{3}{8} = 0.375$
1 corresponds to the 3 possible outcomes (YNN) or (YN) or (NNY)	$\frac{3}{8} = 0.375$
0 corresponds to the one possible outcome (NNN)	$\frac{1}{8} = 0.125$

Figure: Tabular representation of the probability distribution for the random variable in the first example in section 4.1)

Lesson Summary

The **probability distribution** of a discrete random variable is a graph, a table, or a formula that specifies the probability associated with each possible value that the random variable can assume.

All probability distributions must satisfy the following two conditions:

$$P(x \geq 0), \text{ for all values of } X$$

$$\sum P(x) = 1, \text{ for all values of } X$$

Review Questions

1. Consider the following probability distribution:

x	-4	0	1	3
$P(x)$	0.1	0.3	0.4	0.2

- a. What are all the possible values of X ?
 - b. What value of X is most likely to happen?
 - c. What is the probability that $x > 0$?
 - d. What is the probability that $x = -2$?
2. A fair die is tossed twice, and the up face is recorded each time. Let X be the sum of the up faces.
- a. Give the probability distribution for X in tabular form.
 - b. What is $P(x \geq 8)$?
 - c. What is $P(x < 8)$?
 - d. What is the probability that x is odd? What is the probability that x is even?
 - e. What is $P(x = 7)$?
3. If a couple has three children, what is the probability that they have **at least** one boy? At most one girl?

Answers: (1)(a) -4, 0, 1, or 3 (1)(b) $X = 1$ (1)(c) 0.6 (1)(d) 0 (2)(a) (see detailed answers) (2)(b) $5/12$ (2)(c) $7/12$ (2)
 $, \frac{1}{2}$ (2)(e) $1/6$ (3) $P(\text{at least one boy}) = 7/8$. $P(\text{at most one girl}) = \frac{1}{2}$.

4.3 Mean and Standard Deviation of Discrete Random Variables

Learning Objectives

- Know the definition of the mean, or expected value, of a discrete random variable.
- Calculate the mean, standard deviation, and variance of a discrete random variable.

Introduction

In this lesson, you will be presented with the formulas for the mean, variance, and standard deviation of a discrete random variable. You will also be shown many real-world examples of how to use these formulas. In addition, the meaning of expected value will be discussed.

Characteristics of a Probability Distribution

The most important characteristics of any probability distribution are the mean (or average value) and the standard deviation (a measure of how spread out the values are). The example below illustrates how to calculate the mean and the standard deviation of a random variable. A common symbol for the mean is μ (mu), the lowercase m of the Greek alphabet. A common symbol for standard deviation is σ (sigma), the Greek lowercase s .

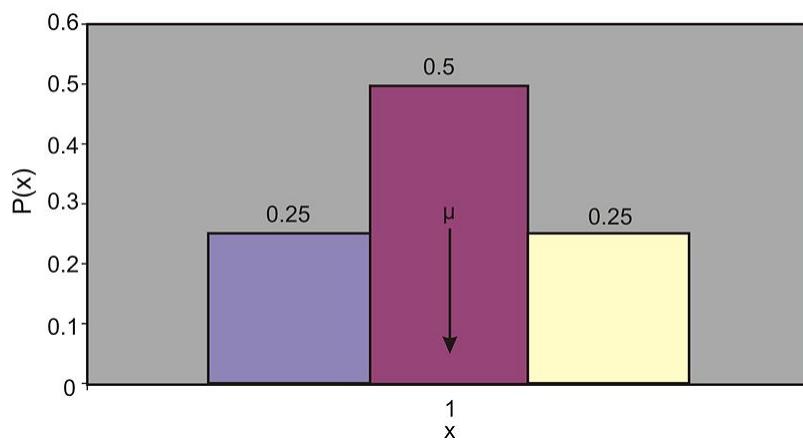
Example: Recall the probability distribution of the 2-coin experiment. Calculate the mean of this distribution.

If we look at the graph of the 2-coin toss experiment (shown below), we can easily reason that the mean value is located right in the middle of the graph, namely, at $x = 1$. This is intuitively true. Here is how we can calculate it:

To calculate the population mean, multiply each possible outcome of the random variable X by its associated probability and then sum over all possible values of X :

$$\mu = (0) \left(\frac{1}{4} \right) + (1) \left(\frac{1}{2} \right) + (2) \left(\frac{1}{4} \right) = 0 + \frac{1}{2} + \frac{1}{2} = 1$$

Probability Distribution for 2-Coin Toss



Mean Value or Expected Value

The mean value, or *expected value*, of a discrete random variable X is given by the following equation:

$$\mu = E(x) = \sum xp(x)$$

This definition is equivalent to the simpler one you have learned before:

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

However, the simpler definition would not be usable for many of the probability distributions in statistics.

Example: An insurance company sells life insurance of \$15,000 for a premium of \$310 per year. Actuarial tables show that the probability of death in the year following the purchase of this policy is 0.1%. What is the expected gain for this type of policy?

There are two simple events here: either the customer will live this year or will die. The probability of death, as given by the problem, is 0.001, and the probability that the customer will live is $1 - 0.001 = 0.999$. The company's expected gain from this policy in the year after the purchase is the random variable, which can have the values shown in the table below.

TABLE 4.5:

Gain, x	Simple Event	Probability
\$310	Live	0.999
-\$14,690	Die	0.001

Figure: Analysis of the possible outcomes of an insurance policy.

Remember, if the customer lives, the company gains \$310 as a profit. If the customer dies, the company "gains" $\$310 - \$15,000 = -\$14,690$, or in other words, it loses \$14,690. Therefore, the expected profit can be calculated as follows:

$$\begin{aligned}\mu &= E(x) = \sum xp(x) \\ \mu &= (310)(99.9\%) + (310 - 15,000)(0.1\%) \\ &= (310)(0.999) + (310 - 15,000)(0.001) \\ &= 309.69 - 14.69 = \$295 \\ \mu &= \$295\end{aligned}$$

This tells us that if the company were to sell a very large number of the 1-year \$15,000 policies to many people, it would make, on average, a profit of \$295 per sale.

Another approach is to calculate the expected payout, not the expected gain:

$$\begin{aligned}\mu &= (0)(99.9\%) + (15,000)(0.1\%) \\ &= 0 + 15 \\ \mu &= \$15\end{aligned}$$

Since the company charges \$310 and expects to pay out \$15, the average profit for the company is \$295 per policy. Sometimes, we are interested in measuring not just the expected value of a random variable, but also the variability and the central tendency of a probability distribution. To do this, we first need to define population variance, or σ^2 . It is the average of the squared distance of the values of the random variable X from the mean value, μ . The formal definitions of variance and standard deviation are shown below.

The Variance

The variance of a discrete random variable is given by the following formula:

$$\sigma^2 = \sum (x - \mu)^2 P(x)$$

The Standard Deviation

The square root of the variance, or, in other words, the square root of σ^2 , is the standard deviation of a discrete random variable:

$$\sigma = \sqrt{\sigma^2}$$

Example: A university medical research center finds out that treatment of skin cancer by the use of chemotherapy has a success rate of 70%. Suppose five patients are treated with chemotherapy. The probability distribution of x successful cures of the five patients is given in the table below:

x	0	1	2	3	4	5
$p(x)$	0.002	0.029	0.132	0.309	0.360	0.168

Figure: Probability distribution of cancer cures of five patients.

- a) Find μ .
- b) Find σ .
- c) Graph $p(x)$ and explain how μ and σ can be used to describe $p(x)$.
- a. To find μ , we use the following formula:

$$\mu = E(x) = \sum x p(x)$$

$$\begin{aligned}\mu &= (0)(0.002) + (1)(0.029) + (2)(0.132) + (3)(0.309) + (4)(0.360) + (5)(0.168) \\ \mu &= 3.50\end{aligned}$$

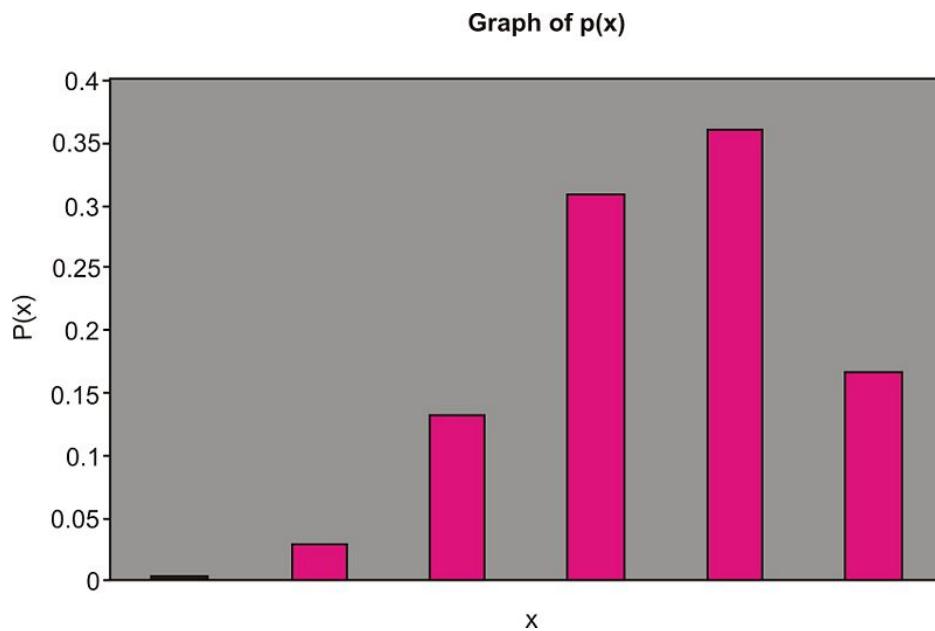
b. To find σ , we first calculate the variance of X :

$$\begin{aligned}\sigma^2 &= \sum(x - \mu)^2 p(x) \\ &= (0 - 3.5)^2(0.002) + (1 - 3.5)^2(0.029) + (2 - 3.5)^2(0.132) \\ &\quad + (3 - 3.5)^2(0.309) + (4 - 3.5)^2(0.36) + (5 - 3.5)^2(0.168) \\ \sigma^2 &= 1.05\end{aligned}$$

Now we calculate the standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} = 1.02$$

c. The graph of $p(x)$ is shown below:



We can use the mean, or μ , and the standard deviation, or σ , to describe $p(x)$ in the same way we used \bar{x} and s to describe the relative frequency distribution. Notice that $\mu = 3.5$ is the center of the probability distribution. In other words, if the five cancer patients receive chemotherapy treatment, we expect the number of them who are cured to be near 3.5. The standard deviation, which is $\sigma = 1.02$ in this case, measures the spread of the probability distribution $p(x)$.

Lesson Summary

The **mean**, or expected value, of the discrete random variable X is given by $\mu = E(x) = \sum xp(x)$.

The **variance** of the discrete random variable X is given by $\sigma^2 = \sum(x - \mu)^2 p(x)$.

The square root of the variance, or, in other words, the square root of σ^2 , is the **standard deviation** of a discrete random variable: $\sigma = \sqrt{\sigma^2}$.

Review Questions

1. Consider the following probability distribution:

x	0	1	2	3	4
$p(x)$	0.1	0.4	0.3	0.1	0.1

- (a) Calculate the mean of the distribution.
(b) Calculate the variance.
(c) Calculate the standard deviation.
2. An officer at a prison questioned each inmate to find out how many times the inmate had been convicted, prior to their current incarceration. The officer came up with the following table that shows the relative frequencies of X , the number of prior convictions:

x	0	1	2	3	4
$p(x)$	0.16	0.53	0.20	0.08	0.03

If we regard the relative frequencies as approximate probabilities, what is the expected value of the number of previous convictions of an inmate?

Answers: (1)(a) 1.7 (1)(b) 1.21 (1)(c) 1.1 (2) 1.29

4.4 Sums and Differences of Independent Random Variables

Learning Objectives

- Create probability distributions of independent random variables.
- Calculate the mean and standard deviation for sums and differences of independent random variables.

Introduction

A probability distribution is the set of values that a random variable can take on. At this time, there are three ways that you can create probability distributions from data. Sometimes previously collected data, relative to the random variable that you are studying, can help to create a probability distribution. In addition to this method, a simulation is also a good way to create an approximate probability distribution. A probability distribution can also be constructed from the basic principles, assumptions, and rules of theoretical probability. The examples in this lesson will lead you to a better understanding of these rules of theoretical probability.

Creating Probability Distributions

It is not difficult to create a probability distribution. Consider the following example.

Example: Silicosis is a pulmonary (lung) disease often afflicting mine workers. In a large study, assume that it is determined, of all silicosis sufferers, that 80% worked in mines. We use the Complement Rule to determine that 20% of silicosis sufferers did not work in mines.

Suppose two silicosis patients are randomly selected from a large population with the disease. We want to create the probability distribution of the possible categorization of the two patients for the source of the disease.

There are four possible outcomes for the two patients. With 'Y' representing "worked in the mines" and 'N' representing "did not work in the mines", the outcomes in the sample space are as follows:

- (N, N) (neither patient worked in the mines)
- (Y, N) (first patient worked in the mines, second patient did not)
- (N, Y) (first patient did not work in the mines, second patient did)
- (Y, Y) (both worked in the mines)

As stated previously, the patients for this survey have been randomly selected from a large population, and therefore, the outcomes are independent. The probability for each outcome can be calculated by using the Multiplication Rule for independent outcomes:

$$P(\text{no for 1}^{\text{st}}) \bullet P(\text{no for 2}^{\text{nd}}) = (0.2)(0.2) = 0.04$$

$$P(\text{yes for 1}^{\text{st}}) \bullet P(\text{no for 2}^{\text{nd}}) = (0.8)(0.2) = 0.16$$

$$P(\text{no for 1}^{\text{st}}) \bullet P(\text{yes for 2}^{\text{nd}}) = (0.2)(0.8) = 0.16$$

$$P(\text{yes for 1}^{\text{st}}) \bullet P(\text{yes for 2}^{\text{nd}}) = (0.8)(0.8) = 0.64$$

Notice that the sum of the 4 probabilities is 1, which is what we would expect for the sum of probabilities in a sample space.

If the random variable X represents the number of silicosis patients who worked in the mines in this random sample, then the first of these outcomes results in $x = 0$, the second and third each result in $x = 1$, and the fourth results in $x = 2$. Because the second and third outcomes are mutually exclusive, their probabilities can be added. The **probability distribution** for X is given in the table below:

TABLE 4.6:

x	Probability, $p(x)$
0	0.04
1	$0.16 + 0.16 = 0.32$
2	0.64

Example: A charity states that 40% of households make a donation of clothing or goods every year. We consider 3 households at random in the community. What is the probability distribution for the number of households, out of these 3, that will make a donation?

We let "Y" and "N" stand for "Yes, donation" and "No donation," respectively.

The possible outcomes are summarized in the table. Each probability is calculated by use of the Multiplication Rule for independent events.

TABLE 4.7:

X = No. of "Yes" Responses	Outcome	Probability $p(X)$
3	(YYY)	$(.4)(.4)(.4) = 0.064$
2	(YYN)	$(.4)(.4)(.6) = 0.096$
2	(YNY)	$(.4)(.6)(.4) = 0.096$
2	(NYY)	$(.6)(.4)(.4) = 0.096$
1	(YNN)	$(.4)(.6)(.6) = 0.144$
1	(NYN)	$(.6)(.4)(.6) = 0.144$
1	(NNY)	$(.6)(.6)(.4) = 0.144$
0	(NNN)	$(.6)(.6)(.6) = 0.216$

We can condense the table just obtained into a proper probability distribution by combining the probabilities for each value of the random variable X , as shown:

TABLE 4.8:

X	$p(X)$
3	0.064
2	0.288
1	0.432
0	0.216

We have now created the probability distribution for the number of households, out of 3, that make a donation of clothing or goods to the charity.

We can now calculate the mean and the standard deviation of this probability distribution. The mean is calculated using the formula $\mu = \sum(X)p(X)$. Here is the work performed in Excel:

Likewise, we can calculate the variance using the formula $\sigma^2 = \sum(X - \mu)^2 p(X)$. Then we take its square root to obtain the standard deviation. Here is the work shown in Excel:

X	p(X)	X p(X)
3	0.064	0.192
2	0.288	0.576
1	0.432	0.432
0	0.216	0
$\mu =$		1.2

FIGURE 4.1

X	p(X)	μ	$[X - \mu]$	$[X - \mu]^2$	$[X - \mu]^2 p(X)$
3	0.064	1.2	1.8	3.24	0.20736
2	0.288	1.2	0.8	0.64	0.18432
1	0.432	1.2	-0.2	0.04	0.01728
0	0.216	1.2	-1.2	1.44	0.31104
				Variance =	0.72
				SD =	0.848528137

FIGURE 4.2

Linear Transformations of X on Mean of X and Standard Deviation of X

Let's revisit a topic from an earlier lesson. A professor has a set of students' exam scores with a mean of 60 and a standard deviation of 10. We previously learned that if the professor adds 10 points to everyone's score, the new mean will be 70 and the standard deviation will be unchanged.

Similarly, if the professor multiplied each score by 2, then the new mean would be 120, with a standard deviation which is now 20.

We summarize this by saying that when the same value is added to every data value, the mean μ is increased by that value, but the standard deviation is unchanged. For multiplication, when each value in a data set is multiplied by a constant, then we multiply both the mean and the standard deviation by that constant.

We now extend this basic rule to a new situation: What happens if we combine two probability distributions? We will use an example to show this.

Example: Suppose a large company is considering combining operations of two customer call centers, in an attempt to save money. The first call center has the following probability distribution:

TABLE 4.9:

X = No. of calls per hour	Probability p(X)
20	0.1
25	0.5
30	0.3
35	0.1

The second call center has the following probability distribution:

TABLE 4.10:

Y = No. of calls per hour	Probability p(Y)
30	0.2
35	0.2
40	0.4
45	0.2

The mean, variance, and standard deviation of each of these 2 probability distributions have been calculated, and here are the **individual results**:

TABLE 4.11:

Probability Distribution	Mean μ	Variance σ^2	Standard Deviation σ
First Call Center (X)	27	16	4
Second Call Center (Y)	38	26	5.1

The company executives need to know if it would make sense to combine the two call centers, but they aren't sure what the probability distribution of the **combined** data would look like. An enterprising young employee, who took a Statistics class (and got an A), dutifully creates a combined probability distribution:

TABLE 4.12:

Z = Combined No. of Calls per Hour	Probability p(Z)
50	0.02
55	0.12
60	0.20
65	0.30
70	0.24
75	0.10
80	0.02

He then calculates the mean, standard deviation, and variance of the combined probability distribution:

The company executives are unsure if this young man is correct in saying that the combined call center would have a mean of 65 calls per hour and a standard deviation of 6.5 calls per hour. They hire a professional statistician to verify his work. The statistician congratulates the young man on his hard work and recommends that the executives give him a promotion! However, the statistician says that the young employee worked much too hard to get the answer. "There is an easier way to do this," says the statistician. "Here are the rules for combining two probability distributions:"

Z	p(Z)	Z p(Z)		Z	p(Z)	μ_z	$[Z - \mu_z]$	$[Z - \mu_z]^2$	$[Z - \mu_z]^2 p(Z)$
50	0.02	1		50	0.02	65	-15	225	4.5
55	0.12	6.6		55	0.12	65	-10	100	12
60	0.2	12		60	0.2	65	-5	25	5
65	0.3	19.5		65	0.3	65	0	0	0
70	0.24	16.8		70	0.24	65	5	25	6
75	0.1	7.5		75	0.1	65	10	100	10
80	0.02	1.6		80	0.02	65	15	225	4.5
$\mu_z = 65$									
Variance = 42									
SD = 6.480740698									

FIGURE 4.3

Rule 1: To get the mean of a combined probability distribution, simply add the means of each of the individual probability distributions.

If you look at the table of individual results, you will see that the mean of the first call center was 27, and the mean of the second call center was 38. We add these two values and get 65 for the mean of the combined call center, which is exactly what our enterprising young employee obtained.

Rule 2: To get the variance of a combined probability distribution, simply add the variances of each of the individual probability distributions.

Again, look at the table of individual results, and you see that the variance of the first call center was 16, and the variance of the second call center was 26. We add these two values and get 42 as the combined variance. This agrees with the employee's calculations.

Rule 3: To get the standard deviation of a combined probability distribution, DO NOT add the standard deviations!!!! Instead, use Rule 2 and add the variances. Then take the square root of the sum of the variances, and you will get the combined standard deviation.

We just saw that the combined variance is 42. So, to get the standard deviation of the combined probability distribution, take the square root of 42, which is rounded to 6.5. Again, this agrees with the calculations of the enterprising young employee.

The company executives are impressed with the young man's initiative and hard work, but instead of giving him a promotion, they give him a small bonus. They justify this by telling him that, while he was correct in his work, he was highly inefficient because he could have gotten the answers they needed much more easily if he had remembered the 3 rules of combining probability distributions.

So did they decide to combine the call centers? Yes, they did. Their reasoning was this: For call center 1, they needed 3 employees to handle the mean of 27 calls per hour. For call center 2, they needed 4 employees to handle the mean of 38 calls per hour. If they combined the 2 call centers, with a combined mean of 65 calls per hour, they decided that they could have 6 employees (instead of 7), and that meant that customers might have to wait a little longer for assistance. But the company would save money by cutting one employee position.

This next example combines all of the rules learned thus far for transforming distributions.

Example: Beth earns \$25.00 an hour for tutoring but spends \$20.00 an hour for piano lessons. She saves the difference between her earnings for tutoring and the cost of the piano lessons. The numbers of hours she spends on each activity in one week vary independently according to the probability distributions shown below. Determine her expected weekly savings and the standard deviation of these savings.

TABLE 4.13: (continued)

Hours of Piano Lessons, x	Probability, $p(x)$
TABLE 4.13:	
Hours of Piano Lessons, x	Probability, $p(x)$
0	0.3
1	0.3
2	0.4

TABLE 4.14:

Hours of Tutoring, y	Probability, $p(y)$
1	0.2
2	0.3
3	0.2
4	0.3

X represents the number of hours per week taking piano lessons, and Y represents the number of hours tutoring per week. The mean and standard deviation for each can be calculated as follows:

$$\begin{aligned} E(x) &= \mu_X = \sum xp(x) & \sigma^2_x &= \sum (x - \mu_X)^2 p(x) \\ \mu_X &= (0)(0.3) + (1)(0.3) + (2)(0.4) & \sigma^2_x &= (0 - 1.1)^2(0.3) + (1 - 1.1)^2(0.3) + (2 - 1.1)^2(0.4) \\ \mu_X &= 1.1 & \sigma^2_x &= 0.69 \\ & & \sigma_x &= 0.831 \end{aligned}$$

$$\begin{aligned} E(y) &= \mu_Y = \sum yp(y) & \sigma^2_y &= \sum (y - \mu_Y)^2 p(y) \\ \mu_Y &= (1)(0.2) + (2)(0.3) + (3)(0.2) + (4)(0.3) & \sigma^2_y &= (1 - 2.6)^2(0.2) + (2 - 2.6)^2(0.3) + (3 - 2.6)^2(0.2) \\ & & & + (4 - 2.6)^2(0.3) \\ \mu_Y &= 2.6 & \sigma^2_y &= 1.24 \\ & & \sigma_y &= 1.11 \end{aligned}$$

The expected number of hours Beth spends on piano lessons is 1.1 with a standard deviation of 0.831 hours. Likewise, the expected number of hours Beth spends tutoring is 2.6 with a standard deviation of 1.11 hours.

Beth spends \$20 for each hour of piano lessons, so her mean weekly cost for piano lessons can be calculated as shown:

$$\mu_{20X} = (20)(\mu_X) = (20)(1.1) = \$22$$

Beth earns \$25 for each hour of tutoring, so her mean weekly earnings from tutoring are as follows:

$$\mu_{25Y} = (25)(\mu_Y) = (25)(2.6) = \$65$$

Thus, Beth's expected weekly savings are:

$$\mu_{25Y} - \mu_{20X} = \$65 - \$22 = \$43$$

The standard deviation of the cost of her piano lessons is:

$$\sigma_{20X} = (20)(0.831) = \$16.62$$

The standard deviation of her earnings from tutoring is:

$$\sigma_{25Y} = (25)(1.11) = \$27.75$$

Finally, the variance and standard deviation of her weekly savings is:

$$\begin{aligned}\sigma^2_{25Y-20X} &= \sigma^2_{25Y} + \sigma^2_{20X} = (27.75)^2 + (16.62)^2 = 1046.2896 \\ \sigma_{25Y-20X} &\approx \$32.35\end{aligned}$$

Note that the new variance was obtained by adding the individual variances, whereas the standard deviation was obtained by taking the square root of the new variance.

Lesson Summary

A chance process can be displayed as a probability distribution that describes all the possible outcomes, x . You can also determine the probability of any set of possible outcomes. A probability distribution table for a random variable, X , consists of a table with all the possible outcomes, along with the probability associated with each of the outcomes. The expected value and the variance of a probability distribution can be calculated using the following formulas:

$$\begin{aligned}E(x) &= \mu_X = \sum xp(x) \\ \sigma^2_x &= \sum (x - \mu_X)^2 p(x)\end{aligned}$$

When we add a constant to every data value, the mean is increased by the value of the constant. (This also works for subtraction.) The standard deviation is unchanged.

When we multiply every data value by a constant, the mean is multiplied by that constant. The standard deviation is also multiplied by that constant.

We also learned 3 rules for combining probability distributions:

Rule 1: To get the mean of a combined probability distribution, simply add the means of each of the individual probability distributions.

Rule 2: To get the variance of a combined probability distribution, simply add the variances of each of the individual probability distributions.

Rule 3: To get the standard deviation of a combined probability distribution, DO NOT add the standard deviations!!!! Instead, use Rule 2 and add the variances. Then take the square root of the sum of the variances, and you will get the combined standard deviation.

Review Questions

- It is estimated that 70% of the students attending a school in a rural area take the bus to school. Suppose you randomly select three students from the population. Construct the probability distribution of the random variable, X , defined as the number of students who take the bus to school. (Hint: Begin by listing all of the possible outcomes.)

2. The Safe Grad Committee at a high school is selling raffle tickets on a Christmas Basket filled with gifts and gift cards. The prize is valued at \$1200, and the committee has decided to sell only 500 tickets.
- Create the probability distribution. What is the expected value of a ticket?
 - If the ticket price is \$5.00, what is the profit per ticket?
 - If the students decide to sell tickets on three monetary prizes – one valued at \$1500 dollars and two valued at \$500 each, what is the expected value of the ticket now?

Students at an elementary school are going on a field trip, and they are allowed to choose their own snack pack, which consists of a beverage and a treat. Based on previous years, the teacher has calculated the following probability distribution for the beverage choice:

TABLE 4.15:

Beverage	$X = \text{Weight (in ounces)}$	Probability $p(X)$
Small	6	0.2
Medium	8	0.5
Large	12	0.3

Here is the probability distribution for the treat:

TABLE 4.16:

Treat	$Y = \text{Weight (in ounces)}$	Probability $p(Y)$
Energy Bar	1	0.1
Chewy Bears	2	0.2
Brownie	3	0.7

Assuming that every child chooses exactly 1 beverage and 1 treat, calculate the mean, variance, and standard deviation for the combined weight of a snack pack for this school.

Answers: (1) see detailed answers (2)(a) \$2.40 (2)(b) \$2.60 (2)(c) expected value is 5 and standard deviation is about 74 (3) mean = 11.4 variance = 5.4 standard deviation = 2.32

4.5 The Binomial Probability Distribution

Learning Objectives

3. List the 4 conditions for a binomial experiment.

- Link the concept of a probability distribution to a binomial probability distribution.
- Define and calculate the mean, the variance, and the standard deviation of a binomial random variable.
- Identify the type of statistical situation to which a binomial distribution can be applied.
- Use a binomial distribution to solve statistical problems.

Introduction

Many experiments result in responses for which there are only two possible outcomes, such as either 'yes' or 'no', 'pass' or 'fail', 'good' or 'defective', 'male' or 'female', etc. A simple example is the toss of a coin. Say, for example, that we toss the coin five times. In each toss, we will observe either a head, H , or a tail, T . We might be interested in the probability distribution of X , the number of heads observed. In this case, the possible values of X range from 0 to 5. It is scenarios like this that we will examine in this lesson

Binomial Experiments

Example: Suppose we select 100 students from a large university campus and ask them whether they are in favor of a certain issue that is being debated on their campus. The students are to answer with either a 'yes' or a 'no'. Here, we are interested in X , the number of students who favor the issue (a 'yes'). If each student is randomly selected from the total population of the university, and the proportion of students who favor the issue is p , then the probability that any randomly selected student favors the issue is p . The probability of a selected student who does not favor the issue is $1 - p$. Sampling 100 students in this way is equivalent to tossing a coin 100 times. This experiment is an example of a **binomial experiment**.

Characteristics of a Binomial Experiment

- The experiment consists of n independent, identical trials.
- There are only two possible outcomes on each trial: S (for success) or F (for failure).
- The probability of S remains constant from trial to trial. We will denote it by p . We will denote the probability of F by q . Thus, $q = 1 - p$.
- The binomial random variable X is the number of successes in n trials.

Example: In the following two examples, decide whether X is a binomial random variable.

Example: Suppose a university decides to give two scholarships to two students. The pool of applicants is ten students: six males and four females. All ten of the applicants are equally qualified, and the university decides to randomly select two. Let X be the number of female students who receive the scholarship.

If the first student selected is a female, then the probability that the second student is a female is $\frac{3}{9}$. Here we have a conditional probability: the success of choosing a female student on the second trial depends on the outcome of the first trial. Therefore, the trials are not independent, and X is not a binomial random variable.

Example: A company decides to conduct a survey of customers to see if its new product, a new brand of shampoo, will sell well. The company chooses 100 randomly selected customers and asks them to state their preference among the new shampoo and another shampoo that has been popular for many years. Let X be the number of the 100 customers who choose the new brand over the old brand.

In this experiment, each customer either states a preference for the new shampoo or does not. The customers' preferences are independent of each other, and therefore, X is a binomial random variable.

Let's examine an actual binomial situation. Suppose we present four people with two cups of coffee (one percolated and one instant) to discover the answer to this question: "If we ask four people which is percolated coffee and none of them really can tell the percolated coffee from the instant coffee, what is the probability that two of the four will guess correctly?" We will present each of four people with percolated and instant coffee and ask them to identify the percolated coffee. The outcomes will be recorded by using C for correctly identifying the percolated coffee and I for incorrectly identifying it. A list of the 16 possible outcomes, all of which are equally likely if none of the four can tell the difference and are merely guessing, is shown below:

TABLE 4.17:

Number Who Correctly Identify Percolated Coffee	Outcomes, C (correct), I (incorrect)	Number of Outcomes
0	III	1
1	IIC IIC IIC IIC	4
2	$ICCI$ $IICC$ $ICIC$ $CIIC$ $CICI$ $CCII$	6
3	$CICC$ $ICCC$ $CCCI$ $CCIC$	4
4	$CCCC$	1

Using the Multiplication Rule for Independent Events, you know that the probability of getting a certain outcome when two people guess correctly, such as $CICI$, is $(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2}) = (\frac{1}{16})$. The table shows six outcomes where two people guessed correctly, so the probability of getting two people who correctly identified the percolated coffee is $\frac{6}{16}$.

A binomial experiment is a probability experiment that satisfies the following conditions:

- Each trial can have only two outcomes—one known as a success, and the other known as a failure.
- There must be a fixed number, n , of trials.
- The outcomes of the trials must be independent of each other. The probability of each success doesn't change, regardless of what occurred previously.
- The probability, p , of a success must remain the same for each trial.

The distribution of the random variable X , where x is the number of successes, is called a *binomial probability distribution*. The probability that you get exactly $x = k$ successes is as follows:

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

To apply the binomial formula to a specific problem, it is useful to have an organized strategy. Such a strategy is presented in the following steps:

- Identify a success.
- Determine p , the probability of success.
- Determine n , the number of experiments or trials.
- Identify x , the number of outcomes of interest for which a probability is desired.

- Use the binomial probability formula or your graphing calculator to calculate the probability of interest.

Example: According to a study conducted by a survey company, the probability is 25% that a randomly chosen household will have a cat. What is the probability that out of three randomly selected households:

- Exactly two will have a cat?
- None will have a cat?

Using the first three steps listed above:

- A success is a household that has a cat.
- The probability of success (cat) is $p = 0.25$. The probability of failure (no cat) is $(1 - p) = 0.75$.
- The number of trials is $n = 3$.

Thus, we can now use the binomial probability formula:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Substituting, we have: $p(x) = \binom{3}{x} (0.25)^x (1 - 0.25)^{3-x}$

- For $x = 2$ cats (out of the 3 households)

$$\begin{aligned} p(x) &= \binom{3}{2} (0.25)^2 (1 - 0.25)^{3-2} \\ &= (3)(0.25)^2 (1 - 0.25)^1 \\ &= 0.14 \end{aligned}$$

Thus, the probability is 0.14 that exactly two out of three randomly selected households own a cat.

- Here, $x = 0$ cats (out of the 3 households). Again, we use the binomial probability formula:

$$\begin{aligned} p(x=0) &= \binom{3}{0} (0.25)^0 (1 - 0.25)^{3-0} \\ &= \frac{3!}{0!(3-0)!} (0.25)^0 (0.75)^3 \\ &= 0.422 \end{aligned}$$

Thus, the probability is 0.422 that none of the three randomly selected households own a cat.

Example: A car dealer knows from past experience that he can make a sale to 20% of the customers who he interacts with. What is the probability that, in five randomly selected interactions, he will make a sale to exactly 3 customers?

A success here is making a sale to a customer. The probability that the car dealer makes a sale to any customer is $p = 0.20$, and the number of trials is $n = 5$. Therefore, the binomial probability formula for this case is:

$$p(x) = \binom{5}{x} (0.2)^x (0.8)^{5-x}$$

- Here we want the probability of exactly 3 sales, so $x = 3$.

$$p(x) = \binom{5}{3} (0.2)^3 (0.8)^{5-3} = 0.051$$

This means that the probability that the car dealer makes exactly three sales in five attempts is 0.051.

Using the TI-83/84 calculator to calculate binomial probabilities

The use of the binomial formula can be bulky. Your calculator will do the arithmetic for you. Let's do the same problem (above) about the car dealer, but we'll use the calculator instead.

1. Turn the calculator on. Press the blue **2^{nd}** button and then the **VARS** button to get to the Distributions menu.
2. Scroll down to Option A: **binompdf(** and press **Enter**.
3. Enter the number of trials: 5. Your value of p is .20, and your x value (the number of sales you are interested in) is equal to 3. Enter these three values, and scroll down to Paste and then press **Enter**.
4. At the new screen, the data inputs will be summarized as **binompdf (5, .2, 3)**. Press **Enter**. The answer will appear as shown here:

```
binompdf(5, .2, 3)
.0512
```

FIGURE 4.4

This answer agrees with the answer obtained by using the binomial formula.

Let's revisit the car dealer. This time we want to calculate the probability that he makes at most one sale out of 5 interactions. This means he will make either no sales ($x = 0$) or 1 sale ($x = 1$). The probability that the car dealer makes a sale to at most one customer can be calculated as follows, using the binomial formula:

$$\begin{aligned} p(x \leq 1) &= p(0) + p(1) \\ &= \binom{5}{0} (0.2)^0 (0.8)^{5-0} + \binom{5}{1} (0.2)^1 (0.8)^{5-1} \\ &= 0.328 + 0.410 = 0.738 \end{aligned}$$

This is a lot of work that your calculator can do for you. You could repeat the calculator instructions above twice, once for $x = 0$ and once again for $x = 1$ to get the answer, but there is an even easier way to calculate multiple binomial probabilities:

1. Press the blue **2^{nd}** button and then the **VARS** button to get to the Distributions menu.
2. This time, scroll down to Option B: **binomcdf(** and press **Enter**. (This is called the binomial *cumulative* distribution function.)

- Enter the number of trials: 5. Your value of p is .20, and your x value will be 1. (When you type in the x value of 1, you are telling the calculator to calculate all probabilities for x values of 1 or less. In this case, you are telling it to calculate the cumulative probabilities for $x = 1$ and for $x = 0$.) Scroll down to Paste and press **Enter**.
- At the new screen, the data inputs will be summarized as $\text{binomcdf}(5, .2, 1)$. Press **Enter**. The answer will appear as shown here:

A screenshot of a TI-Nspire CX CAS calculator's software interface. The input line at the top shows the command `binomcdf(5, .2, 1)`. Below the input line, the result `.73728` is displayed in a large black font. To the left of the result, there is a small icon of a black square.

FIGURE 4.5

Again, this answer agrees with the one obtained by using the binomial formula.

Let's calculate another binomial probability for the car dealer. The probability that the car dealer makes **at least** one sale is the sum of the probabilities of his making 1, 2, 3, 4, or 5 sales, as is shown below:

$$p(x \geq 1) = p(1) + p(2) + p(3) + p(4) + p(5)$$

We can now apply the binomial probability formula to calculate the five probabilities. However, we can save time by calculating the complement of the probability we're looking for and subtracting it from 1 as follows:

$$\begin{aligned} p(x \geq 1) &= 1 - p(x < 1) = 1 - p(x = 0) \\ 1 - p(0) &= 1 - \binom{5}{0} (0.2)^0 (0.8)^{5-0} \\ &= 1 - 0.328 = 0.672 \end{aligned}$$

This tells us that the salesperson has a probability of 0.672 of making at least one sale in five attempts.

We can calculate this probability with the calculator by using Option A (binompdf) for $n = 5$, $p = .2$, and $x = 0$. Calculate the complement of this probability by subtracting it from 1.

We are also asked to determine the probability distribution for the number of sales, X , in five attempts. Therefore, we need to compute $p(x)$ for 0, 1, 2, 3, 4, and 5. We can use the binomial probability formula or our calculator for each value of X . The table below shows the probabilities.

TABLE 4.18:

x	$p(x)$
0	0.328
1	0.410
2	0.205
3	0.051
4	0.006
5	0.00032

Figure: The probability distribution for the number of sales.

Mean and Standard Deviation of a Binomial Distribution

For a random variable X having a binomial distribution with n trials and a probability of success of p , the expected value (mean) and standard deviation for the distribution can be determined by the following formulas:

$$E(x) = \mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)}$$

Example: A poll of twenty voters is taken to determine the number in favor of a certain candidate for mayor. Suppose that 60% of all the city's voters favor this candidate.

- a. Find the mean and the standard deviation of X .
- b. Find the probability of $x \leq 10$.
- c. Find the probability of $x > 12$.
- d. Find the probability of $x = 11$.

a. Since the sample of twenty was randomly selected, it is likely that X is a binomial random variable. Of course, X here would be the number of the twenty who favor the candidate. The probability of success is 0.60, the percentage of the total voters who favor the candidate. Therefore, the mean and the standard deviation can be calculated as shown:

$$\begin{aligned}\mu &= np = (20)(0.6) = 12 \\ \sigma^2 &= np(1-p) = (20)(0.6)(0.4) = 4.8 \\ \sigma &= \sqrt{4.8} = 2.2\end{aligned}$$

b. To calculate the probability that 10 or fewer of the voters favor the candidate, it's possible to add the probabilities that 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 of the voters favor the candidate as follows:

$$p(x \leq 10) = p(0) + p(1) + p(2) + \dots + p(10)$$

or

$$p(x \leq 10) = \sum_{x=0}^{10} p(x) = \sum_{x=0}^{10} \binom{20}{x} (0.6)^x (0.4)^{20-x}$$

As you can see, this would be a very tedious calculation, and it is best to resort to your calculator. Use Option B (the cumulative binomial probability function), and enter $n = 10$, $p = 0.6$, and $x = 10$. The answer is .245.

c. To find the probability that $x > 12$, it's possible to add the probabilities that 13, 14, 15, 16, 17, 18, 19, or 20 of the voters favor the candidate as shown:

$$p(x > 12) = p(13) + p(14) + \dots + p(20) = \sum_{x=13}^{20} p(x)$$

Alternatively, using the Complement Rule, $p(x > 12) = 1 - p(x \leq 12)$.

Using a calculator (Option B, the cumulative binomial probability function) with $n = 20$, $p = 0.6$, and $x = 12$, we get a probability of 0.584 that $x \leq 12$. Thus, $p(x > 12) = 1 - 0.584 = 0.416$.

d. To find the probability that exactly 11 voters favor the candidate, use your calculator Option A, binompdf (probability distribution function), as shown earlier. It calculates a binomial probability for a single outcome. The answer is 0.16.

Lesson Summary

Characteristics of a Binomial Experiment:

- A binomial experiment consists of n identical trials.
- There are only two possible outcomes on each trial: S (for success) or F (for failure).
- The probability of S remains constant from trial to trial. We denote it by p . We denote the probability of F by $(1 - p)$.
- The trials are independent of each other.
- The binomial random variable X is the number of successes in n trials.

The binomial probability distribution is: $p(x) = \binom{n}{x} p^x (1 - p)^{n-x} = \binom{n}{x} p^x q^{n-x}$.

For a binomial random variable, the mean is $\mu = np$.

The variance is $\sigma^2 = npq = np(1 - p)$.

The standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1 - p)}$.

Review Questions

1. Suppose X is a binomial random variable with $n = 4$ and $p = 0.2$. Calculate $p(x)$ for each of the following values of X : 0, 1, 2, 3, 4. Give the probability distribution in tabular form.
2. Suppose X is a binomial random variable with $n = 5$ and $p = 0.2$. Display $p(x)$ in tabular form. Compute the mean and the variance of X .
3. Over the years, a medical researcher has found that 10% of diabetic patients receiving insulin develop antibodies against the hormone, thus requiring a more costly form of medication. The researcher chooses five patients at random.
 - a. Find the probability none will develop antibodies against insulin.
 - b. Find the probability that at least one will develop antibodies.
 - c. Find the probability that fewer than three will develop antibodies.
 - d. Find the probability that more than four will develop antibodies.
4. According to the Canadian census of 2006, the median annual family income for families in Nova Scotia is \$56,400. [Source: Stats Canada. www.statcan.ca] Consider a random sample of 24 Nova Scotia households. (Reminder: the median is the 50th percentile; half are above it and half are below it.)
 - a. What is the expected number of households from the sample that will have annual incomes less than \$56,400?
 - b. What is the standard deviation of the number of households in our sample of 24 that will have annual incomes less than \$56,400?
 - c. What is the probability of getting at least 18 out of the 24 households with annual incomes under \$56,400?

Answers: (1) $p(0) = 0.4096$ $p(1) = 0.4096$ $p(2) = 0.1536$ $p(3) = 0.0256$ $p(4) = 0.0016$ (2) $p(0) = 0.328$ $p(1) = 0.4096$ $p(2) = 0.2048$ $p(3) = 0.0512$ $p(4) = 0.0064$ $p(5) = 0.0003$ $\mu = 1$ $\sigma = 0.894$. (3)(a) 0.59 (3)(b) 0.41 (3)(c) 0.991 (3)(d) 0.0003

4.6 References

1. . . CC BY-NC-SA
2. . . CC BY-NC-SA
3. . . CC BY-NC-SA
4. . . CC BY-NC-SA
5. . . CC BY-NC-SA

CHAPTER**5****Normal Distribution****Chapter Outline**

- 5.1 THE STANDARD NORMAL PROBABILITY DISTRIBUTION**
 - 5.2 THE DENSITY CURVE OF THE NORMAL DISTRIBUTION**
 - 5.3 APPLICATIONS OF THE NORMAL DISTRIBUTION**
-

5.1 The Standard Normal Probability Distribution

Learning Objectives

- Identify the characteristics of a normal distribution.
- Identify and use the Empirical Rule (68-95-99.7 Rule) for normal distributions.
- Calculate a z -score and relate it to probability.
- Determine if a data set corresponds to a normal distribution.

Introduction

Most high schools have a set amount of time in-between classes during which students must get to their next class. If you were to stand at the door of your statistics class and watch the students coming in, think about how the students would enter. Usually, one or two students enter early, then more students come in, then a large group of students enter, and finally, the number of students entering decreases again, with one or two students barely making it on time, or perhaps even coming in late!

Now consider this. Have you ever popped popcorn in a microwave? Think about what happens in terms of the rate at which the kernels pop. For the first few minutes, nothing happens, and then, after a while, a few kernels start popping. This rate increases to the point at which you hear most of the kernels popping, and then it gradually decreases again until just a kernel or two pops.

Here's something else to think about. Try measuring the height, shoe size, or the width of the hands of the students in your class. In most situations, you will probably find that there are a couple of students with very low measurements and a couple with very high measurements, with the majority of students centered on a particular value.

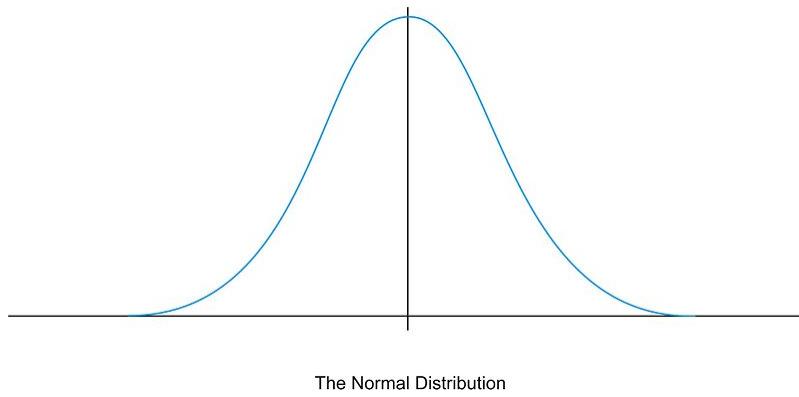


All of these examples show a typical pattern that seems to be a part of many real-life phenomena. In statistics, because this pattern is so pervasive, it seems to fit to call it normal, or more formally, the normal distribution. The normal distribution is an extremely important concept, because it occurs so often in the data we collect from the natural world, as well as in many of the more theoretical ideas that are the foundation of statistics. This chapter explores the details of the normal distribution.

The Characteristics of a Normal Distribution

Shape

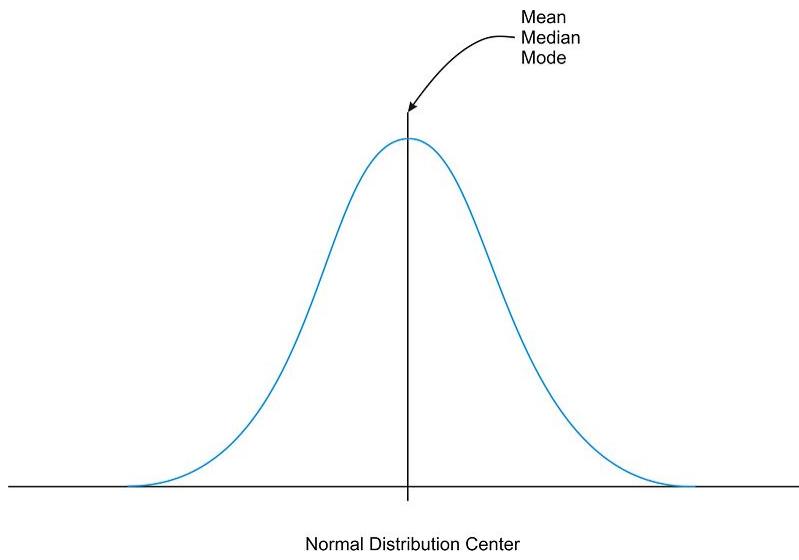
When graphing the data from each of the examples in the introduction, the distributions from each of these situations would be mound-shaped and mostly symmetric. A *normal distribution* is a perfectly symmetric, mound-shaped distribution. It is commonly referred to as a normal curve, or bell curve.



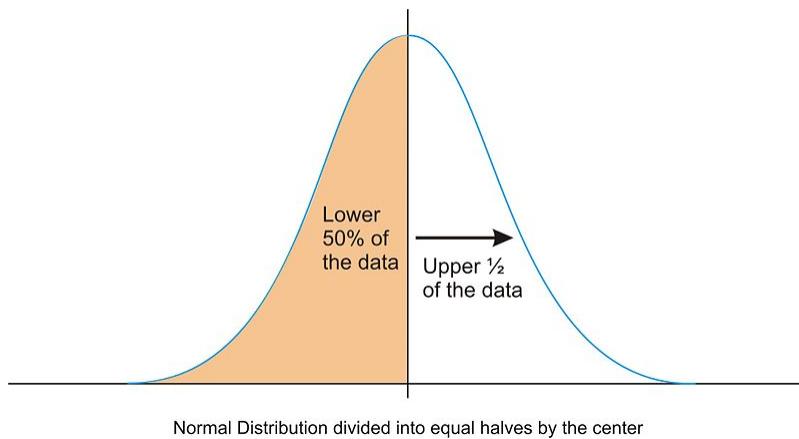
Because so many real data sets closely approximate a normal distribution, we can use the idealized normal curve to learn a great deal about such data. With a practical data collection, the distribution will never be exactly symmetric, so just like situations involving probability, a true normal distribution only results from an infinite collection of data. Also, it is important to note that the normal distribution describes a continuous random variable.

Center

Due to the exact symmetry of a normal curve, the center of a normal distribution, or a data set that approximates a normal distribution, is located at the highest point of the distribution, and all the statistical measures of center we have already studied (the mean, median, and mode) are equal.

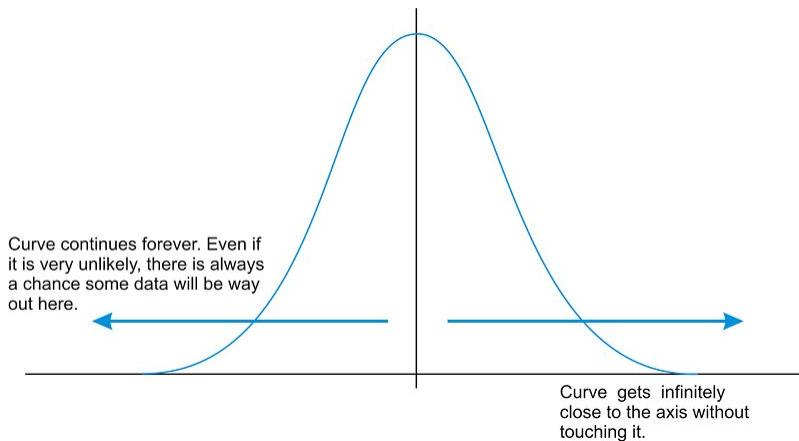


It is also important to realize that this center peak divides the data into two equal parts.

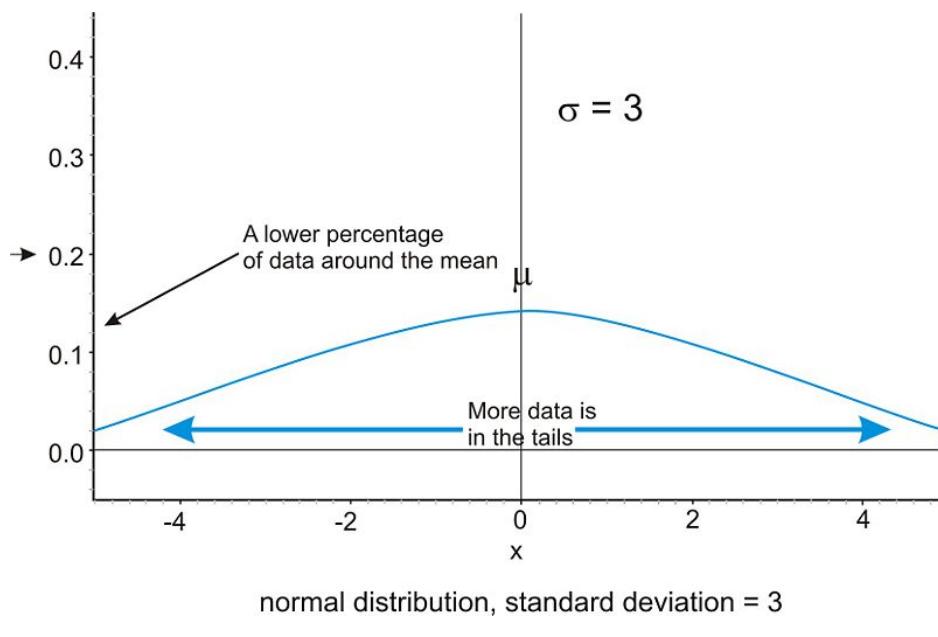
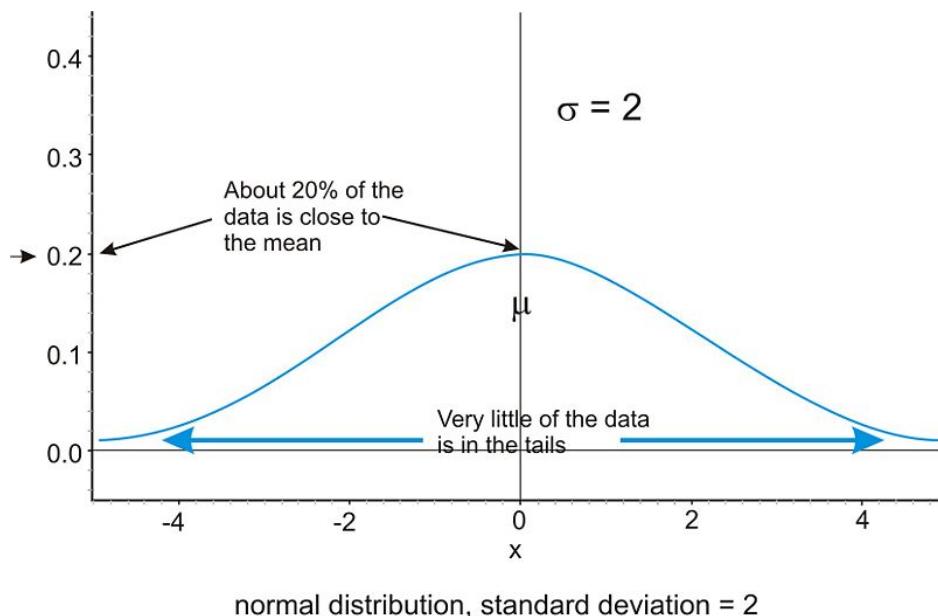


Spread

Let's go back to our popcorn example. The bag advertises a certain time, beyond which you risk burning the popcorn. From experience, the manufacturers know when most of the popcorn will stop popping, but there is still a chance that there are those rare kernels that will require more (or less) time to pop than the time advertised by the manufacturer. The directions usually tell you to stop when the time between popping is a few seconds, but aren't you tempted to keep going so you don't end up with a bag full of un-popped kernels? Because this is a real, and not theoretical, situation, there will be a time when the popcorn will stop popping and start burning, but there is always a chance, no matter how small, that one more kernel will pop if you keep the microwave going. In an idealized normal distribution of a continuous random variable, the distribution continues infinitely in both directions.



Because of this infinite spread, the range would not be a useful statistical measure of spread. The most common way to measure the spread of a normal distribution is with the standard deviation, or the typical distance away from the mean. Because of the symmetry of a normal distribution, the standard deviation indicates how far away from the maximum peak the data will be. Here are two normal distributions with the same center (mean):



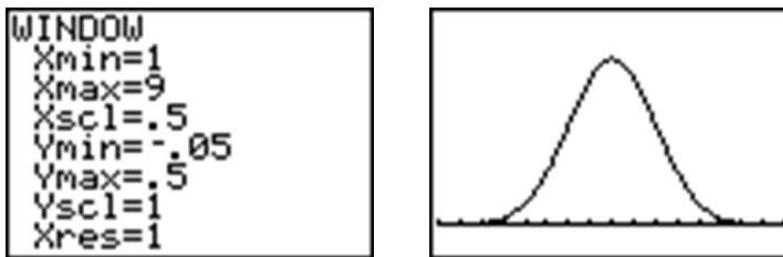
The first distribution pictured above has a smaller standard deviation, and so more of the data are heavily concentrated around the mean than in the second distribution. Also, in the first distribution, there are fewer data values at the extremes than in the second distribution. Because the second distribution has a larger standard deviation, the data are spread farther from the mean value, with more of the data appearing in the tails.

Technology Note: Investigating the Normal Distribution on a TI-83/84 Graphing Calculator

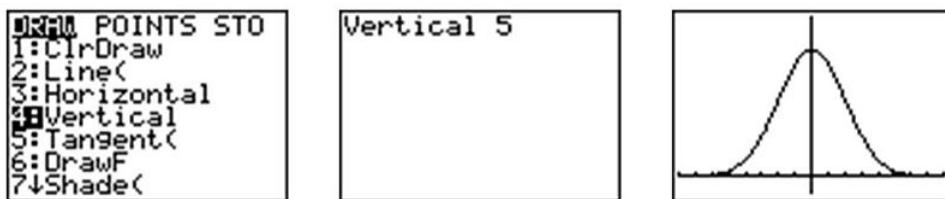
We can graph a normal curve for a probability distribution on the TI-83/84 calculator. To do so, first press [Y=], which is the upper left-most button on the TI-84. To create a normal distribution, we will draw an idealized curve using something called a density function. The command is called 'normalpdf()', and it is found by pressing [2nd][DISTR][1]. Enter an X by pressing the calculator button with the symbols (X,T,θ,n) on it, to represent the random variable, followed by the mean and the standard deviation, all separated by commas. For this example, choose a mean of 5 and a standard deviation of 1.

DRAW DRAW 1: normalpdf(2: normalcdf(3: invNorm(4: invT(5: tpdf(6: tcdf(7: χ^2 pdf(Plot1 Plot2 Plot3 $\text{Y}_1 = \text{normalpdf}(X,$ 5, 1) $\text{Y}_2 =$ $\text{Y}_3 =$ $\text{Y}_4 =$ $\text{Y}_5 =$ $\text{Y}_6 =$
---------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------

Press the [window] button at the top row of buttons. Adjust your window to match the following settings and press [**GRAPH**].



Press [**2ND**][**QUIT**] to go to the home screen. We can draw a vertical line at the mean to show it is in the center of the distribution by pressing [**2ND**][**DRAW**] and choosing 'Vertical'. Enter the mean, which is 5, and press [**ENTER**].

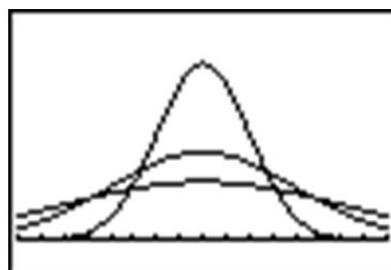


Remember that even though the graph appears to touch the x -axis, it is actually just very close to it.

In your **Y=** Menu, enter the following to graph 3 different normal distributions, each with a different standard deviation:

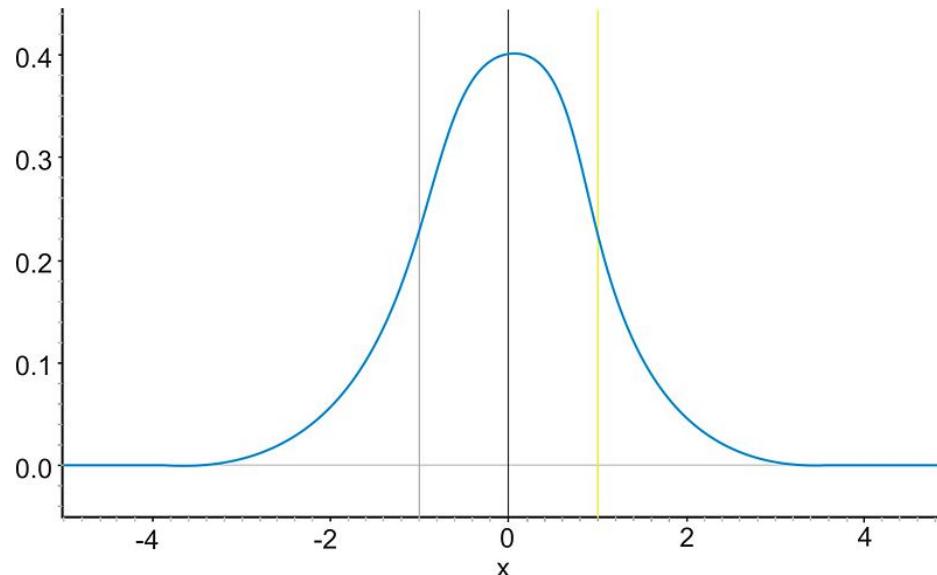
Plot1 Plot2 Plot3 $\text{Y}_1 = \text{normalpdf}(X,$ 5, (1, 2, 3)) $\text{Y}_2 =$ $\text{Y}_3 =$ $\text{Y}_4 =$ $\text{Y}_5 =$ $\text{Y}_6 =$

This makes it easy to see the change in spread when the standard deviation changes.



The Empirical Rule

Because of the similar shape of all normal distributions, we can measure the percentage of data that is a certain distance from the mean no matter what the standard deviation of the data set is. The following graph shows a normal distribution with $\mu = 0$ and $\sigma = 1$. This curve is called a *standard normal curve*. In this case, the values of x represent the number of standard deviations away from the mean.



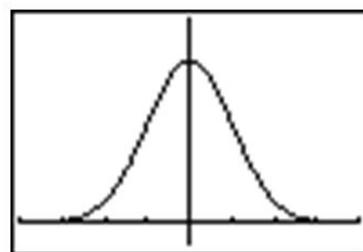
The Standard Normal Distribution

Notice that vertical lines are drawn at points that are exactly one standard deviation to the left and right of the mean. We have consistently described standard deviation as a measure of the typical distance away from the mean. How much of the data is actually within one standard deviation of the mean? To answer this question, think about the space, or area, under the curve. The entire data set, or 100% of it, is contained under the whole curve. What percentage would you estimate is between the two lines? To help estimate the answer, we can use a graphing calculator. Graph a standard normal distribution over an appropriate window. Note that you have to go back to the [Y=] button to change to a mean of 0 and a SD of 1. Also, go to the [Window] button, and change your X settings for a min of -5 and a max value of 5. (Important note: When you enter negative numbers, do not use the subtraction button! Instead use the [(-)] button.

```

Plot1 Plot2 Plot3
Y1=normalpdf(X,
0,1)
Y2=
Y3=
Y4=
Y5=
Y6=

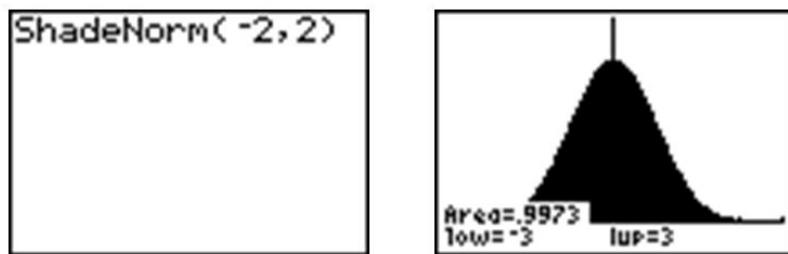
```



Now press [2ND][DISTR], go to the **DRAW** menu, and choose 'ShadeNorm('. Insert '−1, 1' after the 'ShadeNorm(' command and press [ENTER]. It will shade the area within one standard deviation of the mean.



The calculator also gives a very accurate estimate of the area. We can see from the rightmost screenshot above that approximately 68% of the area is within one standard deviation of the mean. If we venture to 2 standard deviations away from the mean, how much of the data should we expect to capture? Make the following changes to the 'ShadeNorm(' command to find out:



Notice from the shading that almost all of the distribution is shaded, and the percentage of data is close to 95%. If you were to venture to 3 standard deviations from the mean, 99.7%, or virtually all of the data, is captured, which tells us that very little of the data in a normal distribution is more than 3 standard deviations from the mean.

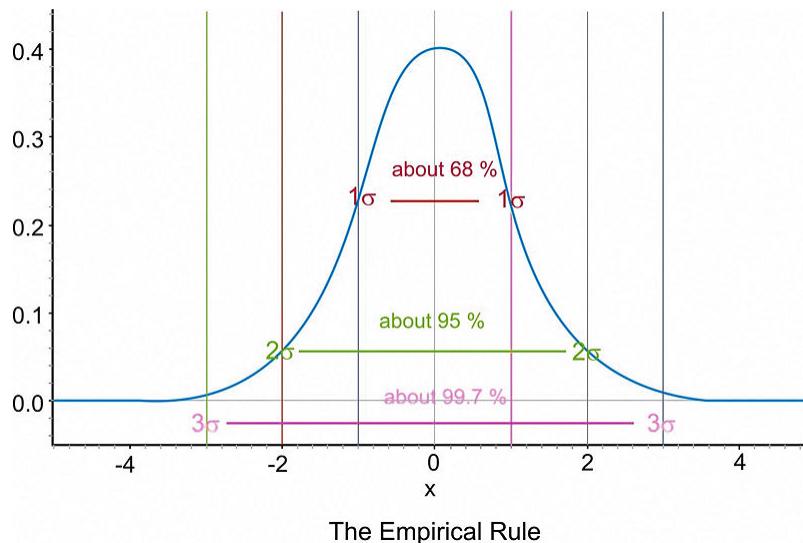


Notice that the calculator actually makes it look like the entire distribution is shaded because of the limitations of the screen resolution, but as we have already discovered, there is still some area under the curve further out than that. These three approximate percentages, 68%, 95%, and 99.7%, are extremely important and are part of what is called the *Empirical Rule*.

The Empirical Rule states that the percentages of data in a normal distribution within 1, 2, and 3 standard deviations of the mean are approximately 68%, 95%, and 99.7%, respectively.

On the Web

<http://tinyurl.com/2ue78u> Explore the Empirical Rule.



z-Scores

A *z-score* is a measure of the number of standard deviations a particular data point is away from the mean. For example, let's say the mean score on a test for your statistics class was an 82, with a standard deviation of 7 points. If your score was an 89, it is exactly one standard deviation to the right of the mean; therefore, your *z-score* would be 1. If, on the other hand, you scored a 75, your score would be exactly one standard deviation below the mean, and your *z-score* would be -1 . All values that are below the mean have negative *z*-scores, while all values that are above the mean have positive *z*-scores. A *z-score* of -2 would represent a value that is exactly 2 standard deviations below the mean, so in this case, the value would be $82 - 14 = 68$.

To calculate a *z-score* for which the numbers are not so obvious, you take the deviation and divide it by the standard deviation.

$$z = \frac{\text{Deviation}}{\text{Standard Deviation}}$$

You may recall that deviation is the mean value of the variable subtracted from the observed value, so in symbolic terms, the *z*-score would be:

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

As previously stated, since σ is always positive, z will be positive when x is greater than μ and negative when x is less than μ . A *z-score* of zero means that the term has the same value as the mean. The value of z represents the number of standard deviations the given value of x is above or below the mean.

Example: What is the *z-score* for an *A* on the test described above, which has a mean score of 82? (Assume that an *A* is a 93.)

The *z*-score can be calculated as follows:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ z &= \frac{93 - 82}{7} \\ z &= \frac{11}{7} \approx 1.57 \end{aligned}$$

If we know that the test scores from the last example are distributed normally, then a z -score can tell us something about how our test score relates to the rest of the class. From the Empirical Rule, we know that about 68% of the students would have scored between a z -score of -1 and 1 , or between a 75 and an 89 , on the test. If 68% of the data is between these two values, then that leaves the remaining 32% in the tail areas. Because of symmetry, half of this, or 16%, would be in each individual tail.

Example: On a nationwide math test, the mean was 65 and the standard deviation was 10 . If Robert scored 81 , what was his z -score?

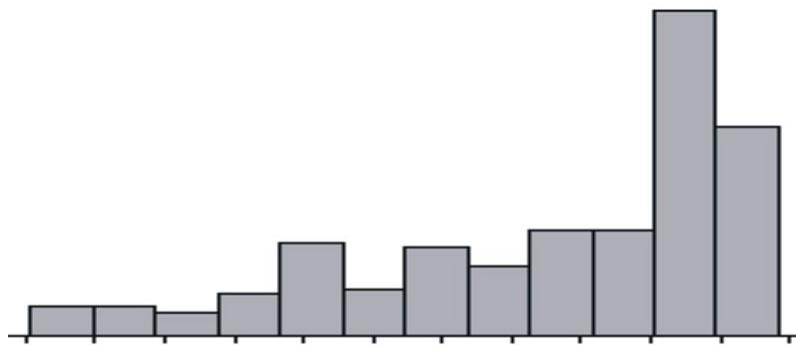
$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ z &= \frac{81 - 65}{10} \\ z &= \frac{16}{10} \\ z &= 1.6 \end{aligned}$$

Example: On a college entrance exam, the mean was 70 , and the standard deviation was 8 . If Helen's z -score was -1.5 , what was her exam score?

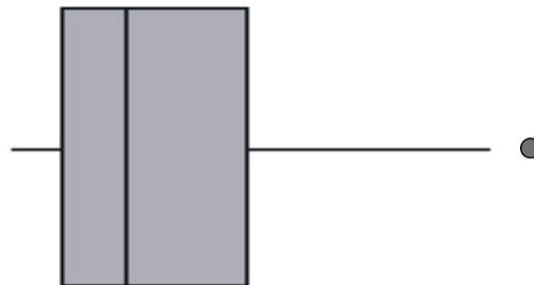
$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ \therefore z \cdot \sigma &= x - \mu \\ x &= \mu + z \cdot \sigma \\ x &= 70 + (-1.5)(8) \\ x &= 58 \end{aligned}$$

Assessing Normality

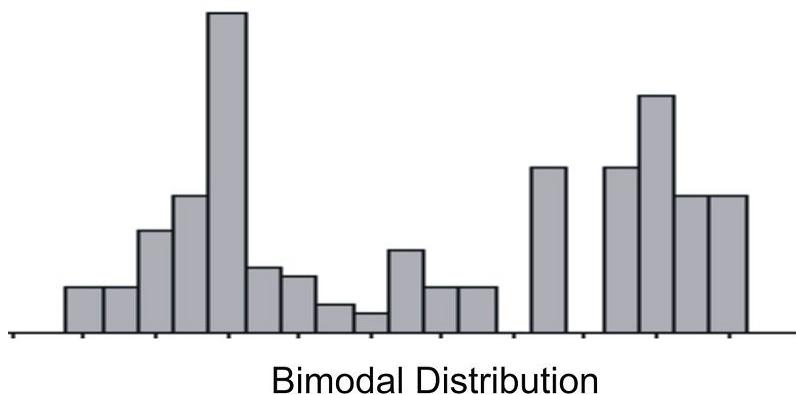
The best way to determine if a data set approximates a normal distribution is to look at a visual representation. Histograms and box plots can be useful indicators of normality, but they are not always definitive. It is often easier to tell if a data set is *not* normal from these plots.



Skewed left distribution



Skewed right distribution with outliers

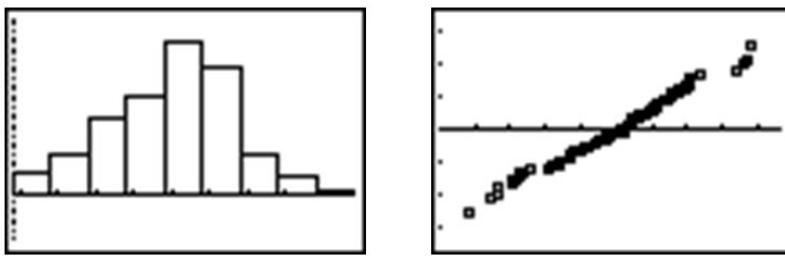


Bimodal Distribution

If a data set is skewed right, it means that the right tail is significantly longer than the left. Similarly, skewed left means the left tail has more weight than the right. A bimodal distribution, on the other hand, has two modes, or peaks. For instance, with a histogram of the heights of American 30-year-old adults, you will see a bimodal distribution—one mode for males and one mode for females.

Now that we know how to calculate z -scores, there is a plot we can use to determine if a distribution is normal. If we calculate the z -scores for a data set and plot them against the actual values, we have what is called a *normal probability plot*, or a *normal quantile plot*. If the data set is normal, then this plot will be perfectly linear. The closer to being linear the normal probability plot is, the more closely the data set approximates a normal distribution.

Look below at the histogram and the normal probability plot for the same data.



The histogram is fairly symmetric and mound-shaped and appears to display the characteristics of a normal distribution. When the z -scores are plotted against the data values, the normal probability plot appears strongly linear, indicating that the data set closely approximates a normal distribution.

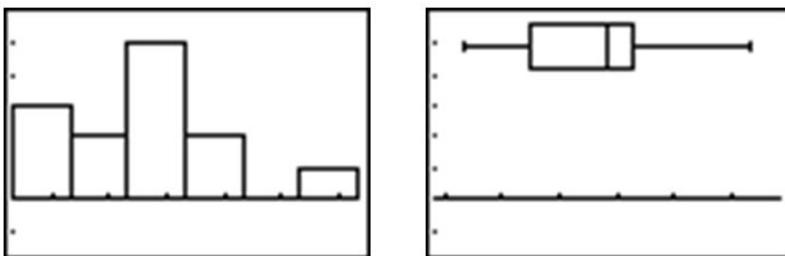
Example: The following data set tracked high school seniors' involvement in traffic accidents. The participants were asked the following question: "During the last 12 months, how many accidents have you had while you were driving (whether or not you were responsible)?"

TABLE 5.1:

Year	Percentage of high school seniors who said they were involved in no traffic accidents
1991	75.7
1992	76.9
1993	76.1
1994	75.7
1995	75.3
1996	74.1
1997	74.4
1998	74.4
1999	75.1
2000	75.1
2001	75.5
2002	75.5
2003	75.8

Figure: Percentage of high school seniors who said they were involved in no traffic accidents. *Source:* Sourcebook of Criminal Justice Statistics: <http://www.albany.edu/sourcebook/pdf/t352.pdf>

Here is a histogram and a box plot of this data:



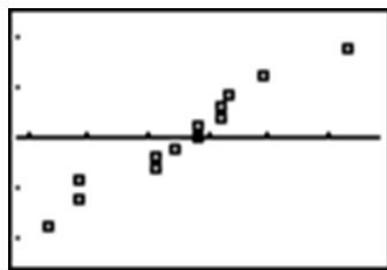
The histogram appears to show a roughly mound-shaped and symmetric distribution. The box plot does not appear to be significantly skewed, but the various sections of the plot also do not appear to be overly symmetric, either. In the following chart, the z -scores for this data set have been calculated. The mean percentage is approximately 75.35%.

TABLE 5.2:

Year	Percentage	<i>z</i>-score
1991	75.7	0.45
1992	76.9	2.03
1993	76.1	0.98
1994	75.7	0.45
1995	75.3	-0.07
1996	74.1	-1.65
1997	74.4	-1.25
1998	74.4	-1.25
1999	75.1	-0.33
2000	75.1	-0.33
2001	75.5	0.19
2002	75.5	0.19
2003	75.8	0.59

Figure: Table of *z*-scores for senior no-accident data.

Here is a plot of the percentages and the *z*-scores, or the normal probability plot:



While not perfectly linear, this plot does have a strong linear pattern, and we would, therefore, conclude that the distribution is reasonably normal.

One additional clue about the data's normality might be gained from investigating the Empirical Rule. Remember that in an idealized normal curve, approximately 68% of the data should be within one standard deviation of the mean. If we count, there are 9 years for which the *z*-scores are between -1 and 1. As a percentage of the total data, $\frac{9}{13}$ is about 69%, or very close to the value indicated by the Empirical Rule. This data set is so small that it is difficult to verify that the other percentages adhere to the Empirical Rule, but they are still not unreasonable. About 92% of the data (all but one of the points) is within 2 standard deviations of the mean, and all of the data (which is in-line with the theoretical 99.7%) is located between *z*-scores of -3 and 3.

Lesson Summary

A **normal distribution** is a perfectly symmetric, mound-shaped distribution that appears in many practical and real data sets. It is an especially important foundation for making conclusions, or inferences, about data. A standard normal distribution is a normal distribution for which the mean is 0 and the standard deviation is 1.

A ***z*-score** is a measure of the number of standard deviations a particular data value is away from the mean. The formula for calculating a *z*-score is:

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

z-scores are useful for comparing two distributions with different centers and/or spreads. When you convert an entire distribution to *z*-scores, you are actually changing it to a standardized distribution. *z*-scores can be calculated for data, even if the underlying population does not follow a normal distribution.

The **Empirical Rule** is the name given to the observation that approximately 68% of a data set is within 1 standard deviation of the mean, about 95% is within 2 standard deviations of the mean, and about 99.7% is within 3 standard deviations of the mean. Some refer to this as the 68-95-99.7 Rule.

You should learn to recognize the normality of a distribution by examining the shape and symmetry of its visual display. A normal probability plot, or normal quantile plot, is a useful tool to help check the normality of a distribution. This graph is a plot of the *z*-scores of a data set against the actual values. If a distribution is normal, this plot will be linear.

Review Questions

1. Which of the following data sets is most likely to be normally distributed? For the other choices, explain why you believe they would not follow a normal distribution.
 - a. The hand span (measured from the tip of the thumb to the tip of the extended 5th finger) of a random sample of high school seniors
 - b. The annual salaries of all employees of a large shipping company
 - c. The annual salaries of a random sample of 50 CEOs of major companies, 25 women and 25 men
 - d. The dates of 100 pennies taken from a cash drawer in a convenience store
2. The grades on a statistics mid-term for a high school are normally distributed, with $\mu = 81$ and $\sigma = 6.3$. Calculate the *z*-scores for each of the following exam grades. Draw and label a sketch for each example. 65, 83, 93, 100
3. Assume that the mean weight of 1-year-old girls in the USA is normally distributed, with a mean of about 9.5 kilograms and a standard deviation of approximately 1.1 kilograms. Without using a calculator, estimate the percentage of 1-year-old girls who meet the following conditions. Draw a sketch and shade the proper region for each problem.
 - a. Less than 8.4 kg
 - b. Between 7.3 kg and 11.7 kg
 - c. More than 12.8 kg
4. For a standard normal distribution, place the following in order from smallest to largest.
 - a. The percentage of data below 1
 - b. The percentage of data below -1
 - c. The mean
 - d. The standard deviation
 - e. The percentage of data above 2
5. The 2007 AP Statistics examination scores were not normally distributed, with $\mu = 2.8$ and $\sigma = 1.34$. What is the approximate *z*-score that corresponds to an exam score of 5? (The scores range from 1 to 5.)
 - a. 0.786
 - b. 1.46
 - c. 1.64
 - d. 2.20
 - e. A *z*-score cannot be calculated because the distribution is not normal.
6. The heights of 5th grade boys in the USA is approximately normally distributed, with a mean height of 143.5 cm and a standard deviation of about 7.1 cm. What is the probability that a randomly chosen 5th grade boy would be taller than 157.7 cm?

¹Data available on the College Board Website: <http://professionals.collegeboard.com/data-reports-research/ap/archived/2007>

5.2 The Density Curve of the Normal Distribution

Learning Objectives

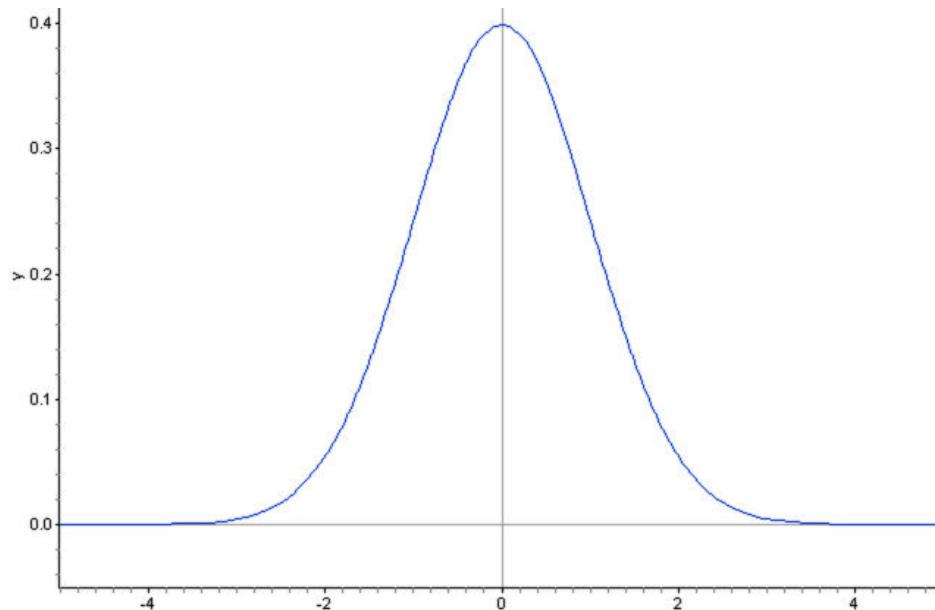
- Identify the properties of a normal density curve and the relationship between concavity and standard deviation.
- Convert between z -scores and areas under a normal probability curve.
- Calculate probabilities that correspond to left, right, and middle areas from a z -score table.
- Calculate probabilities that correspond to left, right, and middle areas using a graphing calculator.

Introduction

In this section, we will continue our investigation of normal distributions to include density curves and learn various methods for calculating probabilities from the normal density curve.

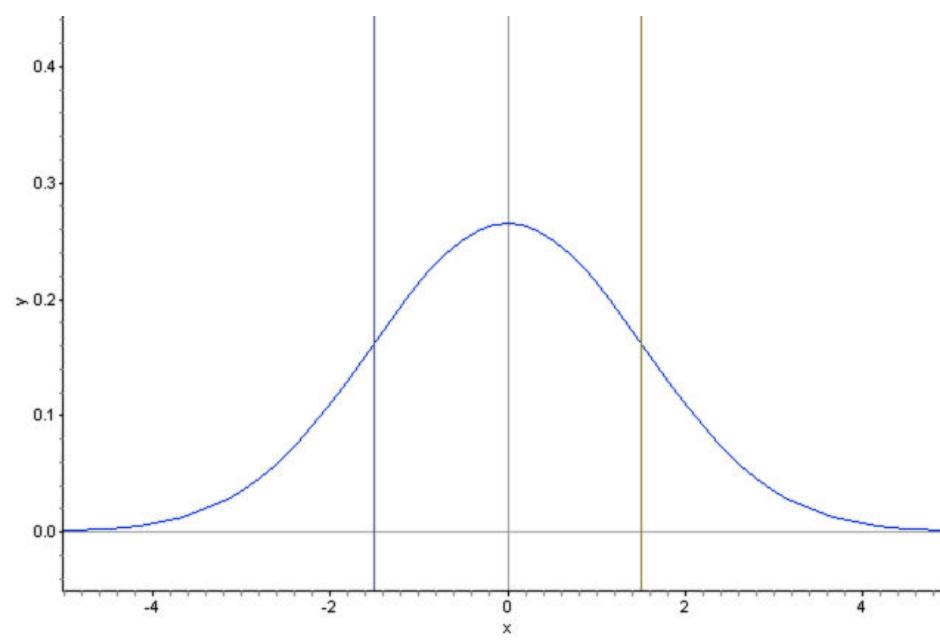
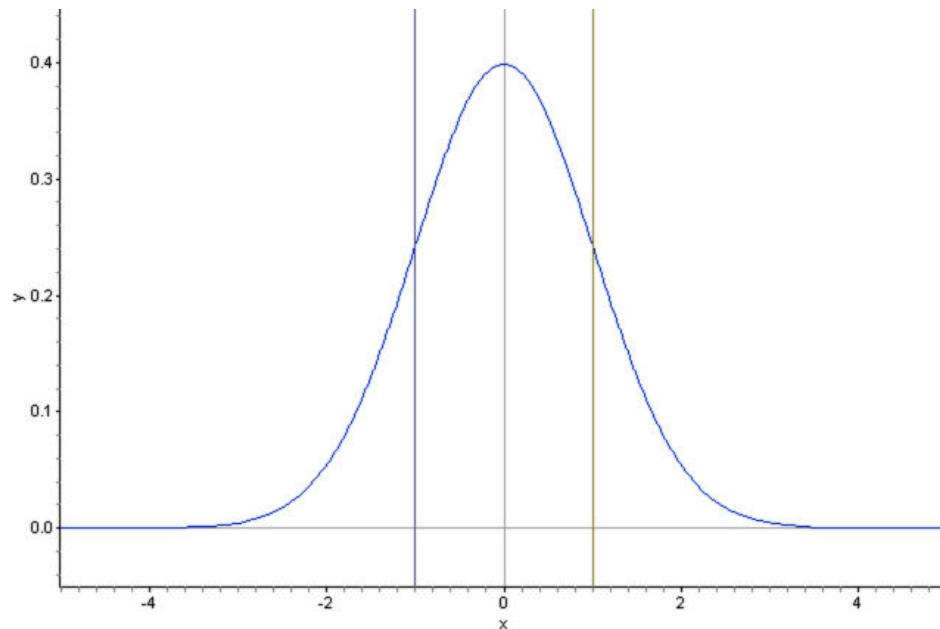
Density Curves

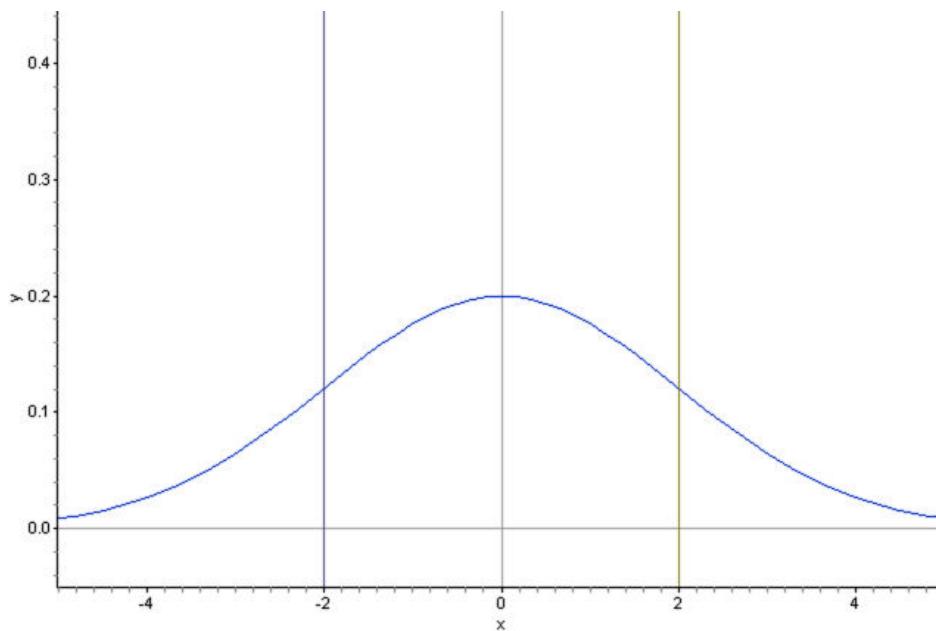
A *density curve* is an idealized representation of a distribution in which the area under the curve is defined to be 1. Density curves need not be normal, but the normal density curve will be the most useful to us.



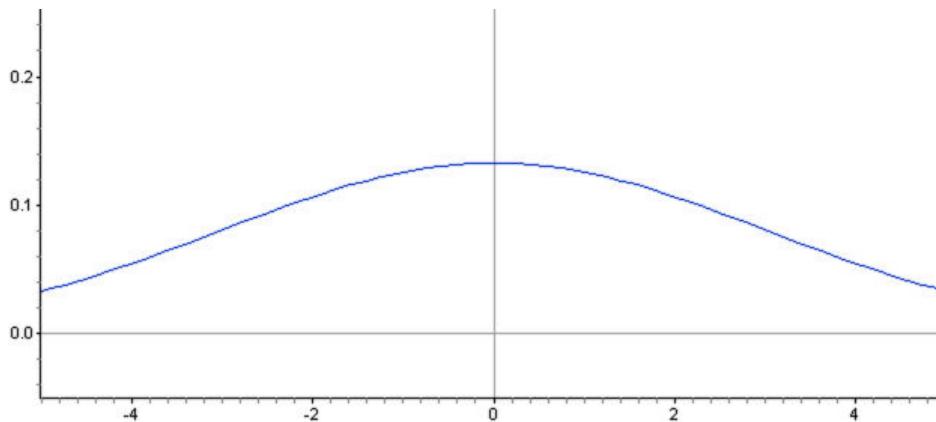
Inflection Points on a Normal Density Curve

We already know from the Empirical Rule that approximately $\frac{2}{3}$ of the data in a normal distribution lies within 1 standard deviation of the mean. With a normal density curve, this means that about 68% of the total area under the curve is within z -scores of ± 1 . Look at the following three density curves:



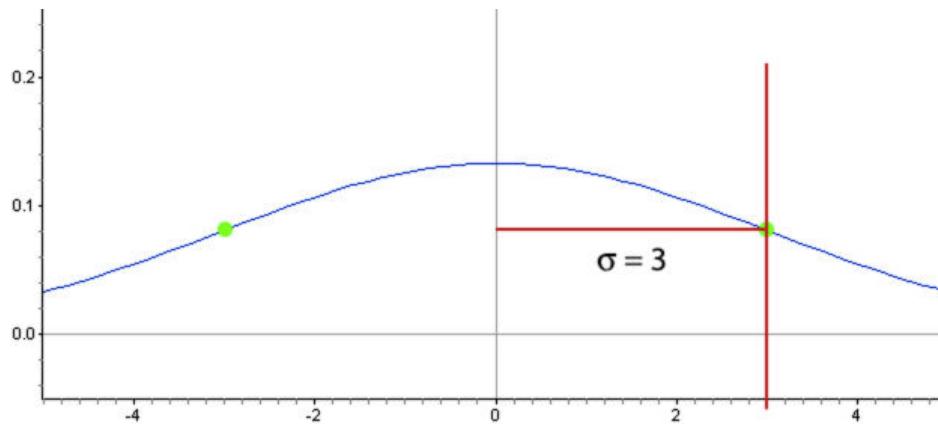


Notice that the curves are spread increasingly wider. Lines have been drawn to show the points that are one standard deviation on either side of the mean. Look at where this happens on each density curve. Here is a normal distribution with an even larger standard deviation.



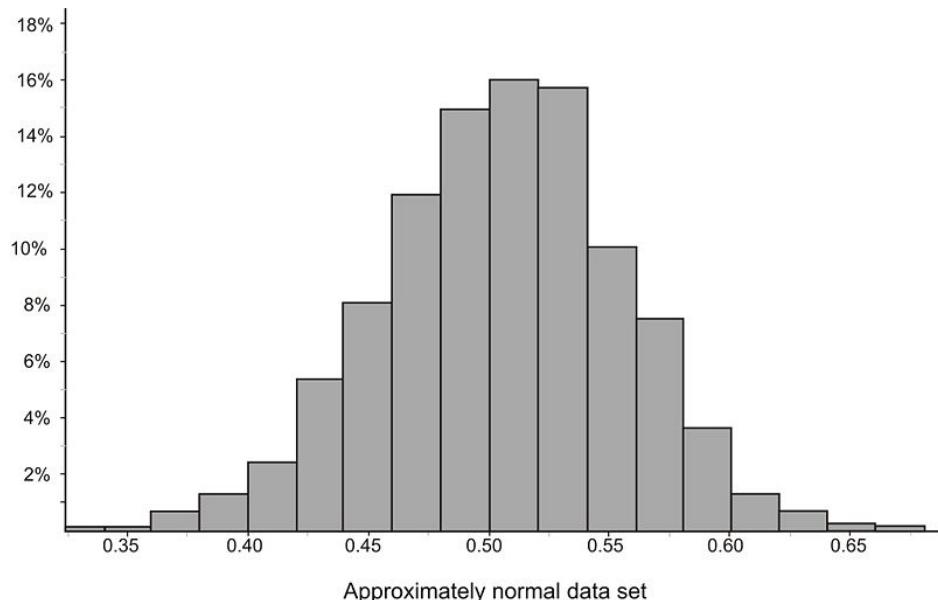
Is it possible to predict the standard deviation of this distribution by estimating the x -coordinate of a point on the density curve? Read on to find out!

You may have noticed that the density curve changes shape at two points in each of our examples. These are the points where the curve changes concavity. Starting from the mean and heading outward to the left and right, the curve is *concave down*. (It looks like a mountain, or 'n' shape.) After passing these points, the curve is *concave up*. (It looks like a valley, or 'u' shape.) The points at which the curve changes from being concave up to being concave down are called the *inflection points*. On a normal density curve, these inflection points are always exactly one standard deviation away from the mean.

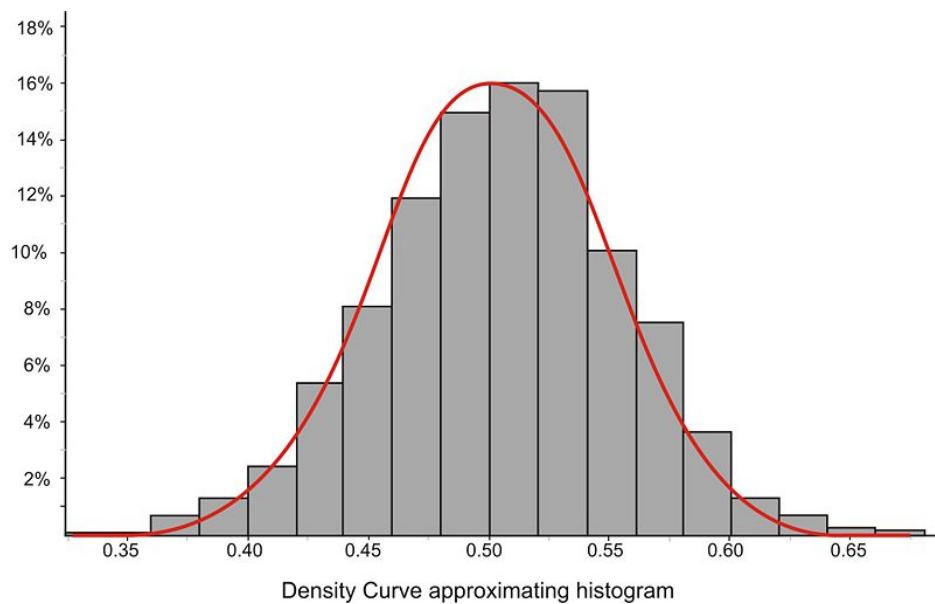


In this example, the standard deviation is 3 units. We can use this concept to estimate the standard deviation of a normally distributed data set.

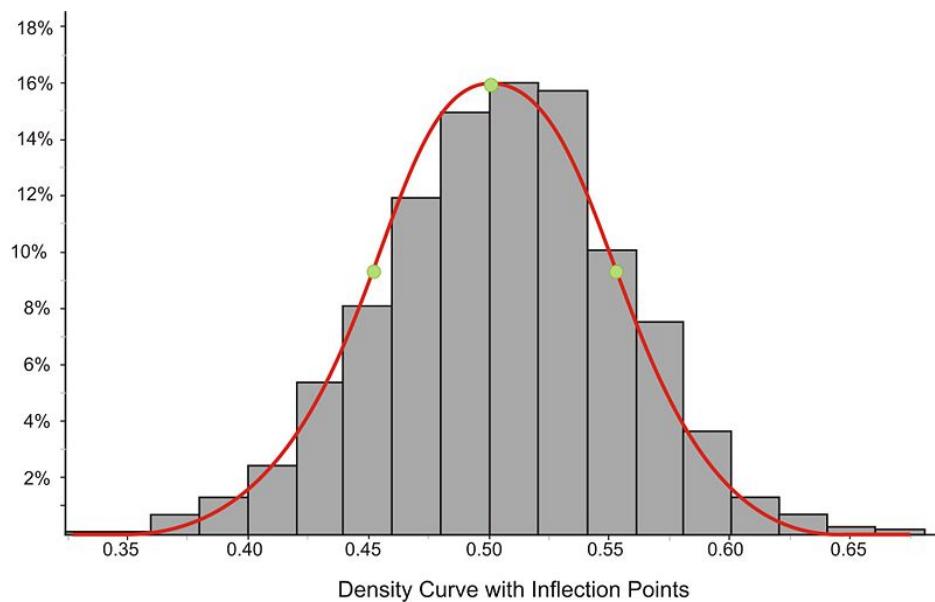
Example: Estimate the standard deviation of the distribution represented by the following histogram.



This distribution is fairly normal, so we could draw a density curve to approximate it as follows:



Now estimate the inflection points as shown below:



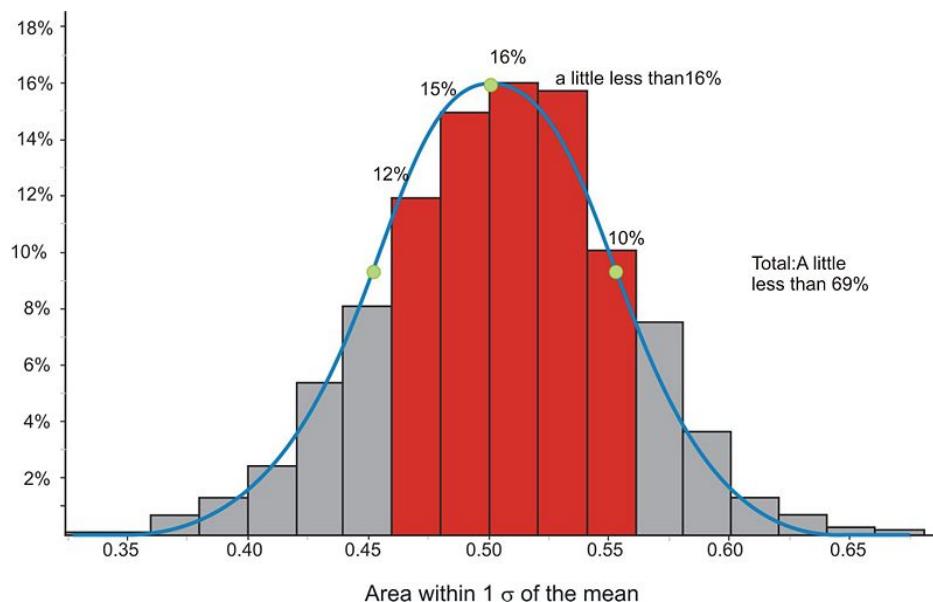
It appears that the mean is about 0.5 and that the x -coordinates of the inflection points are about 0.45 and 0.55, respectively. This would lead to an estimate of about 0.05 for the standard deviation.

The actual statistics for this distribution are as follows:

$$s \approx 0.04988$$

$$\bar{x} \approx 0.04997$$

We can verify these figures by using the expectations from the Empirical Rule. In the following graph, we have highlighted the bins that are contained within one standard deviation of the mean.



If you estimate the relative frequencies from each bin, their total is remarkably close to 68%. Make sure to divide the relative frequencies from the bins on the ends by 2 when performing your calculation.

Calculating Density Curve Areas

While it is convenient to estimate areas under a normal curve using the Empirical Rule, we often need more precise methods to calculate these areas. Luckily, we can use formulas or technology to help us with the calculations.

z-Tables

All normal distributions have the same basic shape, and therefore, rescaling and re-centering can be implemented to change any normal distributions to one with a mean of 0 and a standard deviation of 1. This configuration is referred to as a *standard normal distribution*. In a standard normal distribution, the variable along the horizontal axis is the *z-score*. This score is another measure of the performance of an individual score in a population. To review, the *z*-score measures how many standard deviations a score is away from the mean. The *z*-score of the term x in a population distribution whose mean is μ and whose standard deviation is σ is given by: $z = \frac{x-\mu}{\sigma}$. Since σ is always positive, z will be positive when x is greater than μ and negative when x is less than μ . A *z*-score of 0 means that the term has the same value as the mean. The value of z is the number of standard deviations the given value of x is above or below the mean.

Example: On a nationwide math test, the mean was 65 and the standard deviation was 10. If Robert scored 81, what was his *z*-score?

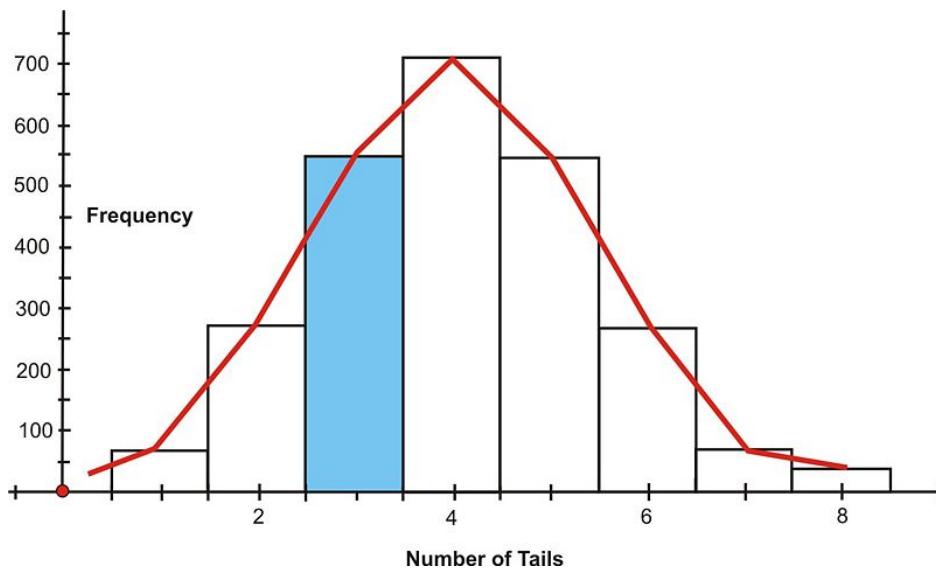
$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ z &= \frac{81-65}{10} \\ z &= \frac{16}{10} \\ z &= 1.6 \end{aligned}$$

Example: On a college entrance exam, the mean was 70 and the standard deviation was 8. If Helen's *z*-score was -1.5 , what was her exam score?

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ \therefore z \cdot \sigma &= x - \mu \\ x &= \mu + z \cdot \sigma \\ x &= (70) + (-1.5)(8) \\ x &= 58 \end{aligned}$$

Now you will see how z -scores are used to determine the probability of an event.

Suppose you were to toss 8 coins 256 times. The following figure shows the histogram and the approximating normal curve for the experiment. The random variable represents the number of tails obtained.



The blue section of the graph represents the probability that exactly 3 of the coins turned up tails. One way to determine this is by the following:

$$\begin{aligned} P(3 \text{ tails}) &= \frac{8C_3}{2^8} \\ P(3 \text{ tails}) &= \frac{56}{256} \\ P(3 \text{ tails}) &\cong 0.2188 \end{aligned}$$

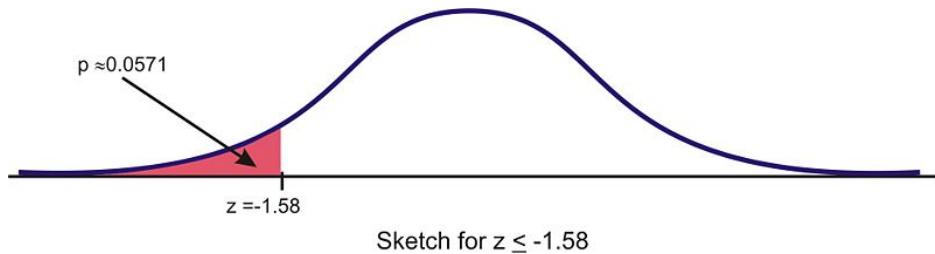
Geometrically, this probability represents the area of the blue shaded bar divided by the total area of the bars. The area of the blue shaded bar is approximately equal to the area under the normal curve from 2.5 to 3.5.

Since areas under normal curves correspond to the probability of an event occurring, a special normal distribution table is used to calculate the probabilities. This table can be found in any statistics book, but it is seldom used today. The following is an example of a table of z -scores and a brief explanation of how it works: <http://tinyurl.com/2ce9ogv>.

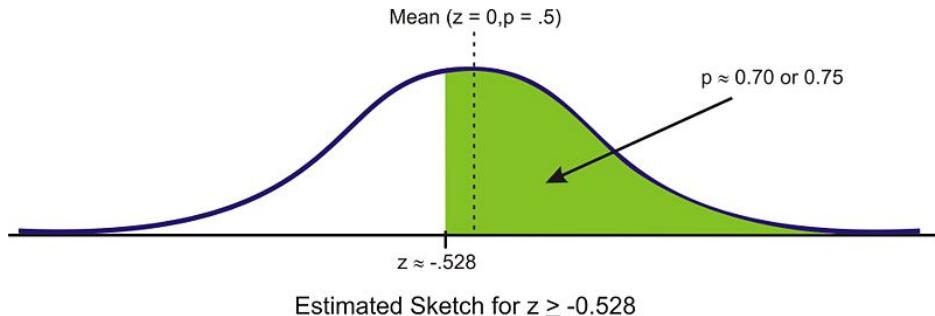
The values inside the given table represent the areas under the standard normal curve for values between 0 and the relative z -score. For example, to determine the area under the curve between z -scores of 0 and 2.36, look in the intersecting cell for the row labeled 2.3 and the column labeled 0.06. The area under the curve is 0.4909. To determine the area between 0 and a negative value, look in the intersecting cell of the row and column which sums

to the absolute value of the number in question. For example, the area under the curve between -1.3 and 0 is equal to the area under the curve between 1.3 and 0 , so look at the cell that is the intersection of the 1.3 row and the 0.00 column. (The area is 0.4032 .)

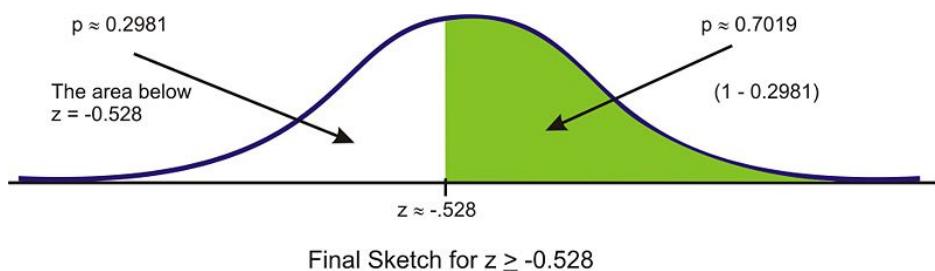
It is extremely important, especially when you first start with these calculations, that you get in the habit of relating it to the normal distribution by drawing a sketch of the situation. In this case, simply draw a sketch of a standard normal curve with the appropriate region shaded and labeled.



Example: Find the probability of choosing a value that is greater than $z = -0.528$. Before even using the table, first draw a sketch and estimate the probability. This z -score is just below the mean, so the answer should be more than 0.5 .



Next, read the table to find the correct probability for the data below this z -score. We must first round this z -score to -0.53 , so this will slightly under-estimate the probability, but it is the best we can do using the table. The table returns a value of $0.5 - 0.2019 = 0.2981$ as the area below this z -score. Because the area under the density curve is equal to 1 , we can subtract this value from 1 to find the correct probability of about 0.7019 .



What about values between two z -scores? While it is an interesting and worthwhile exercise to do this using a table, it is so much simpler using software or a graphing calculator.

Example: Find $P(-2.60 < z < 1.30)$

This probability can be calculated as follows:

$$P(-2.60 < z < 1.30) = P(z < 1.30) - P(z < -2.60) = 0.9032 - 0.0047 = 0.8985$$

It can also be found using the TI-83/84 calculator. Use the 'normalcdf(-2.60, 1.30, 0, 1)' command, and the calculator will return the result 0.898538. The syntax for this command is 'normalcdf(min, max, μ , σ)'. When using this command, you do not need to first standardize. You can use the mean and standard deviation of the given distribution.

Technology Note: The 'normalcdf()' Command on the TI-83/84 Calculator

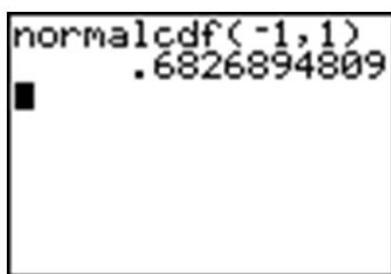
Your graphing calculator has already been programmed to calculate probabilities for a normal density curve using what is called a *cumulative density function*. The command you will use is found in the **DISTR** menu, which you can bring up by pressing [2ND][DISTR].



Press [2] to select the 'normalcdf()' command, which has a syntax of 'normalcdf(lower bound, upper bound, mean, standard deviation)'.

The command has been programmed so that if you do not specify a mean and standard deviation, it will default to the standard normal curve, with $\mu = 0$ and $\sigma = 1$.

For example, entering 'normalcdf(-1, 1)' will specify the area within one standard deviation of the mean, which we already know to be approximately 0.68.



Try verifying the other values from the Empirical Rule.

Summary:

'Normalpdf(x, μ, σ)' gives values of the *probability density function*. This is the function we graphed in Lesson 5.1. If μ and σ are not specified, it is assumed that $\mu = 0$ and $\sigma = 1$.

'Normalcdf (a, b, μ, σ)' gives values of the cumulative normal density function. In other words, it gives the probability of an event occurring between $x = a$ and $x = b$, or the area under the probability density curve between the vertical lines $x = a$ and $x = b$, where the normal distribution has a mean of μ and a standard deviation of σ . If μ and σ are not specified, it is assumed that $\mu = 0$ and $\sigma = 1$.

```
normalcdf( -2,2)
.954499876
normalcdf( -3,3)
.9973000656
```

Example: Find the probability that $x < -1.58$.

The calculator command must have both an upper and lower bound. Technically, though, the density curve does not have a lower bound, as it continues infinitely in both directions. We do know, however, that a very small percentage of the data is below 3 standard deviations to the left of the mean. Use -3 as the lower bound and see what answer you get.

```
normalcdf( -3, -1.
58)
.0557034698
```

The answer is fairly accurate, but you must remember that there is really still some area under the probability density curve, even though it is just a little, that we are leaving out if we stop at -3 . If you look at the z -table, you can see that we are, in fact, leaving out about $0.5 - 0.4987 = 0.0013$. Next, try going out to -4 and -5 .

```
normalcdf( -4, -1.
58)
.057021751
normalcdf( -5, -1.
58)
.0570531499
```

Once we get to -5 , the answer is quite accurate. Since we cannot really capture all the data, entering a sufficiently small value should be enough for any reasonable degree of accuracy. A quick and easy way to handle this is to enter -99999 (or “a bunch of nines”). It really doesn’t matter exactly how many nines you enter. The difference between five and six nines will be beyond the accuracy that even your calculator can display.

```
normalcdf( -99999
, -1.58)
.057053437
normalcdf( -99999
9, -1.58)
.057053437
```

Example: Find the probability for $x \geq -0.528$.

Right away, we are at an advantage using the calculator, because we do not have to round off the z -score. Enter the 'normalcdf(' command, using -0.528 to "a bunch of nines." The nines represent a ridiculously large upper bound that will insure that the unaccounted-for probability will be so small that it will be virtually undetectable.

```
normalcdf( -.528,
9999999)
.7012503533
```

Remember that because of rounding, our answer from the table was slightly too small, so when we subtracted it from 1, our final answer was slightly too large. The calculator answer of about 0.70125 is a more accurate approximation than the answer arrived at by using the table.

Standardizing

In most practical problems involving normal distributions, the curve will not be as we have seen so far, with $\mu = 0$ and $\sigma = 1$. When using a z -table, you will first have to *standardize* the distribution by calculating the z -score(s).

Example: A candy company sells small bags of candy and attempts to keep the number of pieces in each bag the same, though small differences due to random variation in the packaging process lead to different amounts in individual packages. A quality control expert from the company has determined that the mean number of pieces in each bag is normally distributed, with a mean of 57.3 and a standard deviation of 1.2. Endy opened a bag of candy and felt he was cheated. His bag contained only 55 candies. Does Endy have reason to complain?

To determine if Endy was cheated, first calculate the z -score for 55:

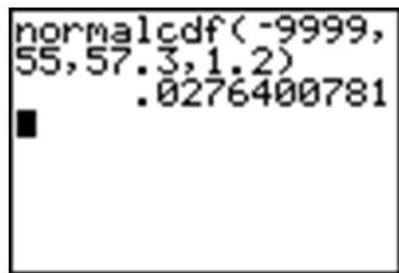
$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ z &= \frac{55 - 57.3}{1.2} \\ z &\approx -1.911666\dots \end{aligned}$$

Using a table, the probability of experiencing a value this low is approximately $0.5 - 0.4719 = 0.0281$. In other words, there is about a 3% chance that you would get a bag of candy with 55 or fewer pieces, so Endy should feel cheated.

Using a graphing calculator, the results would look as follows (the 'Ans' function has been used to avoid rounding off the z -score):

```
55-57.3
Ans/1.2
normalcdf( -99999
99,Ans)
.0276400781
```

However, one of the advantages of using a calculator is that it is unnecessary to standardize. We can simply enter the mean and standard deviation from the original population distribution of candy, avoiding the z -score calculation completely.

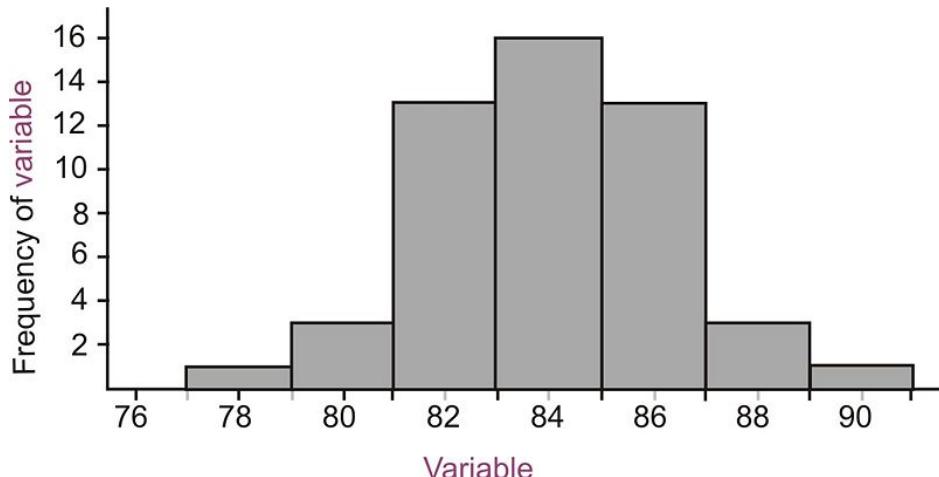


Lesson Summary

A density curve is an idealized representation of a distribution in which the area under the curve is defined as 1, or in terms of percentages, a probability of 100%. A normal density curve is simply a density curve for a normal distribution. Normal density curves have two inflection points, which are the points on the curve where it changes concavity. These points correspond to the points in the normal distribution that are exactly 1 standard deviation away from the mean. Applying the Empirical Rule tells us that the area under the normal density curve between these two points is approximately 0.68. This is most commonly thought of in terms of probability (e.g., the probability of choosing a value at random from this distribution and having it be within 1 standard deviation of the mean is 0.68). Calculating other areas under the curve can be done by using a z -table or by using the 'normalcdf' command on the TI-83/84 calculator. A z -table often provides the area under the standard normal density curve between the mean and a particular z -score. The calculator command allows you to specify two values, either standardized or not, and will calculate the area under the curve between these values.

Review Questions

1. Estimate the standard deviation of the following distribution.



2. A z -table most commonly gives the probabilities below given z -scores, or what are sometimes referred to as left-tail probabilities. Probabilities above certain z -scores are complementary to those below, so all we have to do is subtract the table value from 1 to find the probability above a certain z -score. To calculate the probabilities between two z -scores, calculate the left tail probabilities for both z -scores and subtract the left-most value from the right. Try these using the table only!
 - a. $P(z \geq -0.79)$

- b. $P(-1 \leq z \leq 1)$ Show all work.
c. $P(-1.56 < z < 0.32)$
3. Brielle's statistics class took a quiz, and the results were normally distributed, with a mean of 85 and a standard deviation of 7. She wanted to calculate the percentage of the class that got a *B* (between 80 and 90). She used her calculator and was puzzled by the result. Here is a screen shot of her calculator:



Explain her mistake and the resulting answer on the calculator, and then calculate the correct answer.

4. Which grade is better: A 78 on a test whose mean is 72 and standard deviation is 6.5, or an 83 on a test whose mean is 77 and standard deviation is 8.4. Justify your answer and draw sketches of each distribution.

Answers: (1) about 2 (2)(a) .7852 (2)(b) .6826 (2)(c) .5661 (3) $\text{normalcdf}(80,90,85,7) = .5249$ (4) the score of 78 is better

5.3 Applications of the Normal Distribution

Learning Objective

- Apply the characteristics of a normal distribution to solving problems.

Introduction

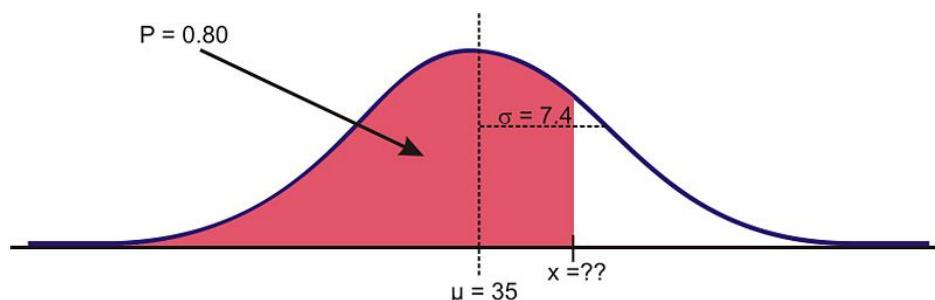
The normal distribution is the foundation for statistical inference and will be an essential part of many of those topics in later chapters. In the meantime, this section will cover some of the types of questions that can be answered using the properties of a normal distribution. The first examples deal with more theoretical questions that will help you master basic understandings and computational skills, while the later problems will provide examples with real data, or at least a real context.

Unknown Value Problems

If you understand the relationship between the area under a density curve and mean, standard deviation, and z -scores, you should be able to solve problems in which you are provided all but one of these values and are asked to calculate the remaining value. In the last lesson, we found the probability that a variable is within a particular range, or the area under a density curve within that range. What if you are asked to find a value that gives a particular probability?

Example: Given the normally-distributed random variable X , with $\mu = 35$ and $\sigma = 7.4$, what is the value of X where the probability of experiencing a value less than it is 80%?

As suggested before, it is important and helpful to sketch the distribution.

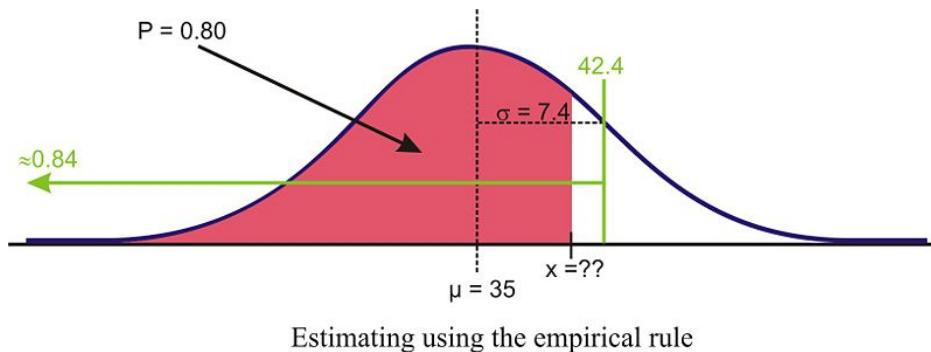


Sketch of distribution

If we had to estimate an actual value first, we know from the Empirical Rule that about 84% of the data is below one standard deviation to the right of the mean.

$$\mu + 1\sigma = 35 + 7.4 = 42.4$$

Therefore, we expect the answer to be slightly below this value.



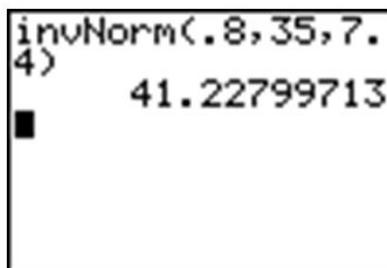
When we were given a value of the variable and were asked to find the percentage or probability, we used a *z*-table or the 'normalcdf' command on a graphing calculator. But how do we find a value given the percentage? Again, the table has its limitations in this case, and graphing calculators and computer software are much more convenient and accurate. The command on the TI-83/84 calculator is 'invNorm'. You may have seen it already in the **DISTR** menu.



The syntax for this command is as follows:

'InvNorm(percentage or probability to the left, mean, standard deviation)'

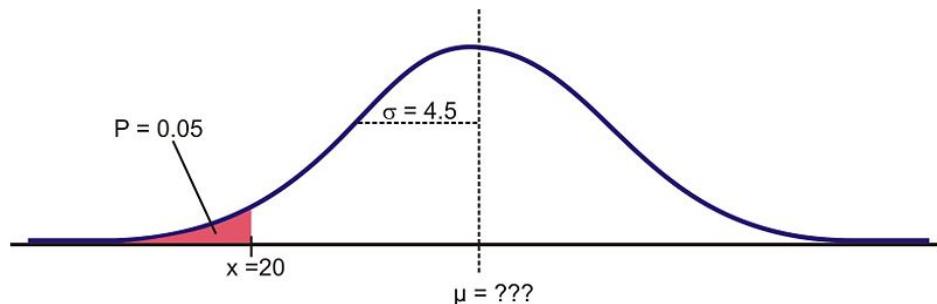
Make sure to enter the values in the correct order, such as in the example below:



Unknown Mean or Standard Deviation

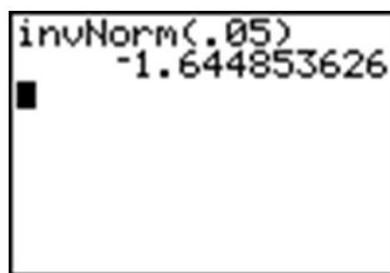
Example: For a normally distributed random variable, $\sigma = 4.5$, $x = 20$, and $p = 0.05$, Estimate μ .

To solve this problem, first draw a sketch:



Remember that about 95% of the data is within 2 standard deviations of the mean. This would leave 2.5% of the data in the lower tail, so our 5% value must be less than 9 units from the mean.

Because we do not know the mean, we have to use the standard normal curve and calculate a z -score using the 'invNorm(' command. The result, -1.645 , confirms the prediction that the value is less than 2 standard deviations from the mean.

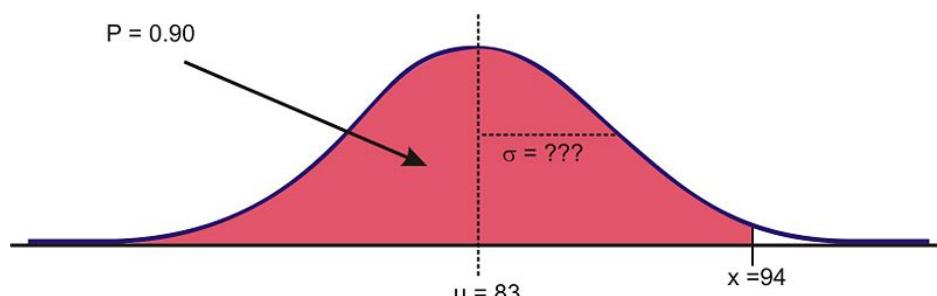


Now, plug in the known quantities into the z -score formula and solve for μ as follows:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ -1.645 &\approx \frac{20 - \mu}{4.5} \\ (-1.645)(4.5) &\approx 20 - \mu \\ -7.402 - 20 &\approx -\mu \\ -27.402 &\approx -\mu \\ \mu &\approx 27.402 \end{aligned}$$

Example: For a normally-distributed random variable, $\mu = 83$, $x = 94$, and $p = 0.90$. Find σ .

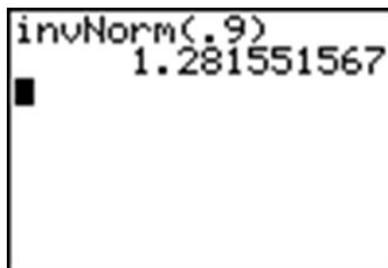
Again, let's first look at a sketch of the distribution.



Sketch of Distribution

Since about 97.5% of the data is below 2 standard deviations, it seems reasonable to estimate that the x value is less than two standard deviations away from the mean and that σ might be around 7 or 8.

Again, the first step to see if our prediction is right is to use 'invNorm(' to calculate the z -score. Remember that since we are not entering a mean or standard deviation, the result is based on the assumption that $\mu = 0$ and $\sigma = 1$.

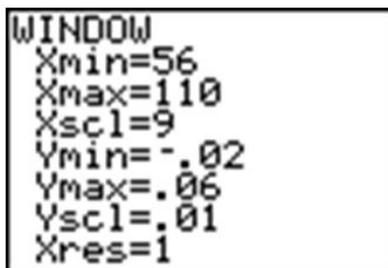


Now, use the z -score formula and solve for σ as follows:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ 1.282 &\approx \frac{94 - 83}{\sigma} \\ \sigma &\approx \frac{11}{1.282} \\ \sigma &\approx 8.583 \end{aligned}$$

Technology Note: Drawing a Distribution on the TI-83/84 Calculator

The TI-83/84 calculator will draw a distribution for you, but before doing so, we need to set an appropriate window (see screen below) and delete or turn off any functions or plots. Let's use the last example and draw the shaded region below 94 under a normal curve with $\mu = 83$ and $\sigma = 8.583$. Remember from the Empirical Rule that we probably want to show about 3 standard deviations away from 83 in either direction. If we use 9 as an estimate for σ , then we should open our window 27 units above and below 83. The y settings can be a bit tricky, but with a little practice, you will get used to determining the maximum percentage of area near the mean.



The reason that we went below the x -axis is to leave room for the text, as you will see.

Now, press [2ND][DISTR] and arrow over to the **DRAW** menu.

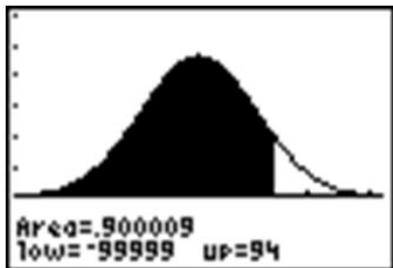
Choose the 'ShadeNorm(' command. With this command, you enter the values just as if you were doing a 'normalcdf(' calculation. The syntax for the 'ShadeNorm(' command is as follows:

'ShadeNorm(lower bound, upper bound, mean, standard deviation)'

Enter the values shown in the following screenshot:

```
ShadeNorm( -99999
,94,83,8.583)
```

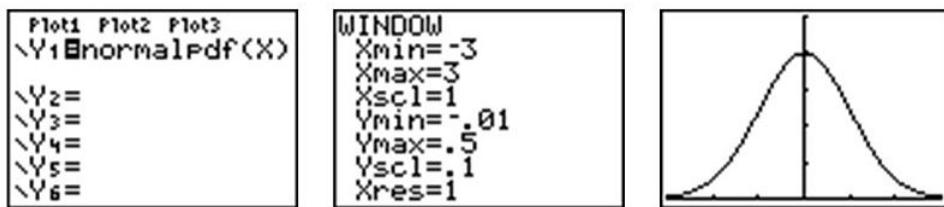
Next, press [ENTER] to see the result. It should appear as follows:



Technology Note: The 'normalpdf(' Command on the TI-83/84 Calculator

You may have noticed that the first option in the **DISTR** menu is 'normalpdf()', which stands for a normal probability density function. It is the option you used in lesson 5.1 to draw the graph of a normal distribution. Many students wonder what this function is for and occasionally even use it by mistake to calculate what they think are cumulative probabilities, but this function is actually the mathematical formula for drawing a normal distribution. You can find this formula in the resources at the end of the lesson if you are interested. The numbers this function returns are not really useful to us statistically. The primary purpose for this function is to draw the normal curve.

To do this, first be sure to turn off any plots and clear out any functions. Then press [**Y=**], insert 'normalpdf()', enter '**X**', and close the parentheses as shown. Because we did not specify a mean and standard deviation, the standard normal curve will be drawn. Finally, enter the following window settings, which are necessary to fit most of the curve on the screen (think about the Empirical Rule when deciding on settings), and press [**GRAPH**]. The normal curve below should appear on your screen.



Normal Distributions with Real Data

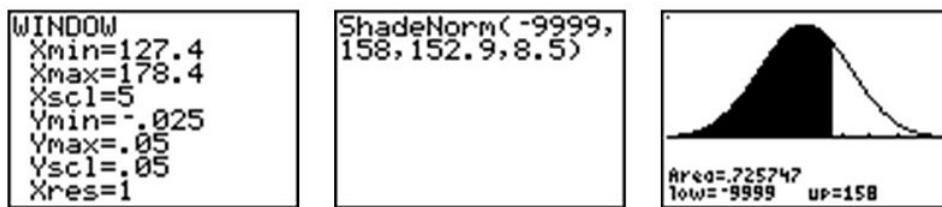
The foundation of performing experiments by collecting surveys and samples is most often based on the normal distribution, as you will learn in greater detail in later chapters. Here are two examples to get you started.

Example: The Information Centre of the National Health Service in Britain collects and publishes a great deal of information and statistics on health issues affecting the population. One such comprehensive data set tracks information about the health of children¹. According to its statistics, in 2006, the mean height of 12-year-old boys was 152.9 cm, with a standard deviation estimate of approximately 8.5 cm. (These are not the exact figures for the

population, and in later chapters, we will learn how they are calculated and how accurate they may be, but for now, we will assume that they are a reasonable estimate of the true parameters.)

If 12-year-old Cecil is 158 cm, approximately what percentage of all 12-year-old boys in Britain is he taller than?

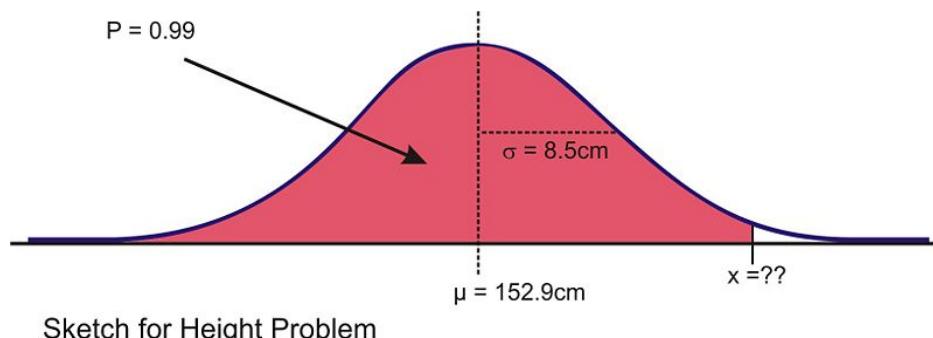
We first must assume that the height of 12-year-old boys in Britain is normally distributed, and this seems like a reasonable assumption to make. As always, draw a sketch and estimate a reasonable answer prior to calculating the percentage. In this case, let's use the calculator to sketch the distribution and the shading. First decide on an appropriate window that includes about 3 standard deviations on either side of the mean. In this case, 3 standard deviations is about 25.5 cm, so add and subtract this value to/from the mean to find the horizontal extremes. Then enter the appropriate 'ShadeNorm(' command as shown:



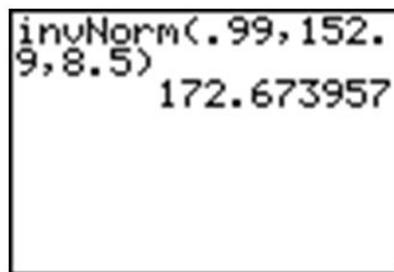
From this data, we would estimate that Cecil is taller than about 73% of 12-year-old boys. We could also phrase our assumption this way: the probability of a randomly selected British 12-year-old boy being shorter than Cecil is about 0.73. Often with data like this, we use percentiles. We would say that Cecil is in the 73rd percentile for height among 12-year-old boys in Britain.

How tall would Cecil need to be in order to be in the top 1% of all 12-year-old boys in Britain?

Here is a sketch:



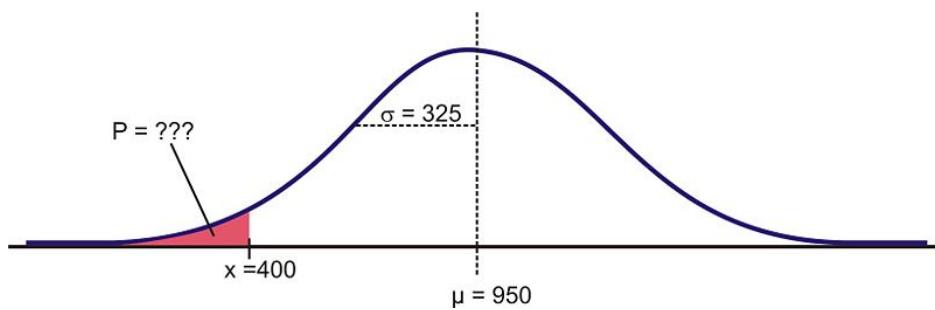
In this case, we are given the percentage, so we need to use the 'invNorm(' command as shown.



Our results indicate that Cecil would need to be about 173 cm tall to be in the top 1% of 12-year-old boys in Britain.

Example: Suppose that the distribution of the masses of female marine iguanas in Puerto Villamil in the Galapagos Islands is approximately normal, with a mean mass of 950 g and a standard deviation of 325 g. There are very few

young marine iguanas in the populated areas of the islands, because feral cats tend to kill them. How rare is it that we would find a female marine iguana with a mass less than 400 g in this area?



Using a graphing calculator, we can approximate the probability of a female marine iguana being less than 400 grams as follows:

```
normalcdf( -9999,
400,950,325)
.045293632
```

With a probability of approximately 0.045, or only about 5%, we could say it is rather unlikely that we would find an iguana this small.

Lesson Summary

In order to find the percentage of data in-between two values (or the probability of a randomly chosen value being between those values) in a normal distribution, we can use the 'normalcdf(' command on the TI-83/84 calculator. When you know the percentage or probability, use the 'invNorm(' command to find a z -score or value of the variable. In order to use these tools in real situations, we need to know that the distribution of the variable in question is approximately normal. When solving problems using normal probabilities, it helps to draw a sketch of the distribution and shade the appropriate region.

Review Questions

- Which of the following intervals contains the middle 95% of the data in a standard normal distribution?
 - $z < 2$
 - $z \leq 1.645$
 - $z \leq 1.96$
 - $-1.645 \leq z \leq 1.645$
 - $-1.96 \leq z \leq 1.96$
- For each of the following problems, X is a continuous random variable with a normal distribution and the given mean and standard deviation. P is the probability of a value of the distribution being less than x . Find

the missing value and sketch and shade the distribution.

mean	Standard deviation	x	P
85	4.5		0.68
mean	Standard deviation	x	P
	1	16	0.05
mean	Standard deviation	x	P
73		85	0.91
mean	Standard deviation	x	P
93	5		0.90

3. What is the z -score for the lower quartile in a standard normal distribution?
4. Suppose that the wrapper of a certain candy bar lists its weight as 2.13 ounces. Naturally, the weights of individual bars vary somewhat. Suppose that the weights of these candy bars vary according to a normal distribution, with $\mu = 2.2$ ounces and $\sigma = 0.04$ ounces.
 - a. What proportion of the candy bars weigh less than the advertised weight?
 - b. What proportion of the candy bars weight between 2.2 and 2.3 ounces?
 - c. A candy bar of what weight would be heavier than all but 1% of the candy bars out there?

Answers: (1) E (2)(a) $x = 87.12$ (b) $\mu = 17.64$ (c) $\sigma = 8.96$ (d) $x = 99.4$ (3) $z = -0.67$ (4)(a) .0401 (4)(b) .4938 (4)(c) 2 ounces

Keywords

Concave down

Starting from the mean and heading outward to the left and right, the curve is *concave down*.

Concave up

After passing these points, the curve is *concave up*.

Cumulative density function

to calculate probabilities for a normal density curve using what is called a *cumulative density function*.

Density curve

A curve where the area under the curve equals exactly one.

Empirical Rule

States what percentages of data in a normal distribution lies within 1, 2, and 3 standard deviations of the mean.

Inflection Points

A point where the curve changes concavity (from concave up to concave down, or concave down to concave up).

Normal distribution

A continuous probability distribution that has a symmetric bell-shaped curve with a single peak.

Normal probability plot

A normal probability plot can also be used to determine normality.

Normal quantile plot

If we calculate the z -scores for a data set and plot them against the actual values, we have what is called a *normal probability plot*, or a *normal quantile plot*. If the data set is normal, then this plot will be perfectly linear.

Probability density function

DISTR menu is 'normalpdf()', which stands for a normal probability density function.

Standard normal curve

We have to use the standard normal curve and calculate a z -score using the 'invNorm()' command.

Standard normal distribution

A normal distribution with $\mu = 0$ and $\sigma = 1$.

Standardize

the curve will not be as we have seen so far, with $\mu = 0$ and $\sigma = 1$. When using a z -table, you will first have to *standardize* the distribution by calculating the z -score(s).

z -score

A measure of the number of standard deviations a particular data point is away from the mean.

CHAPTER

6

Planning and Conducting an Experiment or Study

Chapter Outline

6.1 SURVEYS AND SAMPLING

6.2 EXPERIMENTAL DESIGN

6.1 Surveys and Sampling

Learning Objectives

- Differentiate between a census and a survey or sample.
- Distinguish between sampling error and bias.
- Identify and name potential sources of bias from both real and hypothetical sampling situations.
- Identify and name 4 types of sampling methodologies.

Introduction

The New York Times/CBS News Poll is a well-known regular polling organization that releases results of polls taken to help clarify the opinions of Americans on pending elections, current leaders, or economic or foreign policy issues. In an article entitled “How the Poll Was Conducted” that explains some of the details of a recent poll, the following statements appear¹:

“In theory, in 19 cases out of 20, overall results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking to interview all American adults.”

“In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll. Variation in the wording and order of questions, for example, may lead to somewhat different results.”

These statements illustrate two different potential problems with opinion polls, surveys, observational studies, and experiments. In this chapter, we will investigate these problems and more by looking at sampling in detail.

Census vs. Sample

A *sample* is a representative subset of a population. If a statistician or other researcher wants to know some information about a population, the only way to be truly sure is to conduct a census. In a *census*, every unit in the population being studied is measured or surveyed. In opinion polls, like the *New York Times* poll mentioned above, results are generalized from a sample. If we really wanted to know the true approval rating of the president, for example, we would have to ask every single American adult his or her opinion. There are some obvious reasons why a census is impractical in this case, and in most situations.

First, it would be extremely expensive for the polling organization. They would need an extremely large workforce to try and collect the opinions of every American adult. Also, it would take many workers and many hours to organize, interpret, and display this information. Even if it could be done in several months, by the time the results were published, it would be very probable that recent events had changed peoples’ opinions and that the results would be obsolete.

In addition, a census has the potential to be destructive to the population being studied. For example, many manufacturing companies test their products for quality control. A padlock manufacturer might use a machine to see how much force it can apply to the lock before it breaks. If they did this with every lock, they would have none left to sell! Likewise, it would not be a good idea for a biologist to find the number of fish in a lake by draining the lake and counting them all!

The U.S. Census is probably the largest and longest running census, since the Constitution mandates a complete counting of the population. The first U.S. Census was taken in 1790 and was done by U.S. Marshalls on horseback.

Taken every 10 years, a Census was conducted in 2010, and in a report by the Government Accountability Office in 1994, was estimated to cost \$11 billion. This cost has recently increased as computer problems have forced the forms to be completed by hand³. You can find a great deal of information about the U.S. Census, as well as data from past Censuses, on the Census Bureau's website: <http://www.census.gov/>.

Due to all of the difficulties associated with a census, sampling is much more practical. However, it is important to understand that even the most carefully planned sample will be subject to random variation between the sample and the population. Recall that these differences due to chance are called *sampling error*. We can use the laws of probability to predict the level of accuracy in our sample. Opinion polls, like the *New York Times* poll mentioned in the introduction, tend to refer to this as *margin of error*. The second statement quoted from the *New York Times* article mentions another problem with sampling. That is, it is often difficult to obtain a sample that accurately reflects the total population. It is also possible to make mistakes in selecting the sample and collecting the information. These problems result in a non-representative sample, or one in which our conclusions differ from what they would have been if we had been able to conduct a census.

To help understand these ideas, consider the following theoretical example. A coin is considered fair if the probability, p , of the coin landing on heads is the same as the probability of it landing on tails ($p = 0.5$). The probability is defined as the proportion of heads obtained if the coin were flipped an infinite number of times. Since it is impractical, if not impossible, to flip a coin an infinite number of times, we might try looking at 10 samples, with each sample consisting of 10 flips of the coin. Theoretically, you would expect the coin to land on heads 50% of the time, but it is very possible that, due to chance alone, we would experience results that differ from this. These differences are due to sampling error. As we will investigate in detail in later chapters, we can decrease the sampling error by increasing the sample size (or the number of coin flips in this case). It is also possible that the results we obtain could differ from those expected if we were not careful about the way we flipped the coin or allowed it to land on different surfaces. This would be an example of a non-representative sample.

At the following website, you can see the results of a large number of coin flips: <http://www.mathsonline.co.uk/nonmembers/resource/prob/coins.html>. You can see the random variation among samples by asking for the site to flip 10 coins 10 times. Our results for that experiment produced the following numbers of heads: 3, 3, 4, 4, 4, 4, 5, 6, 6, 6. This seems quite strange, since the expected number is 5. How do your results compare?



Bias in Samples and Surveys

The term most frequently applied to a non-representative sample is *bias*. Bias has many potential sources. It is important when selecting a sample or designing a survey that a statistician make every effort to eliminate potential sources of bias. In this section, we will discuss some of the most common types of bias. While these concepts are universal, the terms used to define them here may be different than those used in other sources.

Sampling Bias

In general, sampling bias refers to the methods used in selecting the sample. The *sampling frame* is the term we use to refer to the group or listing from which the sample is to be chosen. If you wanted to study the population of

students in your school, you could obtain a list of all the students from the office and choose students from the list. This list would be the sampling frame.

Incorrect Sampling Frame

If the list from which you choose your sample does not accurately reflect the characteristics of the population, this is called *incorrect sampling frame*. A sampling frame error occurs when some group from the population does not have the opportunity to be represented in the sample. For example, surveys are often done over the telephone. You could use the telephone book as a sampling frame by choosing numbers from the telephone book. However, in addition to the many other potential problems with telephone polls, some phone numbers are not listed in the telephone book. Also, if your population includes all adults, it is possible that you are leaving out important groups of that population. For example, many younger adults in particular tend to only use their cell phones or computer-based phone services and may not even have traditional phone service. Even if you picked phone numbers randomly, the sampling frame could be incorrect, because there are also people, especially those who may be economically disadvantaged, who have no phone. There is absolutely no chance for these individuals to be represented in your sample. A term often used to describe the problems when a group of the population is not represented in a survey is *undercoverage*. Undercoverage can result from all of the different sampling biases.

One of the most famous examples of sampling frame error occurred during the 1936 U.S. presidential election. The Literary Digest, a popular magazine at the time, conducted a poll and predicted that Alf Landon would win the election that, as it turned out, was won in a landslide by Franklin Delano Roosevelt. The magazine obtained a huge sample of ten million people, and from that pool, 2 million replied. With these numbers, you would typically expect very accurate results. However, the magazine used their subscription list as their sampling frame. During the depression, these individuals would have been only the wealthiest Americans, who tended to vote Republican, and left the majority of typical voters under-covered.

Convenience Sampling

Suppose your statistics teacher gave you an assignment to perform a survey of 20 individuals. You would most likely tend to ask your friends and family to participate, because it would be easy and quick. This is an example of *convenience sampling*, or convenience bias. While it is not always true, your friends are usually people who share common values, interests, and opinions. This could cause those opinions to be over-represented in relation to the true population. Also, have you ever been approached by someone conducting a survey on the street or in a mall? If such a person were just to ask the first 20 people they found, there is the potential that large groups representing various opinions would not be included, resulting in undercoverage.

Judgment Sampling

Judgment sampling occurs when an individual or organization that is usually considered an expert in the field being studied chooses the individuals or group of individuals to be used in the sample. Because it is based on a subjective choice, even by someone considered an expert, it is very susceptible to bias. In some sense, this is what those responsible for the Literary Digest poll did. They incorrectly chose groups they believed would represent the population. If a person wants to do a survey on middle-class Americans, how would this person decide who to include? It would be left to this person's own judgment to create the criteria for those considered middle-class. This individual's judgment might result in a different view of the middle class that might include wealthier individuals that others would not consider part of the population. Similar to judgment sampling, in *quota sampling*, an individual or organization attempts to include the proper proportions of individuals of different subgroups in their sample. While it might sound like a good idea, it is subject to an individual's prejudice and is, therefore, prone to bias.

Size Bias

If one particular subgroup in a population is likely to be over-represented or under-represented due to its size, this is sometimes called *size bias*. If we chose a state at random from a map by closing our eyes and pointing to a particular place, larger states would have a greater chance of being chosen than smaller ones. As another example, suppose that we wanted to do a survey to find out the typical size of a student's math class at a school. The chances are greater that we would choose someone from a larger class for our survey. To understand this, say that you went to a very small school where there are only four math classes, with one class having 35 students, and the other three classes having only 8 students. If you simply choose students at random, it is more likely you will select students for your sample who will say the typical size of a math class is 35, since there are more students in the larger class.

Here's one more example: a person driving on an interstate highway tends to say things like, "Wow, I was going the speed limit, and everyone was just flying by me." The conclusion this person is making about the population of all drivers on this highway is that most of them are traveling faster than the speed limit. This may indeed be true, but let's say that most people on the highway, along with our driver, really are abiding by the speed limit. In a sense, the driver is collecting a sample, and only those few who are close to our driver will be included in the sample. There will be a larger number of drivers going faster in our sample, so they will be over-represented. As you may already see, these definitions are not absolute, and often in a practical example, there are many types of overlapping bias that could be present and contribute to overcoverage or undercoverage. We could also cite incorrect sampling frame or convenience bias as potential problems in this example.

Response Bias

The term *response bias* refers to problems that result from the ways in which the survey or poll is actually presented to the individuals in the sample.

Voluntary Response Bias

Television and radio stations often ask viewers/listeners to call in with opinions about a particular issue they are covering. The websites for these and other organizations also usually include some sort of online poll question of the day. Reality television shows and fan balloting in professional sports to choose all-star players make use of these types of polls as well. All of these polls usually come with a disclaimer stating that, "This is not a scientific poll." While perhaps entertaining, these types of polls are very susceptible to *voluntary response bias*. The people who respond to these types of surveys tend to feel very strongly one way or another about the issue in question, and the results might not reflect the overall population. Those who still have an opinion, but may not feel quite so passionately about the issue, may not be motivated to respond to the poll. This is especially true for phone-in or mail-in surveys in which there is a cost to participate. The effort or cost required tends to weed out much of the population in favor of those who hold extremely polarized views. A news channel might show a report about a child killed in a drive-by shooting and then ask for people to call in and answer a question about tougher criminal sentencing laws. They would most likely receive responses from people who were very moved by the emotional nature of the story and wanted anything to be done to improve the situation. An even bigger problem is present in those types of polls in which there is no control over how many times an individual may respond.

Non-Response Bias

One of the biggest problems in polling is that most people just don't want to be bothered taking the time to respond to a poll of any kind. They hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly past an interviewer on the street. We just don't know how much these individuals' beliefs and opinions reflect those of the general population, and, therefore, almost all surveys could be prone to *non-response bias*.

Questionnaire Bias

Questionnaire bias occurs when the way in which the question is asked influences the response given by the individual. It is possible to ask the same question in two different ways that would lead individuals with the same basic opinions to respond differently. Consider the following two questions about gun control.

"Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?"

"Do you believe that it is reasonable for the government to infringe on an individual's constitutional right to bear arms?"

A gun rights activist might feel very strongly that the government should never be in the position of limiting guns in any way and would answer no to both questions. Someone who is very strongly against gun ownership would similarly answer no to both questions. However, individuals with a more tempered, middle position on the issue might believe in an individual's right to own a gun under some circumstances, while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the wording of a question to obtain a certain response to a poll question. Questionnaire bias is not necessarily always a deliberate action. If a question is poorly worded, confusing, or just plain hard to understand, it could lead to non-representative results. When you ask people to choose between two options, it is even possible that the order in which you list the choices may influence their response!

Incorrect Response Bias

A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individual intentionally responds to a survey with an untruthful answer, this is called *incorrect response bias*. This can occur when asking questions about extremely sensitive or personal issues. For example, a survey conducted about illegal drinking among teens might be prone to this type of bias. Even if guaranteed their responses are confidential, some teenagers may not want to admit to engaging in such behavior at all. Others may want to appear more rebellious than they really are, but in either case, we cannot be sure of the truthfulness of the responses.

Another example is related to the donation of blood. Because the dangers of donated blood being tainted with diseases carrying a negative social stereotype increased in the 1990's, the Red Cross has recently had to deal with incorrect response bias on a constant and especially urgent basis. Individuals who have engaged in behavior that puts them at risk for contracting AIDS or other diseases have the potential to pass these diseases on through donated blood⁴. Screening for at-risk behaviors involves asking many personal questions that some find awkward or insulting and may result in knowingly false answers. The Red Cross has gone to great lengths to devise a system with several opportunities for individuals giving blood to anonymously report the potential danger of their donation.

In using this example, we don't want to give the impression that the blood supply is unsafe. According to the Red Cross, "Like most medical procedures, blood transfusions have associated risk. In the more than fifteen years since March 1985, when the FDA first licensed a test to detect HIV antibodies in donated blood, the Centers for Disease Control and Prevention has reported only 41 cases of AIDS caused by transfusion of blood that tested negative for the AIDS virus. During this time, more than 216 million blood components were transfused in the United States. The tests to detect HIV were designed specifically to screen blood donors. These tests have been regularly upgraded since they were introduced. Although the tests to detect HIV and other blood-borne diseases are extremely accurate, they cannot detect the presence of the virus in the 'window period' of infection, the time before detectable antibodies or antigens are produced. That is why there is still a very slim chance of contracting HIV from blood that tests negative. Research continues to further reduce the very small risk." ⁴ Source:<http://chapters.redcross.org/br/nypennregion/safety/mythsaid.htm>

Reducing Bias by Using Appropriate Sampling Techniques

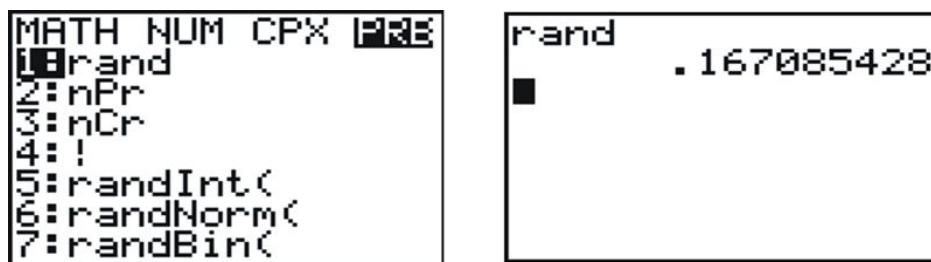
Simple Random Sample (SRS)

The best technique for reducing bias in sampling is *randomization*. When a **simple random sample** of size n (commonly referred to as an SRS) is taken from a population, all possible samples of size n in the population have an equal probability of being selected for the sample. For example, if your statistics teacher wants to choose a student at random for a special prize, he or she could simply place the names of all the students in the class in a hat, mix them up, and choose one. More scientifically, your teacher could assign each student in the class a number from 1 to 25 (assuming there are 25 students in the class) and then use a computer or calculator to generate a random number to choose one student. This would be a simple random sample of size 1.

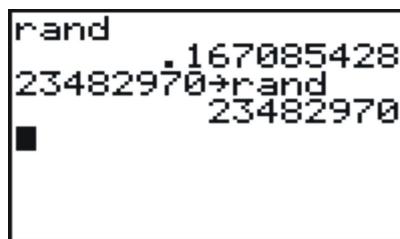
A Note about Randomness

Technology Note: Generating Random Numbers on the TI-83/84 Calculator

Your graphing calculator has a random number generator. Press [MATH] and move over to the **PRB** menu, which stands for probability. (Note: Instead of pressing the right arrow three times, you can just use the left arrow once!) Choose '1:rand' for the random number generator and press [ENTER] twice to produce a random number between 0 and 1. Press [ENTER] a few more times to see more results.



It is important that you understand that there is no such thing as true randomness, especially on a calculator or computer. When you choose the 'rand' function, the calculator has been programmed to return a ten digit decimal that, using a very complicated mathematical formula, simulates randomness. Each digit, in theory, is equally likely to occur in any of the individual decimal places. What this means in practice is that if you had the patience (and the time!) to generate a million of these on your calculator and keep track of the frequencies in a table, you would find there would be an approximately equal number of each digit. However, two brand-new calculators will give the exact same sequences of random numbers! This is because the function that simulates randomness has to start at some number, called a *seed value*. All the calculators are programmed from the factory (or when the memory is reset) to use a seed value of zero. If you want to be sure that your sequence of random digits is different from everyone else's, you need to seed your random number function using a number different from theirs. Type a unique sequence of digits on the home screen, press [STO], enter the 'rand' function, and press [ENTER]. As long as the number you chose to seed the function is different from everyone else's, you will get different results.



Now, back to our example. If we want to choose a student at random between 1 and 25, we need to generate a random integer between 1 and 25. To do this, press [MATH][PRB] and choose the 'randInt(' function.

```
MATH NUM CPX PRE
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(

```

The syntax for this command is as follows:

'RandInt(starting value, ending value, number of random integers)'

The default for the last field is 1, so if you only need a single random digit, you can enter the following:

```
randInt(1, 25) 7
```

In this example, the student chosen would be student number 7. If we wanted to choose 5 students at random, we could enter the command shown below:

```
randInt(1, 25) 7
randInt(1, 25, 5)
{17 21 10 4 10}
```

However, because the probability of any digit being chosen each time is independent from all other times, it is possible that the same student could get chosen twice, as student number 10 did in our example.

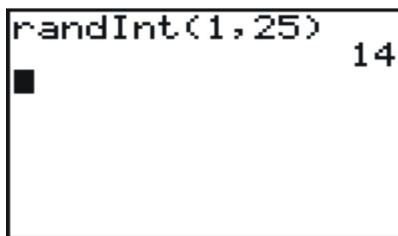
What we can do in this case is ignore any repeated digits. Since student number 10 has already been chosen, we will ignore the second 10. Press [ENTER] again to generate 5 new random numbers, and choose the first one that is not in your original set.

```
randInt(1, 25) 7
randInt(1, 25, 5)
{17 21 10 4 10}
randInt(1, 25, 5)
{4 14 15 16 13}
```

In this example, student number 4 has also already been chosen, so we would select student number 14 as our fifth student.

Systematic Sampling

There are other appropriate sampling techniques besides choosing an SRS, and one of these is a systematic sample. In **systematic sampling**, after choosing a starting point at random, subjects are selected using a jump number. If you have ever chosen teams or groups in gym class by counting off by threes or fours, you were engaged in systematic sampling. The jump number is determined by dividing the population size by the desired sample size to insure that the sample combs through the entire population. If we had a list of everyone in your class of 25 students in alphabetical order, and we wanted to choose 5 of them, we would choose every 5th student. Let's try choosing a starting point at random by generating a random number from 1 to 25 as shown below:



In this case, we would start with student number 14 and then select every 5th student until we had 5 in all. When we came to the end of the list, we would continue the count at number 1. Thus, our chosen students would be: 14, 19, 24, 4, and 9. It is important to note that this is not a simple random sample, as not every possible sample of 5 students has an equal chance of being chosen. For example, it is impossible to have a sample consisting of students 5, 6, 7, 8, and 9.

Cluster Sampling

Cluster sampling is when a naturally occurring group is selected at random, and then either all of that group, or randomly selected individuals from that group, are used for the sample. If we select at random from out of that group, or cluster into smaller subgroups, this is referred to as *multi-stage sampling*. For example, to survey student opinions or study their performance, we could choose 5 schools at random from your state and then use an SRS (simple random sample) from each school. If we wanted a national survey of urban schools, we might first choose 5 major urban areas from around the country at random, and then select 5 schools at random from each of these cities. This would be both cluster and multi-stage sampling. Cluster sampling is often done by selecting a particular block or street at random from within a town or city. It is also used at large public gatherings or rallies. If officials take a picture of a small, representative area of the crowd and count the individuals in just that area, they can use that count to estimate the total crowd in attendance.

Stratified Sampling

In **stratified sampling**, the population is divided into groups, called strata (the singular term is 'stratum'), that have some meaningful relationship. Very often, groups in a population that are similar may respond differently to a survey. In order to help reflect the population, we stratify to insure that each opinion is represented in the sample. For example, we often stratify by gender or race in order to make sure that the often divergent views of these different groups are represented. In a survey of high school students, we might choose to stratify by school to be sure that the opinions of different communities are included. If each school has an approximately equal number of students, then we could simply choose to take an SRS of size 25 from each school. If the numbers in each stratum are different, then it would be more appropriate to choose a fixed sample (100 students, for example) from each school and take a number from each school proportionate to the total school size.

Lesson Summary

If you collect information from every unit in a population, it is called a **census**. Because a census is so difficult to do, we instead take a representative subset of the population, called a **sample**, to try and make conclusions about the entire population. The downside to sampling is that we can never be completely sure that we have captured the truth about the entire population, due to random variation in our sample that is called sampling error. The list of the population from which the sample is chosen is called the **sampling frame**. Poor technique in surveying or choosing a sample can also lead to incorrect conclusions about the population that are generally referred to as bias. Selection bias refers to choosing a sample that results in a subgroup that is not representative of the population. Incorrect sampling frame occurs when the group from which you choose your sample does not include everyone in the population, or at least units that reflect the full diversity of the population. Incorrect sampling frame errors result in undercoverage. This is where a segment of the population containing an important characteristic did not have an opportunity to be chosen for the sample and will be marginalized, or even left out altogether.

We have been introduced to 4 appropriate sampling techniques commonly used by researchers: **simple random sampling, systematic sampling, cluster sampling, and stratified sampling**, all of which will reduce sampling bias.

Points to Consider

- How is the margin of error for a survey calculated?
- What are the effects of sample size on sampling error?

Review Questions

1. Brandy wanted to know which brand of soccer shoe high school soccer players prefer. She decided to ask the girls on her team which brand they liked.
 - a. What is the population in this example?
 - b. What are the units?
 - c. If she asked all high school soccer players this question, what is the statistical term we would use to describe the situation?
 - d. Which group(s) from the population is/are going to be under-represented?
 - e. What type of bias best describes the error in her sample? Why?
 - f. Brandy got a list of all the soccer players in the Colonial conference from her athletic director, Mr. Sprain. This list is called the what?
 - g. If she grouped the list by boys and girls, and chose 40 boys at random and 40 girls at random, what type of sampling best describes her method?
2. Your doorbell rings, and you open the door to find a 6-foot-tall boa constrictor wearing a trench coat and holding a pen and a clip board. He says to you, “I am conducting a survey for a local clothing store. Do you own any boots, purses, or other items made from snake skin?” After recovering from the initial shock of a talking snake being at the door, you quickly and nervously answer, “Of course not,” as the wallet you bought on vacation last summer at Reptile World weighs heavily in your pocket. What type of bias best describes this ridiculous situation? Explain why.

In each of the next two examples, identify the type of sampling that is most evident and explain why you think it applies.

3. In order to estimate the population of moose in a wilderness area, a biologist familiar with that area selects a particular marsh area and spends the month of September, during mating season, cataloging sightings of moose.

4. The local sporting goods store has a promotion where every 1000th customer gets a \$10 gift card.

For questions 5-9, an amusement park wants to know if its new ride, The Pukeinator, is too scary. Explain the type(s) of bias most evident in each sampling technique and/or what sampling method is most evident. Be sure to justify your choice.

5. The first 30 riders on a particular day are asked their opinions of the ride.
 6. The name of a color is selected at random, and only riders wearing that particular color are asked their opinion of the ride.
 7. A flier is passed out inviting interested riders to complete a survey about the ride at 5 pm that evening.
 8. Every 12th teenager exiting the ride is asked in front of his friends: “You didn’t think that ride was scary, did you?”
 9. Five riders are selected at random during each hour of the day, from 9 AM until closing at 5 PM.
-

Answers: (1)(a) all high school soccer players (1)(b) individual soccer players (1)(c) Census (1)(d) male high school soccer players (1)(e) sampling bias (1)(f) sampling frame (1)(g) stratified random sample (2) incorrect response bias (3) cluster sampling (4) systematic sampling (5) convenience sampling (6) convenience sampling (7) voluntary response (8) systematic sampling; questionnaire bias and incorrect response bias (9) stratified random sampling

References

- <http://www.nytimes.com/2008/04/04/us/04pollbox.html>
- <http://www.gao.gov/cgi-bin/getrpt?GAO-04-37>
- <http://www.cnn.com/2008/TECH/04/03/census.problems.ap/>
- http://en.wikipedia.org/wiki/Literary_Digest

6.2 Experimental Design

Learning Objectives

- Identify the important characteristics of an experiment.
- Distinguish between confounding and lurking variables.
- Use a random number generator to randomly assign experimental units to treatment groups.
- Identify experimental situations in which blocking is necessary or appropriate and create a blocking scheme for such experiments.
- Identify experimental situations in which a matched pairs design is necessary or appropriate and explain how such a design could be implemented.
- Identify the reasons for and the advantages of blind experiments.
- Distinguish between correlation and causation.

Introduction

A recent study published by the Royal Society of Britain¹ concluded that there is a relationship between the nutritional habits of mothers around the time of conception and the gender of their children. The study found that women who ate more calories and had a higher intake of essential nutrients and vitamins were more likely to conceive sons. As we learned in the first chapter, this study provides useful evidence of an association between these two variables, but it is only an observational study. It is possible that there is another variable that is actually responsible for the gender differences observed. In order to be able to convincingly conclude that there is a cause and effect relationship between a mother's diet and the gender of her child, we must perform a controlled statistical experiment. This lesson will cover the basic elements of designing a proper statistical experiment.

Confounding and Lurking Variables

In an *observational study* such as the Royal Society's connecting gender and a mother's diet, it is possible that there is a third variable that was not observed that is causing a change in both the explanatory and response variables. A variable that is not included in a study but that may still have an effect on the other variables involved is called a *lurking variable*. Perhaps the existence of this variable is unknown or its effect is not suspected.

Example: It's possible that in the study presented above, the mother's exercise habits caused both her increased consumption of calories and her increased likelihood of having a male child.

A slightly different type of additional variable is called a *confounding variable*. *Confounding variables* are those that affect the response variable and are also related to the explanatory variable. The effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable. They are both observed, but it cannot be distinguished which one is actually causing the change in the response variable.

Example: The study described above also mentions that the habit of skipping breakfast could possibly depress glucose levels and lead to a decreased chance of sustaining a viable male embryo. In an observational study, it is impossible to determine if it is nutritional habits in general, or the act of skipping breakfast, that causes a change in gender birth rates. A well-designed statistical *experiment* has the potential to isolate the effects of these intertwined variables, but there is still no guarantee that we will ever be able to determine if one of these variables, or some other factor, causes a change in gender birth rates.

Observational studies and the public's appetite for finding simplified cause-and-effect relationships between easily

observable factors are especially prone to confounding. The phrase often used by statisticians is, “Correlation (association) does not imply causation.” For example, another recent study published by the Norwegian Institute of Public Health² found that first-time mothers who had a Caesarian section were less likely to have a second child. While the trauma associated with the procedure may cause some women to be more reluctant to have a second child, there is no medical consequence of a Caesarian section that directly causes a woman to be less able to have a child. The 600,000 first-time births over a 30-year time span that were examined are so diverse and unique that there could be a number of underlying causes that might be contributing to this result.

Experiments: Treatments, Randomization, and Replication

There are three elements that are essential to any statistical experiment that can earn the title of a randomized clinical trial. The first is that a **treatment** must be imposed on the subjects of the experiment. In the example of the British study on gender, we would have to prescribe different diets to different women who were attempting to become pregnant, rather than simply observing or having them record the details of their diets during this time, as was done for the study. The next element is that the treatments imposed must be *randomly assigned*. **Random assignment** helps to eliminate other confounding variables. Just as randomization helps to create a representative sample in a survey, if we randomly assign treatments to the subjects, we can increase the likelihood that the treatment groups are equally representative of the population. The other essential element of an experiment is **replication**. The conditions of a well-designed experiment will be able to be replicated by other researchers so that the results can be independently confirmed.

To design an experiment similar to the British study, we would need to use valid sampling techniques to select a representative sample of women who were attempting to conceive. (This might be difficult to accomplish!) The women might then be randomly assigned to one of three groups in which their diets would be strictly controlled. The first group would be required to skip breakfast, the second group would be put on a high-calorie, nutrition-rich diet, and the third group would be put on a low-calorie, low-nutrition diet. This brings up some ethical concerns. An experiment that imposes a treatment which could cause direct harm to the subjects is morally objectionable, and should be avoided. Since skipping breakfast could actually harm the development of the child, it should not be part of an experiment.

It would be important to closely monitor the women for successful conception to be sure that once a viable embryo is established, the mother returns to a properly nutritious pre-natal diet. The gender of the child would eventually be determined, and the results between the three groups would be compared for differences.

Control

Let’s say that your statistics teacher read somewhere that classical music has a positive effect on learning. To impose a treatment in this scenario, she decides to have students listen to an MP3 player very softly playing Mozart string quartets while they sleep for a week prior to administering a unit test. To help minimize the possibility that some other unknown factor might influence student performance on the test, she randomly assigns the class into two groups of students. One group will listen to the music, and the other group will not. When the treatment of interest is actually withheld from one of the treatment groups, it is usually referred to as the **control group**. By randomly assigning subjects to these two groups, we can help improve the chances that each group is representative of the class as a whole.

Placebos and Blind Experiments

In medical studies, the treatment group usually receives some experimental medication or treatment that has the potential to offer a new cure or improvement for some medical condition. This would mean that the control group would not receive the treatment or medication. Many studies and experiments have shown that the expectations of participants can influence the outcomes. This is especially true in clinical medication studies in which participants

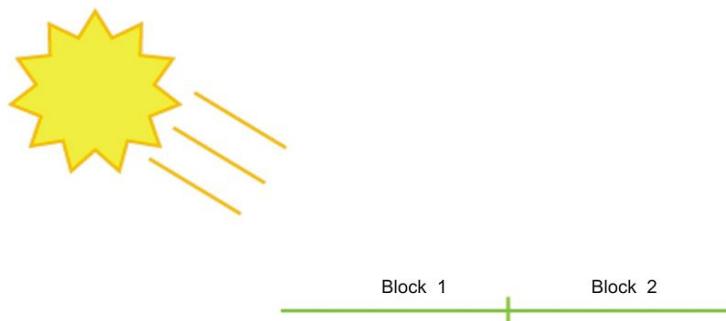
who believe they are receiving a potentially promising new treatment tend to improve. To help minimize these expectations, researchers usually will not tell participants in a medical study if they are receiving a new treatment. In order to help isolate the effects of personal expectations, the control group is typically given a **placebo**. The placebo group would think they are receiving the new medication, but they would, in fact, be given medication with no active ingredient in it. Because neither group would know if they are receiving the treatment or the placebo, any change that might result from the expectation of treatment (this is called the *placebo effect*) should theoretically occur equally in both groups, provided they are randomly assigned. When the subjects in an experiment do not know which treatment they are receiving, it is called a **blind experiment**.

Example: If you wanted to do an experiment to see if people preferred a brand-name bottled water to a generic brand, you would most likely need to conceal the identity of the type of water. A participant might expect the brand-name water to taste better than a generic brand, which would alter the results. Also, sometimes the expectations or prejudices of the researchers conducting the study could affect their ability to objectively report the results, or could cause them to unknowingly give clues to the subjects that would affect the results. To avoid this problem, it is possible to design the experiment so that the researcher also does not know which individuals have been given the treatment or placebo. This is called a **double-blind experiment**. Because drug trials are often conducted or funded by companies that have a financial interest in the success of the drug, in an effort to avoid any appearance of influencing the results, double-blind experiments are considered the gold standard of medical research.

Blocking

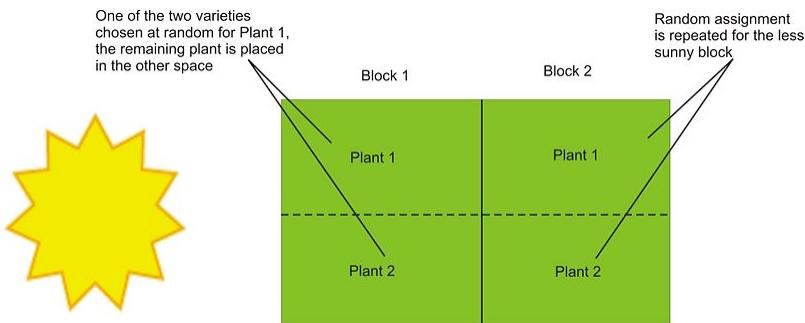
Blocking in an experiment serves a purpose similar to that of stratification in a survey. For example, if we believe men and women might have different opinions about an issue, we must be sure those opinions are properly represented in the sample. The terminology comes from agriculture. In testing different yields for different varieties of crops, researchers would need to plant crops in large fields, or blocks, that could contain variations in conditions, such as soil quality, sunlight exposure, and drainage. It is even possible that a crop's position within a block could affect its yield. Similarly, if there is a sub-group in the population that might respond differently to an imposed treatment, our results could be confounded. Let's say we want to study the effects of listening to classical music on student success in statistics class. It is possible that boys and girls respond differently to the treatment, so if we were to design an experiment to investigate the effect of listening to classical music, we want to be sure that boys and girls were assigned equally to the treatment (listening to classical music) and the control group (not listening to classical music). This procedure would be referred to as blocking on gender. In this manner, any differences that may occur in boys and girls would occur equally under both conditions, and we would be more likely to be able to conclude that differences in student performance were due to the imposed treatment. In blocking, you should attempt to create blocks that are homogenous (the same) for the trait on which you are blocking.

Example: In your garden, you would like to know which of two varieties of tomato plants will have the best yield. There is room in your garden to plant four plants, two of each variety. Because the sun is coming predominately from one direction, it is possible that plants closer to the sun would perform better and shade the other plants. Therefore, it would be a good idea to block on sun exposure by creating two blocks, one sunny and one not.



You would randomly assign one plant from each variety to each block. Then, within each block, you would randomly

assign each variety to one of the two positions.



This type of design is called ***randomized block design***.

Matched Pairs Design

A ***matched pairs design*** is a type of randomized block design in which there are two treatments to apply.

Example: Suppose you were interested in the effectiveness of two different types of running shoes. You might search for volunteers among regular runners using the database of registered participants in a local distance run. After personal interviews, a sample of 50 runners who run a similar distance and pace (average speed) on roadways on a regular basis could be chosen. Suppose that because you feel that the weight of the runners will directly affect the life of the shoe, you decided to block on weight. In a matched pairs design, you could list the weights of all 50 runners in order and then create 25 matched pairs by grouping the weights two at a time. One runner would be randomly assigned shoe A, and the other would be given shoe B. After a sufficient length of time, the amount of wear on the shoes could be compared.

In the previous example, there may be some potential confounding influences. Factors such as running style, foot shape, height, or gender may also cause shoes to wear out too quickly or more slowly. It would be more effective to compare the wear of each shoe on each runner. This is a special type of matched pairs design in which each experimental unit becomes its own matched pair. Because the matched pair is in fact two different observations of the same subject, it is called a ***repeated measures design***. Each runner would use shoe A and shoe B for equal periods of time, and then the wear of the shoes for each individual would be compared. Randomization could still be important, though. Let's say that we have each runner use each shoe type for a period of 3 months. It is possible that the weather during those three months could influence the amount of wear on the shoe. To minimize this, we could randomly assign half the subjects shoe A, with the other half receiving shoe B, and then switch after the first 3 months.

Lesson Summary

The important elements of a statistical experiment are **randomness, imposed treatments, and replication**. The use of these elements is the only effective method for establishing meaningful cause-and-effect relationships. An experiment attempts to isolate, or control, other potential variables that may contribute to changes in the response variable. If these other variables are known quantities but are difficult, or impossible, to distinguish from the other explanatory variables, they are called **confounding variables**. If there is an additional explanatory variable affecting the response variable that was not considered in an experiment, it is called a **lurking variable**. A **treatment** is the term used to refer to a condition imposed on the subjects in an experiment. An experiment will have at least two treatments. When trying to test the effectiveness of a particular treatment, it is often effective to withhold applying that treatment to a group of randomly chosen subjects. This is called a **control group**. If the subjects are aware of the conditions of their treatment, they may have preconceived expectations that could affect the outcome. Especially in medical experiments, the psychological effect of believing you are receiving a potentially effective treatment can

lead to different results. This phenomenon is called the **placebo effect**. When the participants in a clinical trial are led to believe they are receiving the new treatment, when, in fact, they are not, they receive what is called a placebo. If the participants are not aware of the treatment they are receiving, it is called a **blind experiment**, and when neither the participant nor the researcher is aware of which subjects are receiving the treatment and which subjects are receiving a placebo, it is called a **double-blind experiment**.

Blocking is a technique used to control the potential confounding of variables. It is similar to the idea of stratification in sampling. In a **randomized block design**, the researcher creates blocks of subjects that exhibit similar traits that might cause different responses to the treatment and then randomly assigns the different treatments within each block. A **matched pairs design** is a special type of design where there are two treatments. The researcher creates blocks of size 2 on some similar characteristic and then randomly assigns one subject from each pair to each treatment. **Repeated measures designs** are a special matched pairs experiment in which each subject becomes its own matched pair by applying both treatments to the subject and then comparing the results.

Points to Consider

- What are some other ways that researchers design more complicated experiments?
- When one treatment seems to result in a notable difference, how do we know if that difference is statistically significant?
- How can the selection of samples for an experiment affect the validity of the conclusions?

Review Questions

1. As part of an effort to study the effect of intelligence on survival mechanisms, scientists recently compared a group of fruit flies intentionally bred for intelligence to the same species of ordinary flies. When released together in an environment with high competition for food, the percentage of ordinary flies that survived was significantly higher than the percentage of intelligent flies that survived.
 - a. Identify the population of interest and the treatments.
 - b. Based on the information given in this problem, is this an observational study or an experiment?
 - c. Based on the information given in this problem, can you conclude definitively that intelligence decreases survival among animals?
2. In order to find out which brand of cola students in your school prefer, you set up an experiment where each person will taste two brands of cola, and you will record their preference.
 - a. How would you characterize the design of this study?
 - b. If you poured each student a small cup from the original bottles, what threat might that pose to your results? Explain what you would do to avoid this problem, and identify the statistical term for your solution.
 - c. Let's say that one of the two colas leaves a bitter after-taste. What threat might this pose to your results? Explain how you could use randomness to solve this problem.
3. You would like to know if the color of the ink used for a difficult math test affects the stress level of the test taker. The response variable you will use to measure stress is pulse rate. Half the students will be given a test with black ink, and the other half will be given the same test with red ink. Students will be told that this test will have a major impact on their grades in the class. At a point during the test, you will ask the students to stop for a moment and measure their pulse rates. In preparation for this experiment, you measure the at-rest pulse rates of all the students in your class.

Here are those pulse rates in beats per minute:

TABLE 6.1:

Student Number	At Rest Pulse Rate
1	46
2	72
3	64
4	66
5	82
6	44
7	56
8	76
9	60
10	62
11	54
12	76

- (a) Using a matched pairs design, identify the students (by number) that you would place in each pair.
- (b) Use the following table of random digits to randomly assign each student in each pair to a treatment. Explain how you made your assignments. 07081 41008 43563 56934 51119 46109 09931
4. It is a well-known fact that there is a strong association between a student's high school GPA and the student's SAT score. What might be a lurking variable contributing to this association? Explain.
5. People who attend church regularly and are active in their religious lives are known to live longer than those who are not regular attenders and are not active in their religious lives. What confounding factors might be present that also might contribute to the longevity of the religious people?
6. A new blood pressure medication is being tested. A researcher obtains 48 male volunteer subjects, 24 white males and 24 black males, all with essentially the same high blood pressure and similar health issues. He has concerns that there might be differences in the response to the medication due to race, so he decides to block on race.
- (a) Describe an appropriate block experiment. (b) How many groups will he have? (c) How many subjects will be in each group, and what will be the racial breakout of each group? (d) Will he need a control group? (e) What will he measure at the start of the experiment? (f) What will he measure at the end of the experiment? (g) Should the experiment be single-blind, double-blind, or is blinding not necessary?
7. A medical study of breast-feeding shows that babies who are breast-fed tend to have higher IQs than those who are not breast-fed. There is a potential "cause-and-effect" relationship that might be inferred from this study. Explain the possible faulty logic.

Further reading:

<http://www.nytimes.com/2008/05/06/health/research/06epil.html?ref=health>

References

<http://journals.royalsociety.org/content/w260687441pp64w5/>

http://www.fhi.no/eway/default.aspx?pid=238&trg=Area_5954&MainLeft_5812=5954:0:&Area_5954=5825:68516::0:5956:1::0:0

Part One: Multiple Choice

1. A researcher performs an experiment to see if mice can learn their way through a maze better when given a high-protein diet and vitamin supplements. She carefully designs and implements a study with the random

assignment of the mice into treatment groups and observes that the mice on the special diet and supplements have significantly lower maze times than those on normal diets. She obtains a second group of mice and performs the experiment again. This is most appropriately called:

- a. Matched pairs design
 - b. Repeated measures
 - c. Replication
 - d. Randomized block design
 - e. Double blind experiment
2. Which of the following terms does not apply to experimental design?
 - a. Randomization
 - b. Stratification
 - c. Blocking
 - d. Cause and effect relationships
 - e. Placebo
 3. An exit pollster is given training on how to spot the different types of voters who would typically represent a good cross-section of opinions and political preferences for the population of all voters. This type of sampling is called:
 - a. Cluster sampling
 - b. Stratified sampling
 - c. Judgment sampling
 - d. Systematic sampling
 - e. Quota sampling

Use the following scenario to answer questions 4 and 5. A school performs the following procedure to gain information about the effectiveness of an agenda book in improving student performance. In September, 100 students are selected at random from the school's roster. The interviewer then asks the selected students if they intend to use their agenda books regularly to keep track of their assignments. Once the interviewer has 10 students who will use their book and 10 students who will not, the rest of the students are dismissed. Next, the selected students' current averages are recorded. At the end of the year, the grades for each group are compared, and overall, the agenda-book group has higher grades than the non-agenda group. The school concludes that using an agenda book increases student performance.

4. Which of the following is true about this situation?
 - a. The response variable is using an agenda book.
 - b. The explanatory variable is grades.
 - c. This is an experiment, because the participants were chosen randomly.
 - d. The school should have stratified by gender.
 - e. This is an observational study, because no treatment is imposed.
5. Which of the following is not true about this situation?
 - a. The school cannot conclude a cause-and-effect relationship, because there is most likely a lurking variable that is responsible for the differences in grades.
 - b. This is not an example of a matched pairs design.
 - c. The school can safely conclude that the grade improvement is due to the use of an agenda book.
 - d. Blocking on previous grade performance would help isolate the effects of potential confounding variables.
 - e. Incorrect response bias could affect the selection of the sample.

Part Two: Open-Ended Questions

- During the 2004 presidential election, early exit polling indicated that Democratic candidate John Kerry was doing better than expected in some eastern states against incumbent George W. Bush, causing some to even predict that he might win the overall election. These results proved to be incorrect. Again, in the 2008 New Hampshire Democratic primary, pre-election polling showed Senator Barack Obama winning the primary. It was, in fact, Senator Hillary Clinton who comfortably won the contest. These problems with exit polling lead to many reactions, ranging from misunderstanding the science of polling, to mistrust of all statistical data, to vast conspiracy theories. The Daily Show from Comedy Central did a parody of problems with polling. Watch the clip online at the following link. Please note that while “bleeped out,” there is language in this clip that some may consider inappropriate or offensive. <http://www.thedailyshow.com/video/index.jhtml?videoId=156231&title=team-daily-polls> What type of bias is the primary focus of this non-scientific, yet humorous, look at polling?
- Environmental Sex Determination is a scientific phenomenon observed in many reptiles in which air temperature when eggs are growing tends to affect the proportion of eggs that develop into male or female animals. This has implications for attempts to breed endangered species, as an increased number of females can lead to higher birth rates when attempting to repopulate certain areas. Researchers in the Galapagos wanted to see if the Galapagos Giant Tortoise eggs were also prone to this effect. The original study incubated eggs at three different temperatures: 25.50 C, 29.50 C, and 33.50 C. Let's say you had 9 female tortoises, and there was no reason to believe that there was a significant difference in eggs from these tortoises.
 - Explain how you would use a randomized design to assign the treatments and carry out the experiment.
 - If the nine tortoises were composed of three tortoises each of three different species, how would you design the experiment differently if you thought that there might be variations in response to the treatments?
- A researcher who wants to test a new acne medication obtains a group of volunteers who are teenagers taking the same acne medication to participate in a study comparing the new medication with the standard prescription. There are 12 participants in the study. Data on their gender, age, and the severity of their condition are given in the following table:

TABLE 6.2:

Subject Number	Gender	Age	Severity
1	M	14	Mild
2	M	18	Severe
3	M	16	Moderate
4	F	16	Severe
5	F	13	Severe
6	M	17	Moderate
7	F	15	Mild
8	M	14	Severe
9	F	13	Moderate
10	F	17	Moderate
11	F	18	Mild
12	M	15	Mild

- Identify the treatments, and explain how the researcher could use blinding to improve the study.
- Explain how you would use a completely randomized design to assign the subjects to treatment groups.
- The researcher believes that gender and age are not significant factors, but is concerned that the original severity

of the condition may have an effect on the response to the new medication. Explain how you would assign treatment groups while blocking for severity.

(d) If the researcher chose to ignore pre-existing condition and decided that both gender and age could be important factors, he or she might use a matched pairs design. Identify which subjects you would place in each of the 6 matched pairs, and provide a justification of how you made your choice.

(e) Why would you avoid a repeated measures design for this study?

Keywords

Bias

The term most frequently applied to a non-representative sample is *bias*.

Blind experiment

When the subjects in an experiment do not know which treatment they are receiving, it is called a *blind experiment*.

Blocking

Blocking in an experiment serves a purpose similar to that of stratification in a survey.

Census

If you collect information from every unit in a population, it is called a census.

Cluster sampling

Cluster sampling is when a naturally occurring group is selected at random, and then either all of that group, or randomly selected individuals from that group, are used for the sample.

Confounding variables

A slightly different type of additional variable is called a confounding variable. *Confounding variables* are those that affect the response variable and are also related to the explanatory variable.

Control group

The control group is typically given a *placebo*.

Convenience sampling

Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher.

Double blind experiment

A double blind experiment is an experimental method used to ensure impartiality, and avoid errors arising from bias.

Experiment

An *experiment* is a methodical procedure carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis.

Incorrect response bias

When an individual intentionally responds to a survey with an untruthful answer, this is called *incorrect response bias*.

Incorrect sampling frame

If the list from which you choose your sample does not accurately reflect the characteristics of the population, this is called *incorrect sampling frame*.

Judgement sampling

In judgement sampling, the researcher or some other "expert" uses his/her judgement in selecting the units from the population for study based on the population's parameters.

Lurking variable

A variable that is not included in a study but that may still have an effect on the other variables involved is called a *lurking variable*.

Margin of error

The **margin of error** is a statistic expressing the amount of random sampling error in a survey's results.

Matched pairs design

A *matched pairs design* is a type of randomized block design in which there are two treatments to apply.

Multi-stage sampling

If we select at random from out of that group, or cluster into smaller subgroups, this is referred to as *multi-stage sampling*.

Non-response bias

how much these individuals' beliefs and opinions reflect those of the general population, and, therefore, almost all surveys could be prone to *non-response bias*.

Observational study

In an *observational study* such as the Royal Society's connecting gender and a mother's diet, it is possible that there is a third variable that was not observed that is causing a change in both the explanatory and response variables.

Placebo

In order to help isolate the effects of personal expectations, the control group is typically given a *placebo*.

Placebo effect

Especially in medical experiments, the psychological effect of believing you are receiving a potentially effective treatment can lead to different results. This phenomenon is called the placebo effect.

Questionnaire bias

Questionnaire bias occurs when the way in which the question is asked influences the response given by the individual.

Quota sampling

In *quota sampling*, an individual or organization attempts to include the proper proportions of individuals of different subgroups in their sample.

Random sample

In statistical terms a random sample is a set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample.

Randomization

The best technique for reducing bias in sampling is *randomization*.

Randomized block design

The Randomized Block Design is research design's equivalent to stratified random sampling.

Randomly assigned

The next element is that the treatments imposed must be *randomly assigned*.

Repeated measures design

If the matched pair is in fact two different observations of the same subject, it is called a *repeated measures design*.

Replication

The other essential element of an experiment is *replication*. The conditions of a well-designed experiment will be able to be replicated by other researchers so that the results can be independently confirmed.

Response bias

The term *response bias* refers to problems that result from the ways in which the survey or poll is actually presented to the individuals in the sample.

Sample

A *sample* is a representative subset of a population.

Sampling error

The downside to sampling is that we can never be completely sure that we have captured the truth about the entire population, due to random variation in our sample that is called sampling error.

Sampling frame

The *sampling frame* is the term we use to refer to the group or listing from which the sample is to be chosen.

Seed value

The function that simulates randomness has to start at some number, called a *seed value*.

Simple random sample

When a *simple random sample* of size n (commonly referred to as an SRS) is taken from a population, all possible samples of size n in the population have an equal probability of being selected for the sample.

Size bias

If one particular subgroup in a population is likely to be over-represented or under-represented due to its size, this is sometimes called *size bias*.

Stratified sampling

In *stratified sampling*, the population is divided into groups, called strata (the singular term is 'stratum'), that have some meaningful relationship.

Systematic sampling

In *systematic sampling*, after choosing a starting point at random, subjects are selected using a jump number.

Treatment

The first is that a *treatment* must be imposed on the subjects of the experiment.

Undercoverage

A term often used to describe the problems when a group of the population is not represented in a survey is *undercoverage*.

Voluntary response bias

All of these polls usually come with a disclaimer stating that, “This is not a scientific poll.” While perhaps entertaining, these types of polls are very susceptible to *voluntary response bias*.

CHAPTER**7**

Sampling Distributions and Estimations

Chapter Outline

- 7.1 INTRODUCTION TO SAMPLING DISTRIBUTIONS**
 - 7.2 THE CENTRAL LIMIT THEOREM**
 - 7.3 CONFIDENCE INTERVALS WITH Z-VALUES**
 - 7.4 REFERENCES**
-

7.1 Introduction to Sampling Distributions

Learning Objectives

- Define inferential statistics
- Graph a probability distribution for the mean of a discrete variable
- Describe a sampling distribution in terms of "all possible outcomes"
- Describe a sampling distribution in terms of repeated sampling
- Describe the role of sampling distributions in inferential statistics
- Define the standard error of the mean

The concept of a **sampling distribution** is perhaps the most basic concept in inferential statistics. It is also a difficult concept to teach because a sampling distribution is a theoretical distribution rather than an empirical distribution.

This introductory section defines the concept and gives an example for both a discrete and a continuous distribution. It also discusses how sampling distributions are used in inferential statistics.

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would probably not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall that inferential statistics concerns generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter. (In this example, the sample statistics are the sample means and the population parameter is the population mean.)

Discrete Distributions

We will illustrate the concept of sampling distributions with a simple example. Figure 1 shows three pool balls, each with a number on it.



FIGURE 7.1

Figure 1. The pool balls.

Let's first consider the original probability distribution for these three balls. If we draw one ball at random, it is obvious that each ball has a probability of $1/3$ of being chosen. We can graph this probability distribution, and it is a uniform distribution, with all 3 probabilities equal.

Figure 1.5. Probability distribution (uniform distribution) for pool balls, with $n = 1$.

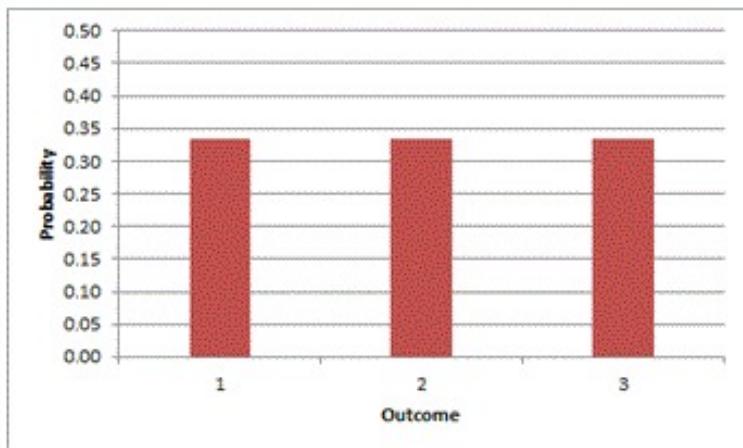


FIGURE 7.2

Now, let's choose two of the balls at random (with replacement) and compute the average of their numbers. All possible outcomes are shown below in Table 1.

Table 1. All possible outcomes when **two** balls are sampled with replacement.

TABLE 7.1:

Outcome	Ball 1	Ball 2	Mean
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in Table 2. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Table 2. Frequencies of means for $N = 2$.

TABLE 7.2:

Mean	Frequency	Relative Frequency
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

Figure 2 shows a relative frequency distribution of the means based on Table 2. This distribution is also a probability distribution since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.

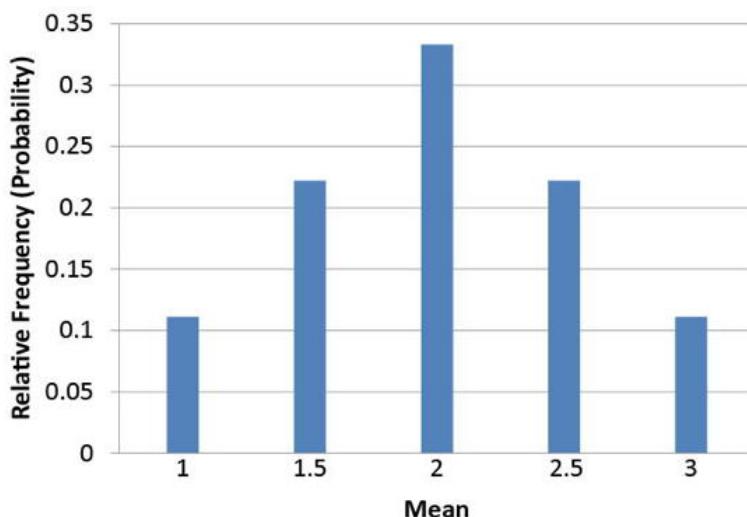


FIGURE 7.3

Figure 2. Distribution of means for $n = 2$.

The distribution shown in Figure 2 is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 ($n = 2$). For this simple example, the distribution of pool balls and the sampling distribution are both **discrete** distributions. The pool balls have only the numbers 1, 2, and 3, and a sample mean can have one of only five possible values.

Now compare the shapes of the two graphs. The first distribution was the distribution resulting from drawing one ball at random. The second distribution resulted from drawing 2 balls (with replacement) and calculating the mean of those two draws. We took all possible samples of size 2, calculated each mean, and then we created the frequency distribution for the 9 means. The shapes of the two (discrete) bar charts are quite different. One is flat (uniform), and the other is more bell-shaped. Figure 1.5 is the original distribution. Figure 2 is the **sampling distribution** of all possible sample means of size $n = 2$.

In the next section we will generalize this result when we learn the Central Limit Theorem.

Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? (Although this distribution is not really continuous, it is close enough to be considered continuous for practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two balls. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this example since there are 1,000,000 possible outcomes (1,000 for the first ball \times 1,000 for the second). Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions, specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed. This idea will be explored more fully in the next section.

Sampling Distributions and Inferential Statistics

As we stated in the beginning of this chapter, sampling distributions are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean was determined. In practice, the process proceeds the other way: you collect sample data and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing

the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the **standard error of the mean**. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Lesson Summary

In this lesson, we have learned about the sampling distribution of the mean for both discrete and continuous random variables. When we study the distribution NOT of the individual data, but rather the distribution of the MEAN of all samples of a particular size, then we have a sampling distribution. In the next section we will study how sample means and sample proportions are distributed when we learn the Central Limit Theorem. We will learn how to calculate probabilities using the sampling distribution of the mean and the sampling distribution of the proportion. Then we will be ready to enter the realm of inferential statistics.

Concept Questions

- Does the mean of the sampling distribution equal the mean of the population?
- If the sampling distribution is normally distributed, is the population normally distributed?
- Are there any restrictions on the size of the sample that is used to estimate the parameters of a population?

7.2 The Central Limit Theorem

Learning Objectives

- State the mean and variance of the sampling distribution of the mean
- Compute the standard error of the mean
- State the Central Limit Theorem
- State the mean and variance of the sampling distribution of the proportion
- Compute the standard error of the proportion

Introduction

In the previous lesson, we learned that sampling is an important tool for determining the characteristics of a population. We used the pool balls to illustrate that when we take all possible samples of size 2 and calculate each sample mean, the shape of the distribution of the means is quite different from the original distribution's shape. It is now time to define some properties of a sampling distribution of sample means and sample proportions and to examine what we can conclude about the entire population based on these properties.

The sampling distribution of the mean was defined in the section introducing sampling distributions. This section reviews some important properties of the sampling distribution of the mean.

Mean

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ , then the mean of the sampling distribution of the mean is also μ . The symbol $\mu_{\bar{x}}$ is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as: $\mu_{\bar{x}} = \mu$. In words, we are saying that the mean of the sampling distribution is the original population mean.

Variance

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

That is, the variance of the sampling distribution of the mean is the population variance divided by n , the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

The **standard error of the mean** is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

The standard error is represented by a σ because it is a standard deviation. The subscript (\bar{x}) indicates that the standard error in question is the standard error of the mean.

Central Limit Theorem

In the pool ball example, we saw that the uniform distribution was flat, but its sampling distribution, even for $n = 2$, was higher in the middle and lower in the tails. We can generalize that observation to the more formal statement of the Central Limit Theorem. We will see more examples to illustrate it.

The Central Limit Theorem states that:

Given a population with mean μ and standard deviation σ , the sampling distribution of the mean approaches a **normal distribution** with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$, which is also called the standard error of the mean.

The expressions for the mean and standard deviation of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as n increases. Figure 1 shows the results of a simulation for $n = 2$ and $n = 10$. The parent population was a uniform distribution, which means that it was flat-topped, with all probabilities equal.

You can see that the distribution for $n = 2$ is far from a normal distribution. Nonetheless, it does show that the scores are denser in the middle than in the tails. For $n = 10$ the distribution is quite close to a normal distribution. Notice that the means of the two distributions are the same, but that the spread of the distribution for $n = 10$ is smaller.

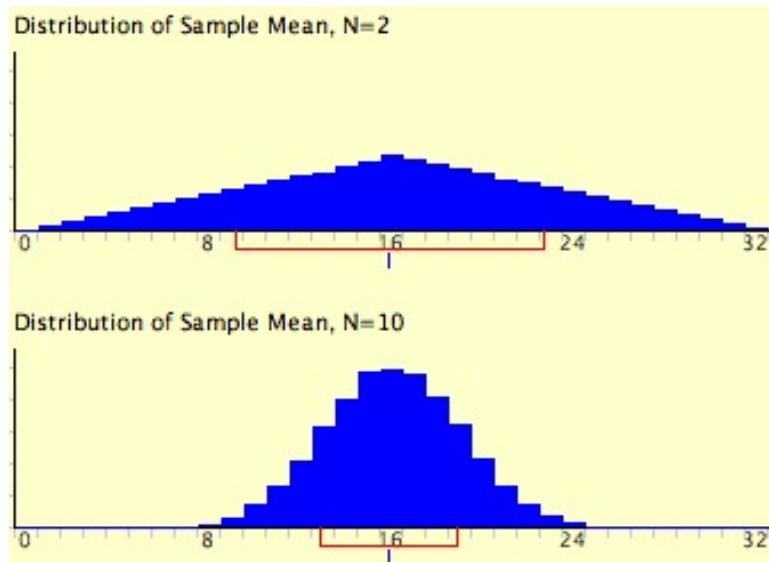


FIGURE 7.4

Figure 1. A simulation of a sampling distribution. The parent population is uniform. The blue line under "16" indicates that 16 is the mean. The red line extends from the mean plus and minus one standard deviation.

From Figure 1, you can see that the mean is the same for each distribution, but the standard deviation is much smaller in the second distribution, when the sample size is larger ($n = 10$ versus $n = 2$).

Figure 2 shows how closely the sampling distribution of the mean approximates a normal distribution even when the parent population is very non-normal. The larger the sample size, the closer the sampling distribution of the mean would be to a normal distribution.

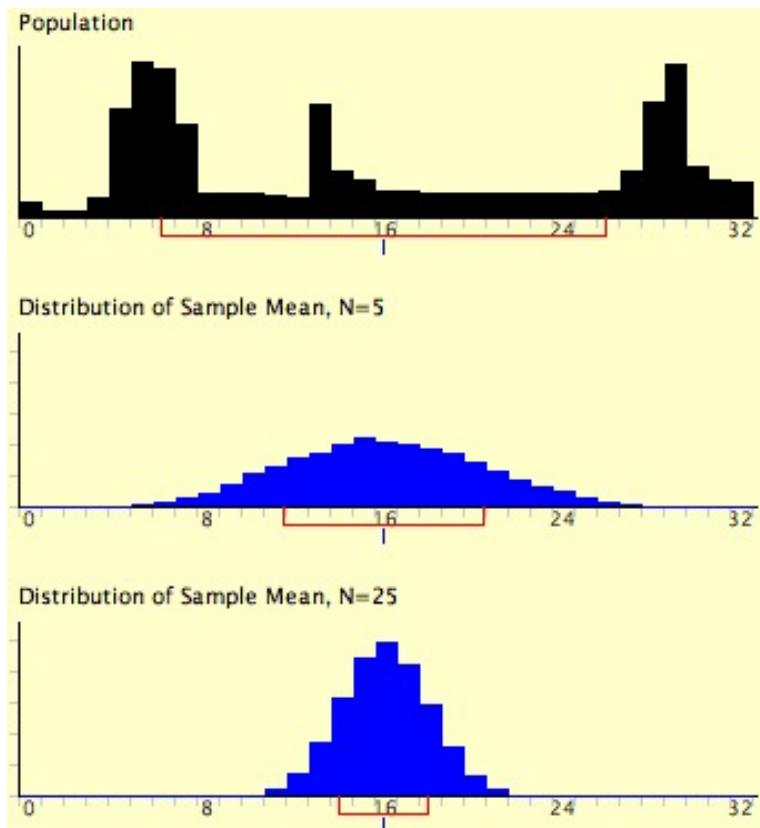


FIGURE 7.5

Figure 2. A simulation of a sampling distribution. The parent population is very non-normal.

The Central Limit Theorem Revisited, with some examples

The *Central Limit Theorem* is a very important theorem in statistics. It basically confirms what might be an intuitive truth to you: that as you increase the sample size for a random variable, the distribution of the sample means better approximates a normal distribution.

Before going any further, you should become familiar with (or reacquaint yourself with) the symbols that are commonly used when dealing with properties of the sampling distribution of sample means. These symbols are shown in the table below:

TABLE 7.3:

	Population Parameter	Sample Statistic	Sampling Distribution
Mean	μ	\bar{x}	$\mu_{\bar{x}}$
Standard Deviation	σ	s	$S_{\bar{x}}$ or $\sigma_{\bar{x}}$
Size	N	n	

As the sample size, n , increases, the resulting sampling distribution would approach a normal distribution with the same mean as the population and with $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The notation $\sigma_{\bar{x}}$ reminds you that this is the standard deviation of the distribution of sample means and not the standard deviation of a single observation.

The Central Limit Theorem states the following:

If samples of size n are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means, \bar{x} , approximates a normal distribution as n increases.

The mean of this sampling distribution approximates the population mean, and the standard deviation of this sampling distribution approximates the standard deviation of the population divided by the square root of the sample size: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

These properties of the sampling distribution of sample means can be applied to determining probabilities. If the sample size is sufficiently large (> 30), the sampling distribution of sample means can be assumed to be approximately normal, even if the parent population is not normally distributed.

Example: Suppose you wanted to answer the question, “What is the probability that a random sample of 20 families in Canada will have an average of 1.5 pets or fewer?” where the mean of the population is 0.8 and the standard deviation of the population is 1.2.

For the sampling distribution, $\mu_{\bar{x}} = \mu = 0.8$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{20}} = 0.268$.

Using technology, a sketch of this problem is as follows:



The shaded area shows the probability that the sample mean is less than 1.5.

The z -score for the value 1.5 is $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1.5 - 0.8}{0.268} \approx 2.6$.

As shown above, the area under the standard normal curve to the left of 1.5 (a z -score of 2.6) is approximately 0.9937. This value can also be determined by using a graphing calculator as follows:

```
normalcdf(-1E99,
1.5,.8,.27)
.9937136558
■
```

Therefore, the probability that the sample mean will be below 1.5 is 0.9937. In other words, with a random sample of 20 families, it is almost definite that the average number of pets per family will be less than 1.5.

Example: A random sample of size 40 is selected from a known population with a mean of 23.5 and a standard deviation of 4.3. Samples of the same size are repeatedly collected, allowing a sampling distribution of sample means to be drawn.

- What is the expected shape of the resulting distribution?
- Where is the sampling distribution of sample means centered?
- What is the approximate standard deviation of the sample means?

The question indicates that multiple samples of size 40 are being collected from a known population, multiple sample means are being calculated, and then the sampling distribution of the sample means is being studied. Therefore, an understanding of the Central Limit Theorem is necessary to answer the question.

Answers:

- a) The sampling distribution of the sample means will be approximately bell-shaped. (Note that the sample size is greater than 30, so this is essentially a guarantee that the sampling distribution will be bell-shaped.)
- b) The sampling distribution of the sample means will be centered about the population mean of 23.5.
- c) The approximate standard deviation of the sample means is 0.68, which can be calculated as shown below:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ \sigma_{\bar{x}} &= \frac{4.3}{\sqrt{40}} \\ \sigma_{\bar{x}} &= 0.68\end{aligned}$$

Example: Multiple samples with a sample size of 40 are taken from a known population, where $\mu = 25$ and $\sigma = 4$. The following chart displays the sample means:

25	25	26	26	26	24	25	25	24	25
26	25	26	25	24	25	25	25	25	25
24	24	24	24	26	26	26	25	25	25
25	25	24	24	25	25	25	24	25	25
25	24	25	25	24	26	24	26	24	26

- a) What is the population mean?
- b) Using technology, determine the mean of the sample means.
- c) What is the population standard deviation?
- d) Using technology, determine the standard deviation of the sample means.
- e) As the sample size increases, what value will the mean of the sample means approach?
- f) As the sample size increases, what (theoretical) value will the standard deviation of the sample means approach?

Answers:

- a) The population mean of 25 was given in the question: $\mu = 25$.
- b) The mean of the sample means is 24.94 and is determined by entering the 40 data values into List 1 and then using '1 Var Stats' on the TI-83/84 calculator: $\mu_{\bar{x}} = 24.94$.
- c) The population standard deviation of 4 was given in the question: $\sigma = 4$.
- d) The standard deviation of the sample means is 0.71 and is determined by using '1 Vars Stat' on the TI-83/84 calculator: $S_{\bar{x}} = 0.71$. Note that the Central Limit Theorem states that the standard deviation should be approximately $\frac{4}{\sqrt{40}} = 0.63$.
- e) The mean of the sample means will approach 25 and is determined by a property of the Central Limit Theorem: $\mu_{\bar{x}} = 25$.
- f) The standard deviation of the sample means will approach $\frac{4}{\sqrt{n}}$ and is determined by a property of the Central Limit Theorem: $\sigma_{\bar{x}} = \frac{4}{\sqrt{n}}$. (Note that as n gets bigger and bigger, the standard deviation will get smaller and smaller, and it actually is equal to 0 when we take a sample so large that it includes every individual in the entire population.)

Sampling Distribution of the Proportion \hat{p}

Thus far, we have used examples and illustrations of the Central Limit Theorem (CLT) using sample means. But we can extend the CLT to sample proportions as well. Here is a discussion and an example:

Assume that in an election race between Candidate A and Candidate B, 0.60 of the voters prefer Candidate A. If a random sample of 10 voters were polled, it is unlikely that exactly 60% of them (6) would prefer Candidate A. By chance the proportion in the sample preferring Candidate A could easily be a little lower than 0.60 or a little higher than 0.60. The sampling distribution of \hat{p} is the distribution that would result if you repeatedly sampled 10 voters and determined the sample proportion for each sample of 10 voters that favored Candidate A.

The sampling distribution of \hat{p} is a special case of the sampling distribution of the mean. Table 1 shows a hypothetical random sample of 10 voters. Those who prefer Candidate A are given scores of 1 and those who prefer Candidate B are given scores of 0. Note that seven of the voters prefer candidate A so the sample proportion (\hat{p}) is

$$\hat{p} = 7/10 = 0.70$$

As you can see, \hat{p} is the mean of the 10 preference scores.

Table 1. Sample of voters.

TABLE 7.4:

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

The distribution of \hat{p} is closely related to the binomial distribution. The binomial distribution is the distribution of the total **number** of successes (favoring Candidate A, for example) whereas the distribution of \hat{p} is the distribution of the **proportion** of successes. The mean, of course, is the total divided by the sample size, n . Therefore, the sampling distribution of \hat{p} and the binomial distribution differ in that \hat{p} is the **mean proportion** of the scores (0.70) and the binomial distribution is dealing with the total **number of successes**(7).

The binomial distribution has a mean of $\mu = np$

Dividing by n to adjust for the fact that the sampling distribution of \hat{p} is dealing with the mean proportion instead of totals, we find that the mean of the sampling distribution of \hat{p} is:

$$\mu_{\hat{p}} = p$$

Another way of saying this is that the mean of all the sample proportions (all of the \hat{p} s) is the population proportion p .

Remember that the standard deviation of the binomial distribution is: $\sigma = \sqrt{np(1-p)}$

Dividing by n because p is a mean and not a total, we find the **standard error** of \hat{p} as $\frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$

Returning to the voter example, $p = 0.60$ and $n = 10$. (Don't confuse $p = 0.60$, the population proportion and $\hat{p} = 0.70$, the sample proportion.) Therefore, the mean of the sampling distribution of \hat{p} is 0.60. The standard error is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-.60)}{10}} = 0.155$$

FIGURE 7.6

The sampling distribution of \hat{p} is a discrete rather than a continuous distribution. For example, with an n of 10, it is possible to have a \hat{p} of 0.50 or a \hat{p} of 0.60 but not a \hat{p} of 0.55.

The sampling distribution of \hat{p} is approximately normally distributed if n is fairly large and p is not close to 0 or 1. A rule of thumb is that the approximation is good if both np and $n(1 - p)$ are greater than 10. The sampling distribution for the voter example is shown in Figure 1. Note that even though $n(1 - p)$ is only 4, the approximation is still quite good.

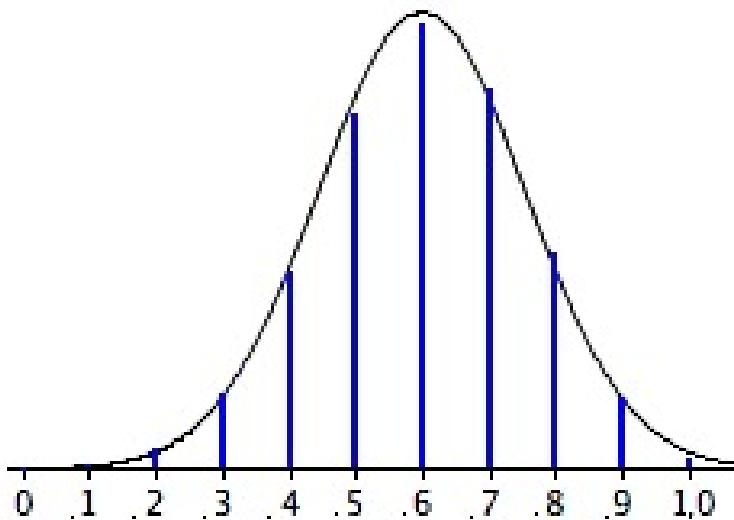


FIGURE 7.7

Figure 1. The sampling distribution of p . Vertical bars are the probabilities; the smooth curve is the normal approximation.

Example:

It is known that in a particular school district, the proportion of parents in favor of combining two elementary schools into one school (to save money) is .43. We draw a random sample of 100 parents. What is the probability that the sample proportion of parents in this sample in favor of combining the 2 schools will be greater than .50?

Answer:

We can use the normal distribution z table to answer this question. We know that $p = .43$ and we want to know the probability that the sample proportion \hat{p} will be greater than .50. We use the familiar z formula as follows:

Now it is easy to use the z table or your calculator to find the probability that z is greater than 1.41, and the answer

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.50 - .43}{\sqrt{\frac{.43(.57)}{100}}} = \frac{.07}{.0495} = 1.41$$

FIGURE 7.8

is .0793. Thus, when the **population** proportion is .43, the probability that our **sample proportion** will be .50 or greater (with a sample size of 100) is about .0793.

This example started with the true population proportion being known. In future lessons, we will not know this value, but we will be able to estimate it by using the sampling distribution of the sample proportion.

Lesson Summary

The Central Limit Theorem confirms the intuitive notion that as the sample size increases for a random variable, the distribution of the sample means will begin to approximate a normal distribution, with the mean equal to the mean of the underlying population and the standard error equal to the standard deviation of the population divided by the square root of the sample size, n .

The sampling distribution of the sample proportion \hat{p} also approximates a normal distribution, with the mean proportion equal to the population proportion, and standard error equal to the "old" binomial standard deviation divided by n .

Point to Consider

- How does sample size affect the variation in sample results?

Review Questions

1. The lifetimes of a certain type of calculator battery are normally distributed. The mean lifetime is 400 days, with a standard deviation of 50 days. For a sample of 100 new batteries, determine the probability that the sample mean will be:
 - greater than 407 days.
 - between 395 and 405 days.
 - less than 390 days.
2. Assume that a human pregnancy has a population mean of 280 days and a population standard deviation of 8 days. Determine the following probabilities:
 - The probability that one randomly chosen pregnant woman exceeds 284 days before giving birth.
 - We take a random sample of $n = 5$ pregnant women. Find the probability that the mean number of days for this group of 5 women exceeds 284 days.
 - Compare your answers for parts a. and b. Why are they different?
 - What was the value of the standard error for part a? For part b? Why are these different?
3. It is known that 35% of freshmen at a large university applied to at least one other university when they were in high school. We draw a random sample of 200 freshmen. What is the probability that our sample will be within 2 percentage points of the true value?

7.3 Confidence Intervals with z-values

Learning Objectives

- Calculate the mean of a sample as a point estimate of the population mean.
- Construct a confidence interval for a population mean based on a sample mean.
- Calculate a sample proportion as a point estimate of the population proportion.
- Construct a confidence interval for a population proportion based on a sample proportion.
- Calculate the margin of error for a point estimate as a function of sample mean or proportion and size.
- Understand the logic of confidence intervals, as well as the meaning of confidence level and confidence intervals.

Introduction

The objective of inferential statistics is to use sample data to increase knowledge about the entire population. In this lesson, we will examine how to use samples to make estimates about the populations from which they came. We will also see how to determine how wide these estimates should be and how confident we should be about them.

One of the major applications of statistics is estimating population parameters from sample statistics . For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium. Out of a random sample of 200 people, 106 say they support the proposition. Thus in the sample, 0.53 of the people supported the proposition. This value of 0.53 is called a **point estimate** of the population proportion. It is called a point estimate because the estimate consists of a single value or point.

Point estimates are usually supplemented by interval estimates called **confidence intervals**. Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time. For example, if the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \mu < 0.60$. The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal. The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%. Another way of expressing this is $53\% \pm 7\%$.

Say you were interested in the mean weight of 10-year-old girls living in the United States. Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds. This sample mean of 90 is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean. For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the

intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.

It seems reasonable to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple. Let's consider the 95% confidence interval calculated above: $72.85 < \mu < 107.15$. The appropriate interpretation of this interval is this: "**We are 95% confident that the true population mean is between 72.85 and 107.15.**" We don't use a probability statement when we interpret confidence intervals. The true value of the parameter either is or is not captured in the interval.

It is important to understand the term "confidence level." This is the percent that is stated in the confidence interval. A 95% confidence interval uses a 95% confidence level. A 92% confidence interval uses a 92% confidence level. Note that a 95% confidence **interval** is a range of values that (we hope) contains the value of the true population parameter, such as μ , whereas a confidence **level** is the value that we use (such as 95% or 90%) for determining the amount of confidence about the interval we have just calculated. Here is a short interpretation of a 95% confidence **level**: If we were to take many, many samples of a specified size, and we calculated a 95% confidence interval for each one of those samples, then about 95% of the computed confidence intervals would capture the true value of the parameter.

When you compute a confidence interval, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean, there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

CONFIDENCE INTERVAL FOR A POPULATION MEAN

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from the section on the sampling distribution of the mean that the mean of the sampling distribution is μ and the standard error of the mean is $\sigma_x = \frac{\sigma}{\sqrt{n}}$. For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of $36/3 = 12$. Note that the standard deviation of a sampling distribution is its standard error. Figure 1 shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48 to 113.52. These limits were computed by adding and subtracting 1.96 standard errors to/from the mean of 90 as follows:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} = 90 - 1.96 \left(\frac{36}{\sqrt{9}} \right) = 66.48 \quad \bar{x} + z^* \frac{\sigma}{\sqrt{n}} = 90 + 1.96 \left(\frac{36}{\sqrt{9}} \right) = 113.52$$

FIGURE 7.9

The 95% confidence interval is: $66.48 \leq \mu \leq 113.52$. Based on our sample mean, we are 95% confident that the true population mean is between 66.48 and 113.52.

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean; 12 is the standard error of the mean.

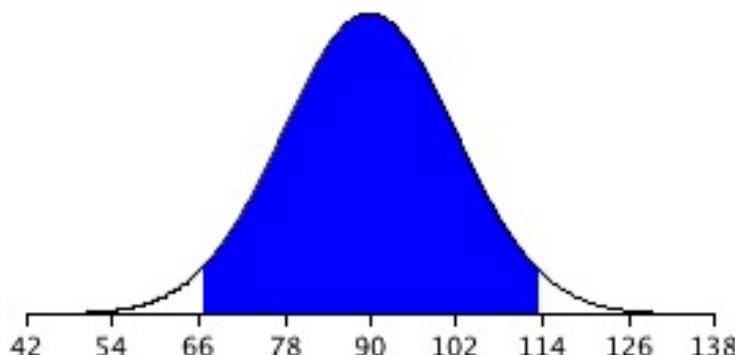


FIGURE 7.10

Figure 1. The sampling distribution of the mean for $n=9$. The middle 95% of the distribution is shaded.

Figure 1 shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the sample mean (\bar{x}) from a sample, and create an interval ranging from $\bar{x} - 23.52$ to $\bar{x} + 23.52$, this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

FIGURE 7.11

$$\text{Lower limit} = \bar{x} - (z_{.95}) (\sigma_x) \quad \text{and} \quad \text{Upper limit} = \bar{x} + (z_{.95}) (\sigma_x)$$

where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area and σ_x is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the population standard deviation (σ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when σ is known is easier than when σ has to be estimated, and serves a pedagogical purpose. Later we will show how to compute a confidence interval for the mean when σ has to be estimated using s , the sample standard deviation.

Suppose the following five numbers were sampled from a normal distribution with a known population standard deviation of 2.5: 2, 3, 5, 6, and 9. To compute the 95% confidence interval, start by computing the mean and standard error:

$$\mu = (2 + 3 + 5 + 6 + 9) / 5 = 5. \quad \text{The standard error is } \sigma_x = 2.5 / \sqrt{5} = 1.118.$$

$Z_{.95}$ can be found using the normal distribution calculator and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points. As shown in Figure 2, the value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been $Z_{.90} = 2.58$.

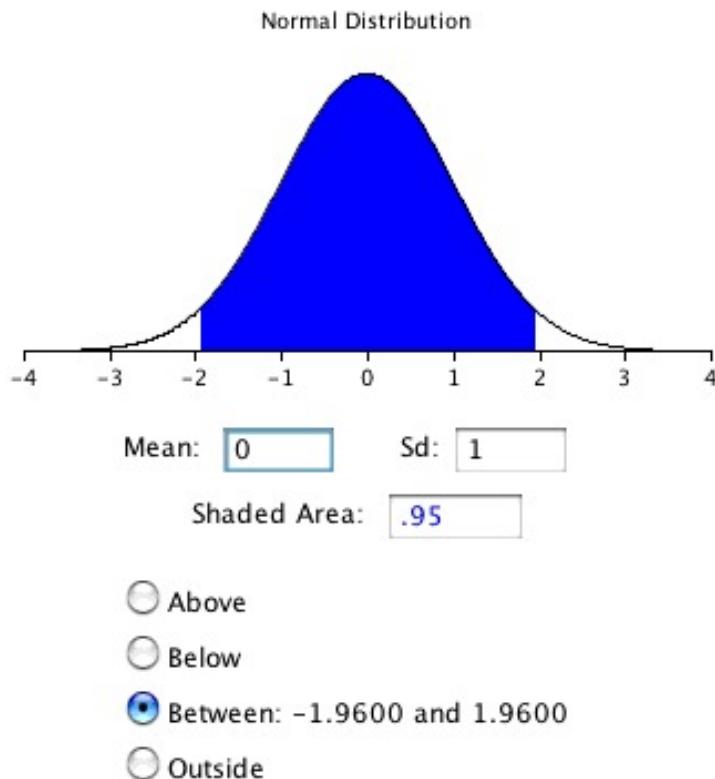


FIGURE 7.12

Figure 2. 95% of the area is between -1.96 and 1.96.

The confidence interval can then be computed as follows:

$$\text{Lower limit} = 5 - (1.96)(1.118) = 2.81 \quad \text{and} \quad \text{Upper limit} = 5 + (1.96)(1.118) = 7.19$$

The 95% confidence interval is: $2.81 \leq \mu \leq 7.19$

Example: Julianne collects four samples of size 60 from a known population with a population standard deviation of 19 and a population mean of 110. Using the four samples, she calculates the four sample means to be:

107 112 109 115

- a) For each sample, determine the 90% confidence interval.
- b) Do all four confidence intervals capture the population mean? Explain.

Answers:

Part a):

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$107 \pm (1.645) \left(\frac{19}{\sqrt{60}} \right)$$

$$107 \pm 4.04$$

$$\text{from } 102.96 \text{ to } 111.04$$

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$112 \pm (1.645) \left(\frac{19}{\sqrt{60}} \right)$$

$$112 \pm 4.04$$

$$\text{from } 107.96 \text{ to } 116.04$$

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$109 \pm (1.645) \left(\frac{19}{\sqrt{60}} \right)$$

$$109 \pm 4.04$$

$$\text{from } 104.96 \text{ to } 113.04$$

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$115 \pm (1.645) \left(\frac{19}{\sqrt{60}} \right)$$

$$115 \pm 4.04$$

$$\text{from } 110.96 \text{ to } 119.04$$

- b) Three of the confidence intervals enclose the population mean. The interval from 110.96 to 119.04 does not enclose the population mean.

In all of the examples shown above, you calculated the confidence intervals for the population mean using the formula $\bar{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$. However, to use this formula, the population standard deviation σ had to be known. If this value is unknown, and if the sample size is large ($n > 30$), the population standard deviation can be replaced with the sample standard deviation. Thus, the formula $\bar{x} \pm z_{\frac{\alpha}{2}} \left(\frac{s_x}{\sqrt{n}} \right)$ can be used as an interval estimator, or confidence interval. We will use this approximation of s for σ extensively in later chapters.

Here is a new name for the quantity $z_{\frac{\alpha}{2}} \left(\frac{s_x}{\sqrt{n}} \right)$. It is called the **margin of error**. Thus, a confidence interval can be thought of simply as: $\bar{x} \pm$ the margin of error.

Technically, this next example deals with proportions instead of sample means, but it is presented to you so that you can use the vocabulary of confidence intervals appropriately. There are no technical calculations required in this example, but you will learn calculations for confidence intervals about proportions just after the example.

Example: A committee set up to field-test questions from a provincial exam randomly selected grade 12 students to answer the test questions. The answers were graded, and the sample mean and sample standard deviation were calculated. Based on the results, the committee predicted that on the same exam, 9 times out of 10, grade 12 students would have an average score of within 3% of 65%.

- a) Are you dealing with a 90%, 95%, or 99% confidence level?
- b) What is the margin of error?
- c) Calculate the confidence interval.
- d) Explain the meaning of the confidence interval.

Answers:

- a) You are dealing with a 90% confidence level. This is indicated by 9 times out of 10.
- b) The margin of error is 3%.
- c) The confidence interval is $\bar{x} \pm$ the margin of error, or 62% to 68%.
- d) There is a 0.90 probability that the method used to produce this interval from 62% to 68% results in a confidence interval that encloses the population mean (the true score for this provincial exam).

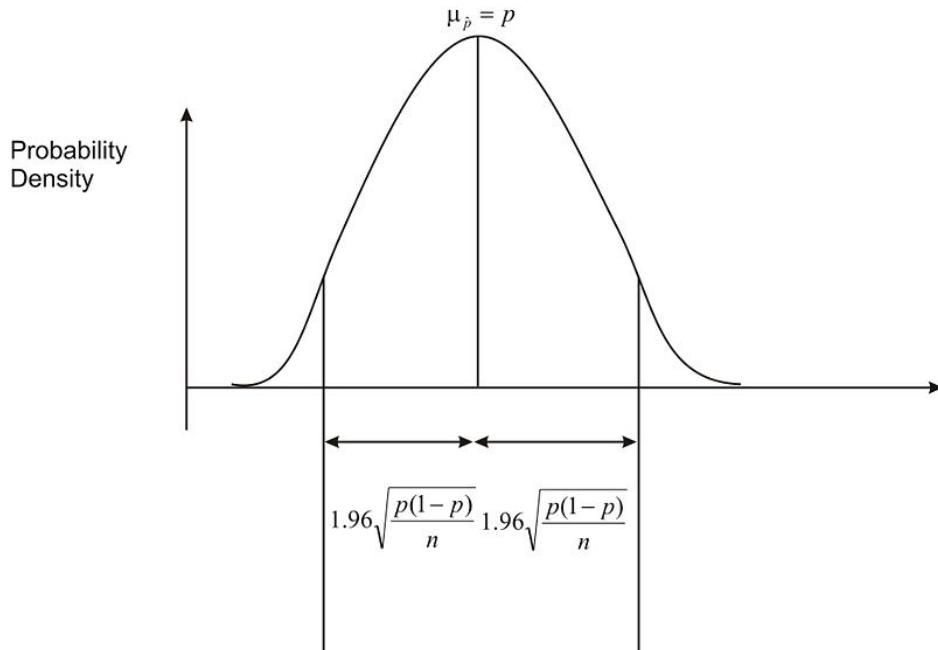
CONFIDENCE INTERVALS FOR POPULATION PROPORTIONS

In estimating a parameter, we can use a point estimate or an interval estimate. The **point estimate** for the population proportion, p , is \hat{p} . We can also find interval estimates for this parameter. These intervals are based on the sampling distributions of \hat{p} .

If we are interested in finding an interval estimate for the population proportion, the following two conditions must be satisfied:

- We must have a random sample.
- The sample size must be large enough ($n\hat{p} > 10$ and $n(1 - \hat{p}) > 10$) that we can use the normal distribution as an approximation to the binomial distribution.

$\sqrt{\frac{p(1-p)}{n}}$ is the standard deviation of the distribution of sample proportions. The distribution of sample proportions is as follows:



Since we do not know the value of p , we must replace it with \hat{p} . We then have the standard error of the sample proportions, $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. If we are interested in a 95% confidence interval, using the Empirical Rule, we are saying that we want the difference between the sample proportion and the population proportion to be within 1.96 standard deviations.

This is a 95% confidence interval for the population proportion. If we generalize for any confidence level, the confidence interval is as follows:

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In other words, the confidence interval is $\hat{p} \pm z_{\alpha/2} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$. Remember that $z_{\alpha/2}$ refers to the positive z -score for a particular confidence interval. Also, \hat{p} is the sample proportion, and n is the sample size. As before, the margin of

error is $z_{\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$, and the confidence interval is $\hat{p} \pm$ the margin of error.

Example: A congressman is trying to decide whether to vote for a bill that would legalize gay marriage. He will decide to vote for the bill only if 70 percent of his constituents favor the bill. In a survey of 300 randomly selected voters, 224 (74.6%) indicated they would favor the bill. The congressman decides that he wants an estimate of the proportion of voters in the population who are likely to favor the bill. Construct a confidence interval for this population proportion.

Answer: Our sample proportion is 0.746, and our standard error of the proportion is 0.0251. We will construct a 95% confidence interval for the population proportion. Under the normal curve, 95% of the area is between $z = -1.96$ and $z = 1.96$. Thus, the confidence interval for this proportion would be:

$$\begin{aligned} & 0.746 \pm (1.96)(0.0251) \\ & 0.697 < p < 0.795 \end{aligned}$$

With respect to the population proportion, we are 95% confident that the interval from 0.697 to 0.795 contains the population proportion. The population proportion is either in this interval, or it is not. When we say that this is a 95% confidence interval, we mean that if we took 100 samples, all of size n , and constructed 95% confidence intervals for each of these samples, about 95 out of the 100 confidence intervals we constructed would capture the population proportion, p .

It would appear that the congressman would be within his comfort zone in voting for the bill.

Example: A large grocery store has been recording data regarding the number of shoppers that use savings coupons at its outlet. Last year, it was reported that 77% of all shoppers used coupons, and 19 times out of 20, these results were considered to be accurate within 2.9%.

- a) Are you dealing with a 90%, 95%, or 99% confidence level?
- b) What is the margin of error?
- c) Calculate the confidence interval.
- d) Explain the meaning of the confidence interval.

Answers:

- a) The statement 19 times out of 20 indicates that you are dealing with a 95% confidence interval.
- b) The results were accurate within 2.9%, so the **margin of error** is 0.029.
- c) The confidence interval is simply $\hat{p} \pm$ the margin of error.

$$77\% - 2.9\% = 74.1\% \quad 77\% + 2.9\% = 79.9\%$$

Thus, the confidence interval is from 0.741 to 0.799.

- d) We are 95% confident that the true population proportion is between .741 and .799.

Lesson Summary

In this lesson, you learned that a sample mean is known as a **point estimate**, because this single number is used as a plausible value of the population mean. In addition to reporting a point estimate, you discovered how to calculate an interval of reasonable values based on the sample data. This interval estimator of the population mean is called

the confidence interval. You can calculate this interval for the population mean by using the formula $\bar{x} \pm z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right)$. The value of $z_{\frac{\alpha}{2}}$ is different for each confidence interval of 90%, 95%, and 99%.

In addition, you learned that you calculate the confidence interval for a population proportion by using the formula $\hat{p} \pm z_{\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$.

Concept Question

- Is there a way to increase the chance of capturing the unknown population mean?

Review Questions

1. In a local teaching district, a technology grant is available to teachers in order to install a cluster of four computers in their classrooms. From the 6,250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. Of those selected, 142 teachers felt that computers were an essential teaching tool.
 - a. Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential teaching tool.
 - b. How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?
2. Josie followed the guidelines presented to her and conducted a binomial experiment. She did 300 trials and reported a sample proportion of 0.61.
 - a. Calculate the 90%, 95%, and 99% confidence intervals for this sample.
 - b. What did you notice about the confidence intervals as the confidence level increased? Offer an explanation for your findings?
 - c. If the population proportion were 0.58, would all three confidence intervals enclose it? Explain.

Keywords

Central Limit Theorem

The distribution of the sample mean (or sample proportion) will approach a normal distribution when the sample size increases.

Confidence interval

Range of possible values the parameter might take.

Confidence level

The probability that the method used to calculate the confidence interval will produce an interval that will enclose the population parameter. Common confidence levels are 90%, 95%, and 99%, but any percent can be used if desired.

Margin of error

The amount that is added to and subtracted from the sample mean (or the sample proportion) to construct the confidence interval.

Parameter

Numerical descriptive measure of a population. Common parameters are μ , σ and p .

Point estimate

A single value that is an estimate of a population parameter. Examples are \bar{x} and \hat{p} .

Sample means

The sampling distribution of the *sample means* is approximately normal, as can be seen by the bell shape in each of the graphs.

Sample proportion

If a procedure gives 48 students who approve of the dress code and 52 who disapprove, the proportion who approve would be 48/100, or 0.48. This statistic is the *sample proportion*, and it is a point estimate.

Sampling distribution

The **sampling distribution** is the **probability distribution** of the statistic.

Standard error

The standard error is the standard deviation of a sample. As the sample size n increases, the standard error decreases.

7.4 References

1. CK-12 Foundation. . CCSA
2. CK-12 Foundation. . CCSA
3. CK-12 Foundation. . CCSA
4. CK-12 Foundation. . CCSA
5. CK-12 Foundation. . CCSA
6. CK-12 Foundation. . CCSA
7. CK-12 Foundation. . CCSA

CHAPTER

8

Hypothesis Testing

Chapter Outline

- 8.1 HYPOTHESIS TESTING AND THE P-VALUE**
 - 8.2 TESTING A PROPORTION HYPOTHESIS**
 - 8.3 TESTING A MEAN HYPOTHESIS**
 - 8.4 STUDENT'S T-DISTRIBUTION**
 - 8.5 TESTING A HYPOTHESIS FOR DEPENDENT AND INDEPENDENT SAMPLES**
 - 8.6 REFERENCES**
-

8.1 Hypothesis Testing and the P-Value

Learning Objectives

- Develop null and alternative hypotheses to test for a given situation.
- Determine the critical regions for one- and two-tailed hypothesis tests.
- Calculate a test statistic to evaluate a hypothesis.
- Test the probability of an event using the P -value.
- Identify type I and type II errors.

Introduction

In this chapter, we will explore hypothesis testing, which involves making conjectures about a population based on a sample drawn from the population. Hypothesis tests are often used in statistics to analyze the likelihood that a population has certain characteristics. For example, we can use hypothesis testing to analyze if a senior class has a particular average SAT score or if a prescription drug has a certain proportion of the active ingredient.

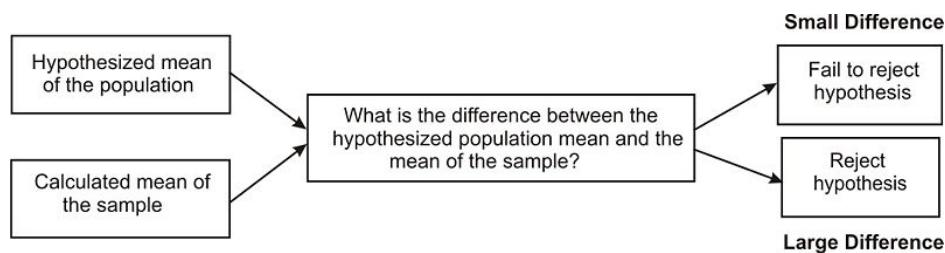
A hypothesis is simply a conjecture about a characteristic or set of facts. When performing statistical analyses, our hypotheses provide the general framework of what we are testing and how to perform the test.

These tests are never certain, and we can never prove or disprove hypotheses with statistics, but the outcomes of these tests provide information that either helps support or refute the hypothesis itself.

In this section, we will learn about different hypothesis tests, how to develop hypotheses, how to calculate statistics to help support or refute the hypotheses, and how to better understand the errors associated with hypothesis testing.

Developing Null and Alternative Hypotheses

Hypothesis testing involves testing the difference between a hypothesized value of a population parameter and the estimate of that parameter, which is calculated from a sample. If the parameter of interest is the mean of the population in hypothesis testing, we are essentially determining the magnitude of the difference between the mean of the sample and the hypothesized mean of the population. If the difference is very large, we reject our hypothesis about the population. If the difference is very small, we do not. Below is an overview of this process.



In statistics, the hypothesis to be tested is called the **null hypothesis** and is given the symbol H_0 . The **alternative hypothesis** is given the symbol H_a .

The null hypothesis defines a specific value of the population parameter that is of interest. Therefore, the null hypothesis always includes the possibility of equality. Consider the following:

$$H_0 : \mu = 3.2$$

$$H_a : \mu \neq 3.2$$

In this situation, if our sample mean, \bar{x} , is very different from 3.2, we would reject H_0 . That is, we would reject H_0 if \bar{x} is much larger than 3.2 or much smaller than 3.2. This is called a *two-tailed test*. An \bar{x} that is very unlikely if H_0 is true is considered to be good evidence that the claim H_0 is not true.

Now consider $H_0 : \mu = 3.2$ and $H_a : \mu > 3.2$. In this situation, we would reject H_0 for very large values of \bar{x} . This is called a *one-tailed test*. If, for this test, our data gives $\bar{x} = 15$, it would be highly unlikely that finding an \bar{x} this different from 3.2 would occur by chance, so we would probably reject the null hypothesis in favor of the alternative hypothesis.

Example: If we were to test the hypothesis that the seniors had a mean SAT score of 1100, our null hypothesis would be that the SAT score would be equal to 1100, or:

$$H_0 : \mu = 1100$$

We test the null hypothesis against an alternative hypothesis, which, as previously stated, is given the symbol H_a and includes the outcomes not covered by the null hypothesis. Basically, the alternative hypothesis states that there is a (large) difference between the hypothesized population mean and the sample mean. The alternative hypothesis can be supported only by rejecting the null hypothesis. In our example above about the SAT scores of graduating seniors, our alternative hypothesis would state the opposite of the null hypothesis, or:

$$H_a : \mu \neq 1100$$

Let's take a look at examples and develop a few null and alternative hypotheses.

Example: We have a medicine that is being manufactured, and each pill is supposed to have 14 milligrams of the active ingredient. What are our null and alternative hypotheses?

Solution:

$$H_0 : \mu = 14$$

$$H_a : \mu \neq 14$$

Our null hypothesis states that the population has a mean equal to 14 milligrams. Our alternative hypothesis states that the population has a mean that differs from 14 milligrams. This is a two-tailed test.

Example: A school principal wants to test if it is true what teachers say—that high school juniors use the computer an average 3.2 hours a day. What are our null and alternative hypotheses?

$$H_0 : \mu = 3.2$$

$$H_a : \mu \neq 3.2$$

Our null hypothesis states that the population has a mean equal to 3.2 hours. Our alternative hypothesis states that the population has a mean that differs from 3.2 hours. This is also a two-tailed test.

Deciding Whether to Reject the Null Hypothesis: One-Tailed and Two-Tailed Hypothesis Tests

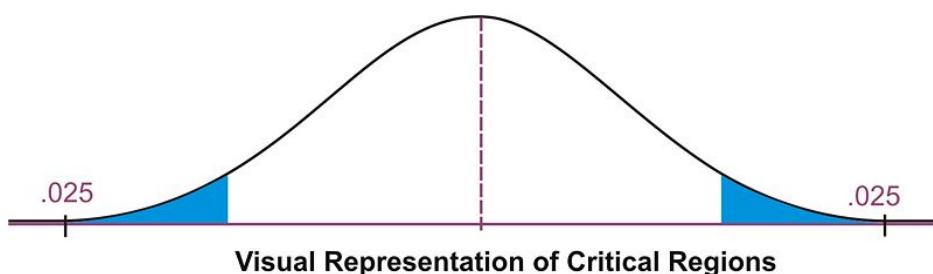
When a hypothesis is tested, a statistician must decide on how much evidence is necessary in order to reject the null hypothesis. For example, if the null hypothesis is that the average height of a population is 64 inches, a statistician wouldn't measure one person who is 66 inches and reject the hypothesis based on this one trial. It is too likely that the discrepancy was merely due to chance. We use a sample mean instead.

We use statistical tests to determine if the sample data give good evidence against the H_0 . The numerical measure that we use to determine the strength of the sample evidence we are willing to consider strong enough to reject H_0 is called the *level of significance*, and it is denoted by α . If we choose, for example, $\alpha = 0.01$, we are saying that the data we have collected would happen 1% or less of the time when H_0 is true. If our experimental outcome were to happen 1% of the time or less, compared to the value stated in the null hypothesis, we would decide that this is such an unusual outcome (very small probability) that it is very likely that the null hypothesis is false.

The most frequently used levels of significance are 0.05 and 0.01. If our data result in a statistic that falls within the region determined by the level of significance, then we reject H_0 . Therefore, the region is called the *critical region or the region of rejection*.

When determining the critical regions for a two-tailed hypothesis test, the level of significance represents the extreme areas under the normal density curve. The figure below illustrates a two-tailed hypothesis test, because the critical region is located in both ends (tails) of the distribution. For example, if there were a significance level of 0.05, the critical region would be the most extreme 5 percent under the curve, with 2.5 percent in each tail of the distribution.

Figure 1. Critical Region for .05 two-tailed test



Therefore, if the mean from the sample taken from the population falls in one of these critical regions, we would conclude that there was too much of a difference between our sample mean and the hypothesized population mean, and we would **reject** the null hypothesis. However, if the mean from the sample falls in the middle of the distribution (in between the critical regions), we would **fail to reject** the null hypothesis.

We calculate the critical region for a single-tail hypothesis test a bit differently. We would use a single-tail hypothesis test when the direction of the results is anticipated or we are only interested in one direction of the results. For example, a single-tail hypothesis test may be used when evaluating whether or not to adopt a new textbook. We would only decide to adopt the textbook if it improved student achievement relative to the old textbook. A single-tail alternative hypothesis simply states that the mean is greater or less than the hypothesized value.

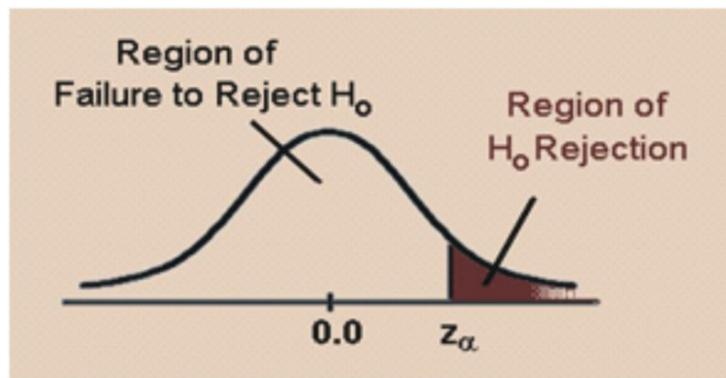
When performing a single-tail hypothesis test, our alternative hypothesis looks a bit different. When developing the alternative hypothesis in a single-tail hypothesis test, we would use the symbols for "greater than" or "less than." Using our example about SAT scores of graduating seniors, our null and alternative hypothesis would look something like:

$$H_0 : \mu = 1100$$

$$H_a : \mu > 1100$$

In this scenario, our null hypothesis states that the mean SAT score would be equal to 1100, while the alternative hypothesis states that the mean SAT score would be greater than 1100. A single-tail hypothesis test also means that

we have only one critical region, because we put the entire region of rejection into just one side of the distribution. When the alternative hypothesis is that the sample mean is "greater than," the critical region is on the right side of the distribution. When the alternative hypothesis is that the sample is "less than," the critical region is on the left side of the distribution (see below).



To calculate the critical regions, we must first find the cut-offs, or the ***critical values***, where the critical regions start. These values are specified by the z -distribution and can be found in a table that lists the areas of each of the tails under a normal distribution. Using this table, we find that for a 0.05 significance level, our critical values would fall at 1.96 standard errors above and below the mean. For a 0.01 significance level, our critical values would fall at 2.57 standard errors above and below the mean. Using the z -distribution, we can find critical values (as specified by standard z -scores) for any level of significance for either single-tailed or two-tailed hypothesis tests.

Example: Determine the critical value for a single-tailed hypothesis test with a 0.05 significance level.

Using the z -distribution table, we find that a significance level of 0.05 corresponds with a critical value of 1.645. If our alternative hypothesis is that the mean is greater than a specified value, the critical value would be 1.645. Due to the symmetry of the normal distribution, if the alternative hypothesis is that the mean is less than a specified value, the critical value would be -1.645 . The critical region would be all values to the left of (less than) -1.645 .

Technology Note: Finding Critical z -Values on the TI-83/84 Calculator

You can also find this critical value using the TI-83/84 calculator as follows: Press [2ND][DISTR], choose 'invNorm(', enter 0.05, 0, and 1, separated by commas, and press [ENTER]. This returns -1.64485 . The syntax for the 'invNorm(' command is 'invNorm (area to the left, mean, standard deviation)'.

Calculating the Test Statistic

Before evaluating our hypotheses by determining the critical region and calculating the *test statistic*, we need to confirm that the distribution is normal and determine the hypothesized mean, μ , of the distribution.

To evaluate the sample mean against the hypothesized population mean, we use the concept of z -scores to determine how different the two means are from each other. Based on the Central Limit Theorem, the sampling distribution of \bar{x} is normal, with mean, μ , and standard deviation, $\frac{\sigma}{\sqrt{n}}$. As we learned in previous lessons, the z -score is calculated by using the following formula:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where:

z is the standardized score.

\bar{x} is the sample mean.

μ is the population mean under the null hypothesis.

σ is the population standard deviation.

n is the sample size

If we do not have the population standard deviation, and if $n \geq 30$, we can use the sample standard deviation, s . If $n < 30$ and we do not have the population sample standard deviation, we use a different distribution, which will be discussed in a future lesson.

Once we calculate the z -score, we can make a decision about whether to reject or to fail to reject the null hypothesis based on the critical values.

The following are the steps you must take when doing a hypothesis test:

- Determine the null and alternative hypotheses.
- Verify that the necessary conditions are satisfied, and choose the α level.
- Set the criteria for rejecting the null hypothesis (critical region).
- Compute the test statistic.
- Make a decision (reject or fail to reject the null hypothesis).
- Interpret the decision in the context of the problem.

Example: College A has an average SAT score of 1500. From a random sample of 125 freshman psychology students, we find the average SAT score to be 1450, with a standard deviation of 100. We want to know if these freshman psychology students are representative of the overall population in terms of SAT scores. What are our hypotheses and test statistic?

Answer:

- Let's first develop our null and alternative hypotheses:

$$\begin{aligned} H_0 : \mu &= 1500 \\ H_a : \mu &\neq 1500 \end{aligned}$$

- We choose $\alpha = 0.05$.

3. This is a two-tailed test. If we choose $\alpha = 0.05$, the critical values will be -1.96 and 1.96 . (Use 'invNorm(0.025,0,1)' and the symmetry of the normal distribution to determine these critical values.) That is, we will reject the null hypothesis if the value of our test statistic is less than -1.96 or greater than 1.96 .

4. The test statistic is $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1450-1500}{\frac{100}{\sqrt{125}}} \approx -5.59$.

5. The value of the test statistic is -5.59 . This is less than -1.96 , so our decision is to reject H_0 .

6. Based on this sample, we believe that freshman psychology students' SAT scores are not representative of the overall SAT score for the college.

Example: A farmer is trying out a planting technique that he hopes will increase the yield of his pea plants. Over the last 5 years, the average number of pods on one of his pea plants was 145 pods, with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of 144 of his plants and finds the average number of pods to be 147. He wonders whether or not this is a statistically significant increase. What are his hypotheses and test statistic?

- First, we develop our null and alternative hypotheses:

$$H_0 : \mu = 145$$

$$H_a : \mu > 145$$

This alternative hypothesis uses the ' $>$ ' symbol, since the farmer believes that there might be a gain in the number of pods.

2. Now we choose $\alpha = 0.05$

3. The critical value will be 1.645. (Use 'invNorm(0.95,0,1)' to determine this critical value.) We will reject the null hypothesis if the test statistic is greater than 1.645. The value of the test statistic is 0.24.

4. Next, we calculate the test statistic for the sample of pea plants:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{147 - 145}{\frac{100}{\sqrt{144}}} \approx 0.24$$

5. The test statistic is less than 1.645, so our decision is to fail to reject H_0 .

6. Based on our sample, we believe the mean number of pea pods on each plant has not significantly increased as a result of his new planting technique.

Exercise: Try this problem, using the 6 steps of hypothesis testing, as outlined above.

A manufacturer states that the weight of Yummy candy bars is 2.5 ounces, with a standard deviation of .1 ounce. A group of students collects 16 candy bars and weighs each bar. The mean of their sample is 2.45 ounces. Is their evidence that the manufacturer is deceiving the public by claiming that the candy bar weighs 2.5 ounces? Use a significance level alpha of .05.

Answer:

1. The null hypothesis is $H_0: \mu = 2.5$ and the alternative hypothesis is $H_a: \mu \neq 2.5$. We chose a 2-sided alternative hypothesis because we have no information about whether the manufacturing process might produce a heavier candy bar or a lighter candy bar than what is stated.

2. Our significance level, also designated as $\alpha = .05$ must be split into 2 pieces, with half in the left tail of the z curve and half in the right tail. (see Figure 1).

3. Use a calculator or Normal Table to determine that the critical value is $z = \pm 1.96$. This means that we will reject the null hypothesis if our test statistic is greater than 1.96 or if it is less than -1.96. The calculated test statistic is -2.

4. We calculate the test statistic $z = \left[\frac{2.45 - 2.5}{0.1 / \sqrt{16}} \right] = \frac{-0.05}{0.025} = -2$

5. The test statistic is less than -1.96, and so we reject the null hypothesis.

6. Based on our sample, the mean weight of Yummy candy bars is not what the manufacturer stated in the null hypothesis. It is likely that the manufacturer is overstating the weight of Yummy candy bars.

Finding the P-Value of an Event

We can also evaluate a hypothesis by asking, "What is the probability of obtaining the value of the test statistic that we did if the null hypothesis is true?" This is called the *P-value*.

Example: Let's use the example of the pea farmer. As we mentioned, the farmer is wondering if the number of pea pods per plant has gone up with his new planting technique and finds that out of a sample of 144 peas, there is an average number of 147 pods per plant (compared to the historical average of 145 pods). To determine the *P-value*,

we ask, "What is $P(z > 0.24)$?" That is, what is the probability of obtaining a z -score greater than 0.24 if the null hypothesis is true?

Answer: Using the 'normcdf(0.24,99999999,0,1)' command on a graphing calculator, or using a table of z -values, we find this probability to be 0.41. This indicates that there is a 41% chance that under the null hypothesis, the peas will produce more than 147 pods. Since the P -value is greater than α , we fail to reject the null hypothesis. This is the same conclusion that was reached earlier.

Example: Suppose your hypotheses are $H_0: \mu = 6$ and $H_a: \mu > 6$. Your calculated test statistic is $z = 2.1$. You are told that your significance level α is .05. Determine the p -value.

Answer: When you are told to calculate the p -value, it is not necessary to find a critical value. Simply use the value of your test statistic, which is $z = 2.1$, and find the area (or probability) to its right. Using a calculator or Normal Table, the p -value is calculated as .0179. Because the p -value is **less than** α of .05, you would **reject** the null hypothesis.

Example: This is the same situation as the previous example, except that a 2-sided alternative hypothesis is used, and we calculate the p -value a bit differently. Suppose your hypotheses are $H_0: \mu = 6$ and $H_a: \mu \neq 6$. Your calculated test statistic is $z = 2.1$. You are told that your significance level α is .05. Determine the p -value.

Answer: We find the area (or probability) to the right of the test statistic, which we previously found as .0179. But because the alternative hypothesis was 2-sided, we need to double the .0179, taking into account that the alternative hypothesis said that we might find our test statistic **either** in the right tail **or** in the left tail. Thus, for a 2-sided alternative, we double the original value. The p -value for this 2-sided test is .0358. This value is still less than .05, so we would still reject the null hypothesis.

(Note: When you use a calculator to perform the hypothesis test, the p -value is automatically doubled by the calculator.)

Type I and Type II Errors

When we decide to reject or not to reject the null hypothesis, we have four possible scenarios:

- The null hypothesis is true, and we reject it.
- The null hypothesis is true, and we do not reject it.
- The null hypothesis is false, and we do not reject it.
- The null hypothesis is false, and we reject it.

Two of these four possible scenarios lead to correct decisions: not rejecting the null hypothesis when it is true and rejecting the null hypothesis when it is false.

Two of these four possible scenarios lead to errors: rejecting the null hypothesis when it is true and not rejecting the null hypothesis when it is false.

Which type of error is more serious depends on the specific research situation, but ideally, both types of errors should be minimized during the analysis.

TABLE 8.1: Below is a table outlining the possible outcomes in hypothesis testing:

	H_0 is true	H_0 is false
Not Reject H_0	Good Decision	Error (type II)
Reject H_0	Error (type I)	Good Decision

The general approach to hypothesis testing focuses on the *type I error*: rejecting the null hypothesis when it is true. The level of significance, also known as the *alpha level*, is defined as the probability of making a type I error when

testing a null hypothesis. For example, at the 0.05 level, we know that the decision to reject the hypothesis may be incorrect 5 percent of the time.

$$\alpha = P(\text{rejecting } H_0 | H_0 \text{ is true}) = P(\text{making a type I error})$$

Calculating the probability of making a *type II error* is not as straightforward as calculating the probability of making a type I error. The probability of making a type II error can only be determined when values have been specified for the alternative hypothesis. The probability of making a type II error is denoted by β .

$$\beta = P(\text{not rejecting } H_0 | H_0 \text{ is false}) = P(\text{making a type II error})$$

Once the value for the alternative hypothesis has been specified, it is possible to determine the probability of making a correct decision, which is $1 - \beta$. This quantity, $1 - \beta$, is called the **power of a test**.

The goal in hypothesis testing is to minimize the potential of both type I and type II errors. However, there is a relationship between these two types of errors. As the level of significance, or alpha level, increases, the probability of making a type II error (β) decreases, and vice versa.

Often we establish the alpha level based on the severity of the consequences of making a type I error. If the consequences are not that serious, we could set an alpha level at 0.10 or 0.20. However, in a field like medical research, we would set the alpha level very low (at 0.001, for example) if there was potential bodily harm to patients. We can also attempt minimize the type II errors by setting higher alpha levels in situations that do not have grave or costly consequences.

Lesson Summary

Hypothesis testing involves making a conjecture about a population based on a sample drawn from the population. We establish critical regions based on level of significance, or α level. If the value of the test statistic falls in one of these critical regions, we make the decision to reject the null hypothesis.

To evaluate the sample mean against the hypothesized population mean, we use the concept of z -scores to determine how different the two means are from each other.

When we make a decision about a hypothesis, there are four different possible outcomes and two different types of errors. A type I error is when we reject the null hypothesis when it is true, and a type II error is when we do not reject the null hypothesis, even when it is false. The level of significance of the test, α , is the probability of rejecting the null hypothesis when, in fact, it is true (an error).

The power of a test is defined as the probability of rejecting the null hypothesis when it is false (in other words, making the correct decision).

Review Questions

1. If the difference between the hypothesized population mean and the mean of a sample is large, we ___ the null hypothesis. If the difference between the hypothesized population mean and the mean of a sample is small, we ___ the null hypothesis.
2. At the Chrysler manufacturing plant, there is a part that is supposed to weigh precisely 19 pounds. The engineers take a sample of the parts and want to know if they meet the weight specifications. What are our null and alternative hypotheses?

3. In a hypothesis test, if the difference between the sample mean and the hypothesized mean divided by the standard error falls in the middle of the distribution and in-between the critical values, we ___ the null hypothesis. If this number falls in the critical regions and beyond the critical values, we ___ the null hypothesis.
4. Use a z -distribution table to determine the critical value for a single-tailed hypothesis test with a 0.01 significance level.
5. Sacramento County high school seniors have an average SAT score of 1020. From a random sample of 144 Sacramento high school students, we find the average SAT score to be 1100 with a standard deviation of 144. We want to know if these high school students are representative of the overall population. What are our hypotheses and test statistic?
6. During hypothesis testing, we use the P -value to predict the ___ of an event occurring.
7. The IQ scores of a standardized IQ test for children are normally-distributed and are known to have a mean of 100 and a standard deviation of 15. A group of 16 children who were breast-fed as babies are administered this IQ test. The sample mean of the group is 108.
 - (a) State the hypotheses for this study.
 - (b) Calculate the value of the test statistic.
 - (c) Using a level of significance of .05, determine the critical value(s).
 - (d) Sketch the critical region (region of rejection) and identify the location of the test statistic on the sketch.
 - (e) State the statistical decision.
 - (f) State your conclusion.
 - (g) Calculate the p -value for this hypothesis test.
8. Fill in the types of errors missing from the table below:

TABLE 8.2:

Decision Made	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	(1) ___	Correct Decision
Do not Reject Null Hypothesis	Correct Decision	(2) ___

Fill in the name and the symbol for blanks (1) and (2) in the chart.

Answers: (1) Reject; do not reject (2) $H_0: \mu = 19$ $H_a: \mu \neq 19$ (3) do not reject; reject (4) $z = 2.33$ or $z = -2.33$ (5) $H_0: \mu = 1020$ $H_a: \mu \neq 1020$; test statistic $z = 6.67$. (6) probability (7)(a) $H_0: \mu = 100$ and $H_a: \mu \neq 100$ (7)(b) $z = 2.13$ (7)(c) -1.96 and +1.96 (7)(d) see detailed solution (7)(e) reject H_0 (7)(f) see detailed solution (7)(g) .0332 (8)(1) α Type I error (2) β Type II error

8.2 Testing a Proportion Hypothesis

Learning Objectives

- Test a hypothesis about a population proportion
- Test a hypothesis about a population proportion using the P -value.

Introduction

In the previous section, we studied the test statistic that is used when you are testing hypotheses about the mean of a population and you have a large sample ($n > 30$).

In addition to the mean, statisticians are often interested in making inferences about a population proportion. For example, when we look at election results, we often look at the proportion of people who vote and who these voters choose. Typically, we call these proportions percentages, and we would say something like, “Approximately 68 percent of the population voted in this election, and 48 percent of these voters voted for Barack Obama.”

So how do we test hypotheses about proportions? We use the same process as we did when testing hypotheses about means, but we must include sample proportions as part of the analysis. This lesson will address how we investigate hypotheses around population proportions and how to construct confidence intervals around our results.

Hypothesis Testing about Population Proportions

We could perform tests of population proportions to answer the following questions:

- What percentage of graduating seniors will attend a 4-year college?
- What proportion of voters will vote for John McCain?
- What percentage of people will choose Diet Pepsi over Diet Coke?

To test questions like these, we make hypotheses about population proportions. For example, here are some hypotheses we could make:

H_0 : 35% of graduating seniors will attend a 4-year college.

H_0 : 42% of voters will vote for John McCain.

H_0 : 26% of people will choose Diet Pepsi over Diet Coke.

To test these hypotheses, we might design an experiment or study, as follows:

- Hypothesize a value for the population proportion, p , like we did above.
- Randomly select a sample.
- Use the sample proportion, \hat{p} , to test the stated hypothesis.

To determine the test statistic, we need to know the sampling distribution of the sample proportions. We use the binomial distribution, which is appropriate for situations in which two outcomes are possible (for example, voting for a candidate and not voting for a candidate). Therefore, the test statistic can be calculated as follows:

$$z = \frac{\text{sample estimate} - \text{value under the null hypothesis}}{\text{standard error under the null hypothesis}}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

\hat{p} is the sample proportion.

p_0 is the hypothesized value of the proportion under the null hypothesis.

n is the sample size.

Example: We want to test a hypothesis that 60 percent of the 400 seniors graduating from a certain California high school will enroll in a two- or four-year college upon graduation. What would be our hypotheses and test statistic?

Since we want to test the proportion of graduating seniors, and we think that proportion is around 60 percent, our hypotheses are:

$$H_0 : p = 0.6$$

$$H_a : p \neq 0.6$$

Also, the test statistic would be $z = \frac{\hat{p}-0.6}{\sqrt{\frac{0.6(1-0.6)}{400}}}$. To complete this calculation, we would have to have a value for the sample proportion.

Testing a Proportion Hypothesis

Similar to testing hypotheses dealing with population means, we use a similar set of steps when testing proportion hypotheses.

- Determine and state the null and alternative hypotheses.
- Determine the significance level α .
- Determine the critical value and the critical region(s).
- Calculate the value of the test statistic.
- Decide whether to reject or fail to reject the null hypothesis.
- Interpret the decision within the context of the problem.

Example: A congressman is trying to decide on whether to vote for a bill that would legalize gay marriage. He will decide to vote for the bill only if more than 70 percent of his constituents favor the bill. In a survey of 300 randomly selected voters, 224 (74.6%) indicated that they would favor the bill. Should he or should he not vote for the bill?

First, we develop our null and alternative hypotheses:

$$H_0 : p = 0.7$$

$$H_a : p > 0.7$$

Next, we set the criterion for rejecting the null hypothesis. Choose $\alpha = 0.05$, and since the alternative hypothesis is $p > 0.7$, make this a one-tailed test. Using a standard z -table or the TI-83/84 calculator, we find the critical value for

a one-tailed test at an alpha level of 0.05 to be 1.645. Thus, the critical region is a z-value of 1.645 or greater. A sketch can be useful.

Next, we calculate the value of the test statistic:

$$\text{The test statistic is } z = \frac{0.74 - 0.7}{\sqrt{\frac{(0.7)(1-0.7)}{300}}} \approx 1.51.$$

Since our critical value is 1.645 and our test statistic is 1.51, we cannot reject the null hypothesis. This means that we cannot statistically conclude that the population proportion is greater than 0.70. In other words, given this information, it is not safe to conclude that at least 70 percent of the voters would favor this bill with any degree of certainty. Even though the sample proportion of voters supporting the bill is over 70 percent, this could be due to chance and is not statistically significant.

We can calculate the *p*-value for this hypothesis test. We use the test statistic $z = 1.51$ and determine the area (probability) to its right. Using the calculator or table of *z*-values, we find that this area (probability) is $1 - .9345 = .0655$. With a one-sided alternative, we do not double the value, and so we state that our *p*-value is .0655. Because this value is large (greater than the significance level of .05), we see that we cannot reject the null hypothesis.

Example: Admission staff from a local university are conducting a survey to determine the proportion of incoming freshman who will need financial aid. A survey on housing needs, financial aid, and academic interests is collected from 400 of the incoming freshman. Staff hypothesized that 30 percent of freshman will need financial aid, and the sample from the survey indicated that 101 (25.3%) would need financial aid. Is 30 percent an accurate guess?

First, we develop our null and alternative hypotheses:

$$H_0 : p = 0.3$$

$$H_a : p \neq 0.3$$

Next, we set the criterion for rejecting the null hypothesis. The 0.05 alpha level is used, and for a two-tailed test, the critical values of the test statistic are 1.96 and -1.96 .

The test statistic can be calculated as follows:

$$z = \frac{0.253 - 0.3}{\sqrt{\frac{0.3(1-0.3)}{400}}} \approx -2.05$$

Since our critical values are ± 1.96 , and since $-2.05 < -1.96$, we can reject the null hypothesis. This means that we can conclude that the population of freshman needing financial aid is significantly more than 30 percent or less than 30 percent. Since the test statistic is negative, we can say that in the population of incoming freshman, less than 30 percent of the students will need financial aid.

We can calculate the *p*-value for this hypothesis test. We use the test statistic $z = -2.05$ and determine the area (probability) to its left. Using the calculator or table of *z*-values, we find that this area (probability) is .0202. To get the *p*-value, we double this value (because the alternative is two-sided) and obtain the *p*-value of .0404. Because this value is small (less than the significance level of .05), we see that the null hypothesis is rejected.

Lesson Summary

In statistics, we also make inferences about proportions of a population. We use the same process as in testing hypotheses about means of populations, but we must include hypotheses about proportions and the proportions of the sample in the analysis. To calculate the test statistic needed to evaluate the population proportion hypothesis, we must also calculate the standard error of the proportion, which is defined as $s_p = \sqrt{\frac{p_0(1-p_0)}{n}}$.

The formula for calculating the test statistic for a population proportion is as follows:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

\hat{p} is the sample proportion.

p_0 is the hypothesized population proportion.

n is the sample size.

We establish critical regions based on level of significance, or α level. If the value of the test statistic falls in one of these critical regions, we make the decision to reject the null hypothesis.

Review Questions

1. The test statistic helps us determine ____.
 2. True or false: In statistics, we are able to study and make inferences about proportions, or percentages, of a population.
 3. A state senator cannot decide how to vote on an environmental protection bill. The senator decides to request her own survey, and if the proportion of registered voters supporting the bill exceeds 0.60, she will vote for it. A random sample of 750 voters is selected, and 495 are found to support the bill.
 - a. What are the null and alternative hypotheses for this problem?
 - b. What is the observed value of the sample proportion?
 - c. What is the standard error of the proportion?
 - d. What is the test statistic for this scenario?
 - e. What decision would you make about the null hypothesis if you had an alpha level of 0.01?
 - f. What is the p -value for this problem?
-

Answers: (1) how unusual the experimental outcome is. (2) True (3)(a) $H_0: p = .60$ and $H_a: p > .60$ (3)(b) .66 (3)(c) .018 (3)(d) 3.33 (3)(e) reject H_0 (3)(f) p -value is .0005

8.3 Testing a Mean Hypothesis

Learning Outcome

- Use the 6-step hypothesis-testing procedure to test the mean for large samples.

Evaluating Hypotheses for Population Means using Large Samples

When testing a hypothesis for the mean of a normal distribution, we follow a series of six basic steps:

- a. State the null and alternative hypotheses.
- b. Choose an α level.
- c. Set the criterion (critical values) for rejecting the null hypothesis.
- d. Compute the test statistic.
- e. Make a decision (reject or fail to reject the null hypothesis).
- f. Interpret the result.

If we reject the null hypothesis, we are saying that the difference between the observed sample mean and the hypothesized population mean is too great to be attributed to chance. When we fail to reject the null hypothesis, we are saying that the difference between the observed sample mean and the hypothesized population mean is probable if the null hypothesis is true. Essentially, we are willing to attribute this difference to sampling error.

Example: The school nurse was wondering if the average height of 7th graders has been increasing. Over the last 5 years, the average height of a 7th grader was 145 cm, with a standard deviation of 20 cm. The school nurse takes a random sample of 200 students and finds that the average height this year is 147 cm. Conduct a single-tailed hypothesis test using a 0.05 significance level to evaluate the null hypothesis.

First, we develop our null and alternative hypotheses:

$$\begin{aligned}H_0 &: \mu = 145 \\H_a &: \mu > 145\end{aligned}$$

Next, we choose $\alpha = 0.05$. The critical value for this one-tailed test is 1.64. Therefore, any test statistic greater than 1.64 will be in the rejection region.

Finally, we calculate the test statistic for the sample of 7th graders as follows:

$$z = \frac{147 - 145}{\sqrt{\frac{20}{200}}} \approx 1.414$$

Since the calculated z -score of 1.414 is smaller than 1.64, it does not fall in the critical region. Thus, our decision is to fail to reject the null hypothesis and to conclude that the probability of obtaining a sample mean equal to 147 if the mean of the population is 145 is likely to have been due to chance. There is not compelling evidence that the average height of 7th graders has increased.

Here is a sketch of the test statistic, the critical value, and the region of rejection for the height hypothesis test. We see that the test statistic is less than 1.64, and so it is not in the region of rejection. Therefore we do not reject the null hypothesis.

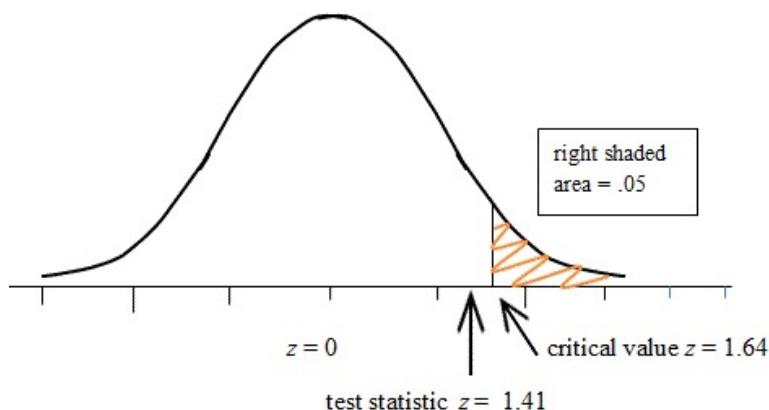


FIGURE 8.1

Figure 1. Sketch of the Critical Value and Test Statistic for Height Hypothesis Test

Again, when testing a hypothesis for the mean of a distribution, follow the six basic steps. Commit these steps to memory.

Review Questions

- In hypothesis testing, when we work with large samples, we use the ___ distribution. When working with small samples (typically samples under 30), we use the ___ distribution.
- The dean from UCLA is concerned that the students' grade point averages have changed dramatically in recent years. The graduating seniors' mean GPA over the last five years is 2.75. The dean randomly samples 256 seniors from the last graduating class and finds that their mean GPA is 2.85, with a sample standard deviation of 0.65.
 - What would the null and alternative hypotheses be for this scenario?
 - What would the standard error be for this particular scenario?
 - Describe in your own words how you would set the critical regions and what they would be at an alpha level of 0.05.
 - Test the null hypothesis and explain your decision.
- For each of the following pairs of scenarios, state which option is more likely to lead to the rejection of the null hypothesis.
 - A one-tailed or two-tailed test, each with level of significance of .05.
 - A 0.05 or 0.01 level of significance
 - A sample size of $n = 144$ or $n = 444$

Answers: (1) z distribution, t distribution (2)(a) $H_0: \mu = 2.75$ and $H_a: \mu \neq 2.75$ (2)(b) 0.041 (2)(c) -1.96 and +1.96 (2)(d) $z = 2.44$; reject H_0 (3)(a) one-tailed (3)(b) .05 (3)(c) $n = 444$

8.4 Student's t-Distribution

Learning Objectives

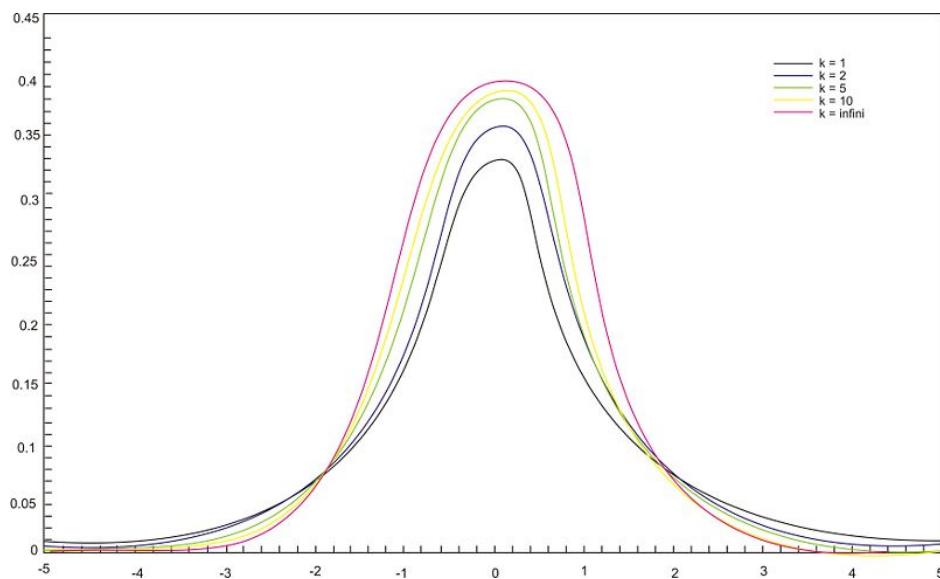
- Use Student's t -distribution to perform a small-sample hypothesis test about a population mean.
- Understand how the shape of Student's t -distribution corresponds to the sample size (which corresponds to a measure called the degrees of freedom).
- Use Student's t -distribution to calculate a confidence interval for the population mean.

Introduction

Back in the early 1900s, a chemist at a brewery in Ireland discovered that when he was working with very small samples, the distributions of the means differed significantly from the normal distribution. He noticed that as his sample sizes changed, the shape of the distribution changed as well. He published his results under the pseudonym 'Student', and this concept and the distributions for small sample sizes are now known as Student's t -distributions.

Hypothesis Testing with Small Sample Sizes

Student's t-distributions are a family of distributions that, like the normal distribution, are symmetrical, bell-shaped, and centered on a mean. However, the distribution shape changes as the sample size changes. Therefore, there is a specific shape, or distribution, for every sample of a given size (see figure below; each distribution has a different value of k , the number of *degrees of freedom*, which is 1 less than the size of the sample).



We use Student's t -distributions in hypothesis testing the same way that we use the normal distribution. Each row in the t -distribution table represents a different t -distribution, and each distribution is associated with a unique number of degrees of freedom (the number of observations minus one). The column headings in the table represent the portion of the area in the tails of the distribution. We use the numbers in the table just as we use z -scores.

As the number of observations gets larger, the t -distribution approaches the shape of the normal distribution. In general, once the sample size is large enough—usually about 120—we would use the normal distribution or a z -table instead.

In calculating the t -test statistic, we use the following formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where:

t is the test statistic and has $n - 1$ degrees of freedom.

\bar{x} is the sample mean.

μ_0 is the population mean under the null hypothesis.

s is the sample standard deviation.

n is the sample size.

$\frac{s}{\sqrt{n}}$ is the estimated standard error.

Example: A high school athletic director is asked if football players are doing as well academically as the other student athletes at his school. We know from a previous study that the average GPA for the student athletes is 3.10 and that the standard deviation of the sample is 0.54. After an initiative to help improve the GPA of student athletes, the athletic director samples 20 football players and finds that their GPA is 3.18. Is there a significant improvement? Use a 0.05 significance level.

Answer:

First, we establish our null and alternative hypotheses:

$$\begin{aligned} H_0 : \mu &= 3.10 \\ H_a : \mu &\neq 3.10 \end{aligned}$$

Next, we use our alpha level of 0.05 and the t -distribution table to find our critical values. For a two-tailed test with 19 degrees of freedom and a 0.05 level of significance, our critical values are equal to ± 2.093 .

Finally, in calculating the test statistic, we use the formula as shown:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.18 - 3.10}{\frac{0.54}{\sqrt{20}}} \approx 0.66$$

This means that the observed sample mean of the GPA of football players of 3.18 is 0.66 standard errors above the hypothesized value of 3.10. Because the value of the test statistic is less than the critical value of 2.093, we fail to reject the null hypothesis.

Therefore, we can conclude that the difference between the sample mean and the hypothesized value is not sufficient to attribute it to anything other than sampling error. Thus, the athletic director can conclude that the mean academic performance of football players does not differ from the mean performance of other student athletes.

Example: The masses of newly-produced bus tokens are supposed to have a mean of 3.16 grams. A random sample of 11 tokens was removed from the production line, and the mean weight of the tokens was calculated to be 3.21 grams, with a standard deviation of 0.067. What is the value of the test statistic for a test to determine if the machine is producing bus tokens with the specified mass?

The test statistic for this problem can be calculated as follows:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

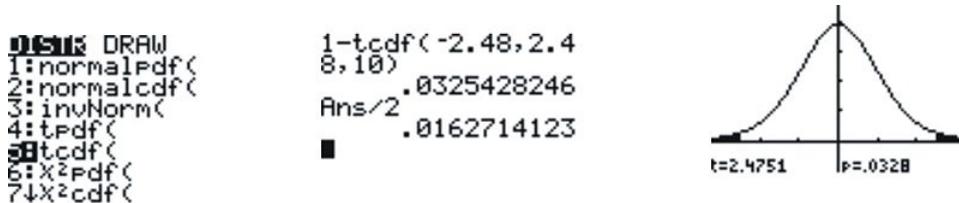
$$t = \frac{3.21 - 3.16}{\frac{0.067}{\sqrt{11}}}$$

$$t \approx 2.48$$

The P -value for a two-sided test is the area under the t -distribution with degrees of freedom of $11 - 1 = 10$ that lies above $t = 2.48$ and below $t = -2.48$. This P -value can be calculated by using technology.

Technology Note: Using the 'tcdf' Command on the TI-83/84 Calculator to Calculate Probabilities Associated with the t-Distribution

Press [2ND][DIST] and use the down arrow to select 'tcdf('. The syntax for this command is 'tcdf(lower bound, upper bound, degrees of freedom)'. This command will return the total area under both tails. To calculate the area under one tail, divide by 2 as shown below:



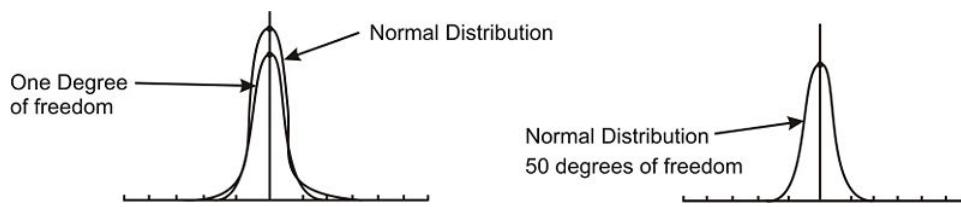
This means that there is only a 0.016 chance of getting a value of t as large as or even larger than the one from this sample. The small P -value tells us that the sample is inconsistent with the null hypothesis. There are problems with the production of the bus tokens; the sample shows that it is highly likely that the machine is producing tokens that are underweight.

When the P -value is small (less than .05 is usually considered "small"), there is strong evidence against the null hypothesis. On the other hand, when the P -value is large, the result from the sample is consistent with the estimated or hypothesized mean, and there is no evidence against the null hypothesis.

A visual picture of the P -value can be obtained by using a graphing calculator as follows:



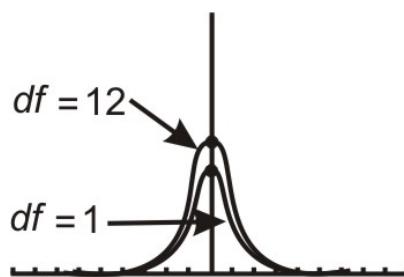
The spread of any t -distribution is greater than that of a standard normal distribution. This is due to the fact that in the denominator of the formula, σ has been replaced with s . Since s is a random quantity changing with various samples, the variability in t is greater, resulting in a larger spread.



Notice that in the first distribution graph shown above, the spread of the inner curve is small, but in the second graph, both distributions are basically overlapping and are roughly normal. This is due to the increase in the degrees of freedom.

To further illustrate this point, the t -distributions for 1 and 12 degrees of freedom can be graphed on a graphing calculator. To do so, first press **[Y=][2ND][DISTR]**, choose the 'tpdf' command, enter 'X' and 1, separated by commas, and close the parentheses. Then go down to **Y2** and repeat the process, this time entering 12 instead of 1. Finally, make sure your window is set correctly and press **[GRAPH]**.

The t -distributions for 1 and 12 degrees of freedom should look similar to the ones shown below (df denotes degrees of freedom):



Notice the difference in the two distributions. The one with 12 degrees of freedom approximates a normal curve.

The t -distribution can be used with any statistic having a bell-shaped distribution. We already know that the Central Limit Theorem states that the sampling distribution of a statistic will be close to normal with a large enough sample size, but, in fact, the Central Limit Theorem predicts a roughly normal distribution under any of the following conditions:

- The population distribution is normal.
- The sampling distribution is symmetric and the sample size is ≤ 15 .
- The sampling distribution is moderately skewed and the sample size is $16 \leq n \leq 30$.
- The sample size is greater than 30, without outliers.

In addition to the fact that the t -distribution can be used with any bell-shaped distribution, it also has some unique properties. These properties are as follows:

- The mean of the distribution equals zero.
- The population standard deviation is unknown.
- Although the t -distribution is bell-shaped, the smaller sample sizes produce a flatter curve. The distribution is not as mound-shaped as a normal distribution, and the tails are thicker. As the sample size increases and approaches 30, the distribution approaches a normal distribution.
- The population is unimodal and symmetric.

Example: Duracell manufactures batteries that the CEO claims will last 300 hours under normal use. A researcher randomly selected 15 batteries from the production line and tested these batteries. The tested batteries had a mean life span of 290 hours, with a standard deviation of 50 hours. If the CEO's claim were true, what is the probability that 15 randomly selected batteries would have a mean life span of no more than 290 hours?

Answer:

The null hypothesis is $H_0: \mu = 300$ and the alternative hypothesis is $H_a: \mu < 300$. This is a one-sided hypothesis test.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ There are } n - 1 = 15 - 1 = 14 \text{ degrees of freedom.}$$

$$t = \frac{290 - 300}{\frac{50}{\sqrt{15}}}$$

$$t = \frac{-10}{12.9099}$$

$$t = -0.7745967$$

Using a graphing calculator or a t -table, the p -value is shown to be 0.226, which means that if the true life span of a battery were 300 hours, there is a 22.6% chance that the mean life span of the 15 tested batteries would be less than or equal to 290 hours. This is not a high enough level of confidence to reject the null hypothesis and count the discrepancy as significant.

Technology: Using the TI-84 Plus Calculator to Perform a t-test

You can use your calculator to perform a one-sample t-test as follows: We will use the problem just completed as an example. Press [STAT] and scroll right to TESTS. Scroll down to Option 2 (T-Test) and press [ENTER] to get to the input screen. In the first line, scroll to "Stats" and press [ENTER]. The next line asks for the value of the hypothesized μ , which is 300. In the next line, enter the sample mean of 290. Enter the value of the standard deviation in the next line, which is 50, and then enter the sample size $n = 15$ in the next line. In the next line you have 3 choices for the alternative hypothesis. In our example it was a one-sided, less-than alternative, so scroll to $<\mu_0$ and press [ENTER]. Your input screen should look like this:

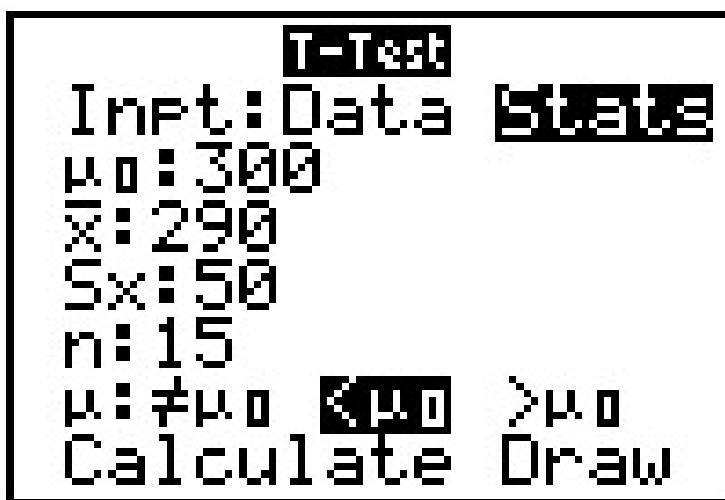
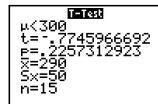


FIGURE 8.2

Now scroll down to Calculate and press [ENTER]. A new screen appears:



The output screen shows the identical results as the previous use of the formula did: The t test statistic is $-.775$ and the p -value is 0.2257 .

Example: You have just taken ownership of a pizza shop. The previous owner told you that you would save money if you bought the mozzarella cheese in a 4.5-pound slab. Each time you purchase a slab of cheese, you weigh it to ensure that you are receiving 72 ounces of cheese. The results of 7 random measurements are 70, 69, 73, 68, 71, 69 and 71 ounces, respectively. Find the test statistic for this scenario.

Begin the problem by determining the mean of the sample and the sample standard deviation. This can be done using a graphing calculator. You should find that $\bar{x} = 70.143$ and $s = 1.676$. Now calculate the test statistic as follows:

$$\begin{aligned}t &= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \\t &= \frac{70.143 - 72}{\frac{1.676}{\sqrt{7}}} \\t &\approx -2.9315\end{aligned}$$

Example: In the last example, the test statistic for testing that the mean weight of the cheese wasn't 72 ounces was computed. Find and interpret the P -value.

The test statistic computed in the last example was -2.9315 . Using technology or a t -table, the P -value is 0.0262 . In other words, the probability that 7 random measurements would give a value of t greater than 2.9315 or less than -2.9315 is about 0.0262 .

Example: In the previous example, the P -value for testing that the mean weight of cheese was 72 ounces was determined.

- State the hypotheses.
- Would the null hypothesis be rejected at the 10% level? The 5% level? The 1% level?

Answers:

a)

$$\begin{aligned}H_0 : \mu &= 72 \\H_a : \mu &\neq 72\end{aligned}$$

- b) Because the P -value of 0.0262 is less than both 0.10 and 0.05 , the null hypothesis would be rejected at these levels. However, the P -value is greater than 0.01 , so the null hypothesis would not be rejected if this level of confidence was required.

CONFIDENCE INTERVAL USING THE t DISTRIBUTION

Remember that a confidence interval gives us a range of values for the value of the unknown population mean. We earlier used the formula below for calculating a 95% confidence interval when we knew the population standard deviation.

FIGURE 8.3

$$\text{Lower limit} = \bar{x} - (z_{95}) (\sigma_x) \quad \text{and} \quad \text{Upper limit} = \bar{x} + (z_{95}) (\sigma_x)$$

You should use the t distribution rather than the normal distribution when the population standard deviation is not known and has to be estimated from sample data. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78 standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table 1. The first column, df, stands for degrees of freedom, and for confidence intervals on the mean, df is equal to $n - 1$, where n is the sample size.

Table 1. Abbreviated t table.

TABLE 8.3:

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Assume that the following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9 and that the standard deviation is not known. The first steps are to compute the sample mean and sample standard deviation: $\bar{x} = 5$ $s = 2.74$. The next step is to estimate the standard error of the mean. We don't know the value of σ , so we will use the sample standard deviation of s as its estimate. Therefore, the standard error will be $\frac{s}{\sqrt{n}} = \frac{2.74}{\sqrt{5}} = 1.225$.

The next step is to find the value of t . The number of degrees of freedom is $(n - 1) = (5 - 1) = 4$. As you can see from Table 1, the value for the 95% interval for $df = 4$ is 2.776. The confidence interval is then computed just as it was for a large sample, except that we are using the t table and we are estimating the unknown σ with s .

$$\text{Lower limit} = 5 - (2.776)(1.225) = 1.60 \quad \text{Upper limit} = 5 + (2.776)(1.225) = 8.40$$

More generally, the formula for the 95% confidence interval on the mean is:

FIGURE 8.4

$$\text{Lower limit} = \bar{x} - (t^*) \left(\frac{s}{\sqrt{n}} \right) \quad \text{Upper limit} = \bar{x} + (t^*) \left(\frac{s}{\sqrt{n}} \right)$$

where \bar{x} is the sample mean, t^* is the t for the confidence level desired (0.95 in the above example), and $\frac{s}{\sqrt{n}}$ is the estimated standard error of the mean.

Using the Calculator to Compute a t Confidence Interval

Let's use the previous example to illustrate the use of the calculator to obtain the 95% confidence interval for the data values 2, 3, 5, 6, 9. Press [STAT] and press [ENTER] to get to the List-making screen. Into List 1, enter the 5 data values, pressing [ENTER] after each data value is typed. Then press [STAT] and scroll right to TESTS. Scroll down to Option 8 (T Interval) and press [ENTER]. On the first line, scroll to Data and press [ENTER]. On line 2, press the blue [2nd] button on your calculator keypad and then the number [1] button. This results in L_1 appearing on the second line of the input screen. For Freq: enter 1, and for C-Level enter .95. Scroll down to Calculate. Your screen should look like this:

```
TInterval
Inpt:Data Stats
List:L1
Freq:1
C-Level:.95
Calculate
```

Now scroll down to Calculate and press [ENTER]. The new screen shows:

```
TIinterval
(1.5996, 8.4004)
x̄=5
Sx=2.738612788
n=5
```

FIGURE 8.5

The 2 numbers enclosed in parentheses are the lower and upper limits of the 95% confidence interval for the mean. These values agree with the previous manual calculations.

Lesson Summary

A test of significance is done when a claim is made about the value of a population parameter. The test can only be conducted if the random sample taken from the population came from a distribution that is normal or approximately normal. When the sample size is small, you must use t instead of z to complete the significance test for a mean.

Hypothesis testing with a small sample size uses the same steps as testing with a large sample. Only the formula has changed to reflect the fact that the population standard deviation is unknown and s is used as an estimate of σ .

Confidence intervals with small samples are similar to large-sample confidence intervals. Only the formula has changed to reflect the fact that the population standard deviation is unknown and s is used as an estimate of σ .

Points to Consider

- Is there a way to determine where the *t*-statistic lies on a distribution?
- If a way does exist, what is the meaning of its placement?

Review Questions

1. The dean of a university is concerned that the students' grade point averages have changed dramatically in recent years. The graduating seniors' mean GPA over the last five years is 2.75. The dean randomly samples 25 seniors from the last graduating class and finds that their mean GPA is 2.85, with a sample standard deviation of 0.65. Would a *t*-distribution now be the appropriate sampling distribution for the mean? Why or why not?
2. Using the appropriate *t*-distribution, test the same null hypothesis with a sample of 25.
3. With a sample size of 30, do you need to have a larger or smaller difference between the hypothesized population mean and the sample mean than with a sample size of 256 to obtain statistical significance? Explain your answer.
4. Calculate a 90% confidence interval for the mean systolic blood pressure for the following 6 readings

(a) 132 117 122 127 109 114

Answers: (1) see detailed answers (2) test statistic $z = .769$. Do not reject. (3) see detailed answers (4) $113.2 \leq \mu \leq 127.2$

8.5 Testing a Hypothesis for Dependent and Independent Samples

Learning Objectives

- Identify situations that contain dependent or independent samples.
- Calculate the test statistic to test hypotheses about dependent data pairs.
- Calculate the test statistic to test hypotheses about independent data pairs for both large and small samples.
- Calculate the test statistic to test hypotheses about the difference of proportions between two independent samples.

Introduction

In the previous lessons, we learned about hypothesis testing for proportions and means in large and small samples. However, in the examples in those lessons, only one sample was involved. In this lesson, we will apply the principles of hypothesis testing to situations involving two samples. There are many situations in everyday life where we would perform statistical analysis involving two samples. For example, suppose that we wanted to test a hypothesis about the effect of two medications on curing an illness. Or we may want to test the difference between the means of males and females on the SAT. In both of these cases, we would analyze both samples, and the hypothesis would address the difference between the two sample means.

In this lesson, we will identify situations with different types of samples, learn to calculate the test statistic, calculate the estimate for population variance for both samples, and calculate the test statistic to test hypotheses about the difference of proportions or means between samples.

Dependent and Independent Samples

When we are working with one sample, we know that we have to randomly select the sample from the population, measure that sample's statistics, and then make a hypothesis about the population based on that sample. When we work with two *independent samples*, we assume that if the samples are selected at random or are randomly assigned to a group, the two samples will vary only by chance, and any difference obtained will be because of some underlying true difference between the groups.

Independent samples can occur in two scenarios.

In one, when testing the difference of the means between two fixed populations, we test the differences between samples from each population. When both samples are randomly selected, we can make inferences about the populations.

In the other, when working with subjects (people, pets, etc.), if we select a random sample and then randomly assign half of the subjects to one group and half to another, we can make inferences about the population.

Dependent samples are a bit different. Two samples of data are dependent when each observation in one sample is *paired* with a specific observation in the other sample. This gives rise to the term "matched-pairs," which is often used to describe this type of experimental situation. In short, these types of samples are related to each other. Dependent samples can occur in two scenarios. In one, each member will be measured twice, such as in a pre-test/post-test situation (scores on a test before and after the lesson). The other scenario is one in which an observation in one sample is matched with an observation in the second sample. Individuals in each pair in a blood pressure experiment might be matched on age, gender, and whether they smoke or not.

To distinguish between tests of hypotheses for independent and dependent samples, we use a different symbol for hypotheses with dependent samples. For dependent sample hypotheses, we use the symbol μ_d to symbolize the mean difference between the two samples. Therefore, in our null hypothesis, we state that the mean difference of the two samples is equal to 0, or $\mu_d = 0$.

Testing Hypotheses with Dependent Samples

Dependent samples generally occur with what is called matched pairs. Examples are pre-test scores versus post-test scores for students taking a CPR course, testing the effectiveness of a weight-loss program over 6 months, or taking a person's blood pressure before and after they run 1 mile. In these examples, the same person is providing both measurements. What we will test is if the before versus after score or measurement is the same.

Let's look at the weight loss example. If the weight-loss program is *not* effective, we expect the mean weight loss to be 0. If we have 10 people using the weight-loss program, we weigh each person at the beginning of the diet and then again at the end of a 6-month period. The weight loss for each person is the difference between the starting weight and the ending weight. Each person actually has two measurements, but we consider the difference (net weight loss) as a *single data value* for each person. For a sample of 10 people, we would have 20 measurements, but we would be interested in only the 10 values of the net weight loss.

Continuing with the weight-loss example, consider what the situation would be if the weight-loss program were bogus. If this were the case, then the expected weight loss for the participants would be 0. This gives rise to the null hypothesis for a matched-pair test. The null hypothesis would be

$$H_0: \mu_d = 0$$

This says that the mean difference (before versus after) of the 10 subjects would be 0, indicating that if the program were worthless, then we would expect no weight to be lost.

The alternative hypothesis would reflect that either the program leads to weight loss, or possibly it could even lead to weight gain. This would be a two-sided alternative hypothesis, and we would express it as

$$H_a: \mu_d \neq 0$$

The test statistic for the matched-pair (dependent samples) procedure is

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where \bar{d} is the mean difference

μ_d is the hypothesized mean difference (which is usually 0)

s_d is the sample standard deviation of the differences

n is the sample size

The formula for s_d is provided below. It is easier to use a calculator to determine the value of s_d .

Since our population standard is unknown, we estimate it by first using the following formula for the standard deviation of the samples:

$$s_d^2 = \frac{\sum(d - \bar{d})^2}{n - 1}$$

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n - 1}}$$

where:

s_d^2 is the variance of the difference data.

d is the difference between corresponding pairs.

\bar{d} is the mean difference of the data.

n is the number of pairs in the sample.

s_d is the standard deviation of the differences.

With the standard deviation of the samples, we can calculate the standard error of the difference between the two samples using the following formula:

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

After we calculate the standard error, we can use the general formula for the test statistic as shown below:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

The number of degrees of freedom will be $(n - 1)$, where n is the number of pairs in the experiment.

Example: A math teacher wants to determine the effectiveness of her statistics lesson and gives a pre-test and a post-test to 9 students in her class. Our null hypothesis is that the mean difference of the student scores is 0, and our alternative hypothesis is that the mean difference of student scores is not equal to 0. Here are the hypotheses:

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

The results for the pre-test and post-test are shown below:

TABLE 8.4:

Subject	Pre-test Score	Post-test Score	d difference	d^2
1	78	80	2	4
2	67	69	2	4
3	56	70	14	196
4	78	79	1	1
5	96	96	0	0
6	82	84	2	4
7	84	88	4	16
8	90	92	2	4
9	87	92	5	25
Sum	718	750	32	254
Mean	79.7	83.3	3.6	

Using the information from the table above, we can first solve for the standard deviation of the samples, then the standard error of the difference between the two samples, and finally the test statistic.

Standard Deviation:

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{254 - \frac{(32)^2}{9}}{8}} \approx 4.19$$

Standard Error of the Difference:

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{4.19}{\sqrt{9}} = 1.40$$

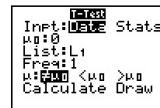
Test Statistic (*t*-test):

$$t = \frac{\bar{d} - 0}{s_{\bar{d}}} = \frac{3.6 - 0}{1.40} \approx 2.57$$

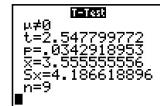
With 8 degrees of freedom (number of observations - 1) and a significance level of 0.05, we find our critical values to be ± 2.31 . Since our test statistic of 2.57 exceeds 2.31, we can **reject** the null hypothesis, which says that the mean before-versus-after difference is 0, and we then conclude that the lesson had an effect on student scores.

Using a Calculator to Perform a Matched-Pair (Dependent Samples) *t*-Test

If you use a graphing calculator to perform the matched-pair test, enter the difference data into List 1, then press [STAT][TESTS], and scroll to Option 2: T-Test and press [ENTER]. Your calculator screen should look like this:



Now scroll to Calculate and press [ENTER]. Your screen will now display a summary of the hypothesis test:



You can compare these calculator results to the manual ones from the example. The results are the same.

Testing Hypotheses with 2 Independent Samples

With 2 independent samples, we consider such things as the comparison of SAT scores for males versus females, the average MPGs for two different car brands, or the mean lengths of all fish caught at two different fishing tournaments. There is no matching of subjects; each sample is independently drawn from its population. Each sample mean is calculated, and the two means are compared.

When testing hypotheses with two independent samples, we follow steps similar to those when testing one random sample:

- State the null and alternative hypotheses.
- Choose α .
- Set the criterion (critical values) for rejecting the null hypothesis.
- Compute the test statistic.
- Make a decision: reject or fail to reject the null hypothesis.
- Interpret the decision within the context of the problem.

When stating the null hypothesis, we assume there is no difference between the means of the two independent samples. Therefore, our null hypothesis in this case would be the following:

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

Similar to the one-sample test, the critical values that we set to evaluate these hypotheses depend on our alpha level, and our decision regarding the null hypothesis is carried out in the same manner. However, since we have two samples, we calculate the test statistic a bit differently and use the formula shown below:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

where:

$\bar{x}_1 - \bar{x}_2$ is the difference between the sample means.

$\mu_1 - \mu_2$ is the difference between the hypothesized population means.

$s_{\bar{x}_1 - \bar{x}_2}$ is the standard error of the difference between sample means.

Example: The head of the English department is interested in the difference in writing scores between remedial freshman English students who are taught by different teachers. The incoming freshmen needing remedial services are randomly assigned to one of two English teachers and are given a standardized writing test after the first semester. We take a sample of eight students from one class and nine from the other. Is there a difference in achievement on the writing test between the two classes? Use a 0.05 significance level.

First, we would generate our hypotheses based on the two samples as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

Also, this is a two-tailed test, and for this example, we have two independent samples from the population and have a total of 17 students who we are examining. Since our sample size is so low, we use the t -distribution. In this example, we have 15 degrees of freedom, which is the number in the samples minus 2. With a 0.05 significance level and the t -distribution, we find that our critical values are 2.13 standard scores above and below the mean.

To calculate the test statistic, we first need to find the pooled estimate of variance from our sample. The data from the two groups are as follows:

TABLE 8.5:

Sample 1	Sample 2
35	52
51	87
66	76
42	62
37	81
46	71
60	55
55	67
53	

From this sample, we can calculate several descriptive statistics that will help us solve for the pooled estimate of variance:

TABLE 8.6:

Descriptive Statistic	Sample 1	Sample 2
Number (n)	9	8
Sum of Observations ($\sum x$)	445	551
Mean of Observations (\bar{x})	49.44	68.88
Sum of Squared Deviations $(\sum_{i=1}^n (x_i - \bar{x})^2)$	862.22	1058.88

Therefore, the pooled estimate of variance can be calculated as shown:

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = 128.07$$

This means that the standard error of the difference of the sample means can be calculated as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{128.07 \left(\frac{1}{9} + \frac{1}{8} \right)} \approx 5.50$$

Using this information, we can finally solve for the test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(49.44 - 68.88) - (0)}{5.50} \approx -3.53$$

Since -3.53 is less than the critical value of -2.13 , we decide to reject the null hypothesis and conclude that there is a significant difference in the achievement of the students assigned to different teachers.

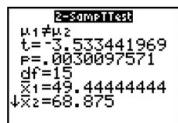
Using a Calculator to Perform a Test for Two Independent Samples

The calculations involved in 2-sample hypothesis testing are quite complicated. It is advisable that you use your graphing calculator to perform the calculations for the hypothesis test. For the above example, enter the 2 sets of data into Lists 1 and 2. Then press [STAT] and scroll to [TESTS]. Go to Option 4, 2-Samp T-test, and press [ENTER]. Here is a screenshot of what you should enter into your calculator:



(Note: In the screen above, the "Data" option is to be highlighted, because you have raw data to analyze. Sometimes, however, you will have summary statistics (the two sample means, the two sample standard deviations, and the two samples sizes) provided to you. If this is the case, you will highlight "Stats" and then the screen will ask you to input the summary data.)

Scroll down to "Calculate" and press [ENTER]. The results will appear on this screen:



Compare the calculator results to the manual calculations. They are identical.

Testing Hypotheses about the Difference in Proportions between Two Independent Samples

Suppose we want to test if there is a difference between proportions of two independent samples. As discussed in the previous lesson, proportions are used extensively in polling and surveys, especially by people trying to predict election results. It is possible to test a hypothesis about the proportions of two independent samples by using a method similar to that described above. We might perform these hypotheses tests in the following scenarios:

- When examining the proportions of children living in poverty in two different towns.
- When investigating the proportions of freshman and sophomore students who report test anxiety.
- When testing if the proportions of high school boys and girls who smoke cigarettes is equal.

In testing hypotheses about the difference in proportions of two independent samples, we state the hypotheses and set the criterion for rejecting the null hypothesis in similar ways as the other hypotheses tests. In these types of tests, we set the proportions of the samples equal to each other in the null hypothesis, $H_0 : p_1 = p_2$, and use the appropriate standard table to determine the critical values. Remember that we would use the binomial distribution for small sample probabilities, as we did earlier in this course. For large samples we use the normal approximation to the binomial, and thus, we use the z -table to calculate probabilities.

When solving for the test statistic in large samples, we use the following formula:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{p_1-p_2}}$$

where:

\hat{p}_1 and \hat{p}_2 are the observed sample proportions.

p_1 and p_2 are the population proportions under the null hypothesis. Generally, this quantity will be equal to 0.

$s_{p_1-p_2}$ is the standard error of the difference between independent proportions.

Similar to the standard error of the difference between independent means, we need to do a bit of work to calculate the standard error of the difference between independent proportions. To find the standard error under the null hypothesis, we assume that $p_1 - p_2 = p$, and we use all the data to calculate \hat{p} as an estimate for p as follows:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Now the standard error of the difference between independent proportions is $\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$.

This means that the test statistic is now $z = \frac{(\hat{p}_1 - \hat{p}_2) - (0)}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$.

Example: Suppose that we are interested in finding out which of two cities is more satisfied with the services provided by the city government. We take a survey and find the following results:

TABLE 8.7:

Number Satisfied	City 1	City 2
Yes	122	84
No	78	66
Sample Size	$n_1 = 200$	$n_2 = 150$
Proportion Who Said Yes	0.61	0.56

Is there a statistical difference in the proportions of citizens who are satisfied with the services provided by the city government? Use a 0.05 level of significance.

First, we establish the null and alternative hypotheses:

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

Since we have large sample sizes, we will use the z -distribution. At a 0.05 level of significance, our critical values are ± 1.96 . To solve for the test statistic, we must first solve for the standard error of the difference between proportions:

$$\hat{p} = \frac{(200)(0.61) + (150)(0.56)}{350} = 0.589$$

$$s_{p_1 - p_2} = \sqrt{(0.589)(0.411) \left(\frac{1}{200} + \frac{1}{150} \right)} \approx 0.053$$

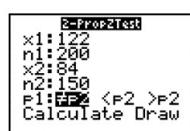
Therefore, the test statistic can be calculated as shown:

$$z = \frac{(0.61 - 0.56) - (0)}{0.053} \approx 0.94$$

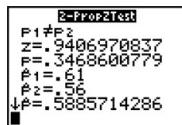
Since 0.94 does not exceed the critical value of 1.96, the null hypothesis is not rejected. Therefore, we can conclude that the difference in the proportions could have occurred by chance and that there is no difference in the level of satisfaction between citizens of the two cities.

Using a Calculator to Perform a Test for Two Proportions

Press [STAT], scroll to [TESTS] and then scroll down to Option 6: 2-Prop Z Test. Press [ENTER]. Make appropriate entries; your screen should look like this:



Now scroll to 'Calculate' and press [ENTER]. The calculator display has performed the computations necessary for conducting the hypothesis test. Here is the screenshot:



The value of the z test statistic is 0.94, which is the same answer as the manual calculation. The p -value for the test is .347. Since this value is greater than the significance level of .05, we would **not reject** the null hypothesis.

Lesson Summary

In addition to testing single samples associated with a mean, we can also perform hypothesis tests with two samples. We can test two independent samples, which are samples that do not affect one another, or dependent samples, which are samples that are related to each other.

Dependent samples are also called matched pairs. From each pair we collect 2 measurements, and we analyze the difference between these values. To calculate the test statistic for two dependent samples, we use the following formula:

$$t = \frac{\bar{d} - 0}{s_d} \text{ with } s_d = \sqrt{\frac{\sum d^2 - (\sum d)^2/n}{n-1}}$$

When testing a hypothesis about **two independent samples**, we follow a similar process as when testing one random sample. However, when computing the test statistic, we need to calculate the estimated standard error of the difference between sample means, which is found by using the following formula:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \text{ with } s^2 = \frac{ss_1 + ss_2}{n_1 + n_2 - 2}$$

We carry out the test on the means of two independent samples in way similar to that of testing one random sample. However, we use the following formula to calculate the test statistic, with the standard error defined above:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

We can also test the proportions associated with two independent samples. In order to calculate the test statistic associated with two independent samples, we use the formula shown below:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (0)}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ with } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Review Questions

1. In hypothesis testing, we have scenarios that have both dependent and independent samples. Give an example of an experiment with dependent samples and an experiment with two independent samples.
2. True or False: When we test the difference between the means of males and females on the SAT, we are using independent samples.

3. A study is conducted on the effectiveness of a drug on the hyperactivity of laboratory rats. Two random samples of rats are used for the study. One group is given Drug A, and the other group is given Drug B. The number of times that each rat pushes a lever is recorded. The following results for this test were calculated:

TABLE 8.8:

	Drug A	Drug B
\bar{x}	75.6	72.8
n	18	24
s^2	12.25	10.24
s	3.5	3.2

- (a) Does this scenario involve dependent or independent samples? Explain.
 (b) What would the hypotheses be for this scenario?
 (c) Is this a z -test or a t -test? Defend your choice.
 (d) Use your calculator to calculate the test statistic and the p -value. What is the test statistic, and at an alpha level of 0.05, what conclusions would you make about the null hypothesis?
4. A survey is conducted on attitudes towards drinking. A random sample of eight married couples is selected, and the husbands and wives respond to an attitude-toward-drinking scale. The scores are as follows:

TABLE 8.9:

Husbands	Wives
16	15
20	18
10	13
15	10
8	12
19	16
14	11
15	12

- (a) Why is a matched-pair test appropriate for these data? What are the hypotheses for this scenario?
 (b) Is this a z -test or a t -test? Defend your choice.
 (c) Use your calculator to perform the t test. What is the test statistic, and at an alpha level of 0.05, what conclusions would you make about the null hypothesis?
5. For two high schools, we are interested in comparing the proportion of students who drive to school. We ask a random sample of 100 students from each school. At Beach High, 65 of the students in the sample drive to school. At Island High, 58 of the students drive to school. State the hypotheses and perform the hypothesis test. At the .05 level of significance, do we have evidence to say that there is a difference in the proportions of students who drive to school from the two high schools?

Answers: (1) see detailed solutions (2) True (3)(a) Independent (3)(b) $H_0: \mu_A = \mu_B$ and $H_a: \mu_A \neq \mu_B$ (3)(c) 2-sample t test; small sample size (3)(d) $t = 2.70$ and p -value = .0102 (4)(a) see detailed solution; $H_0: \mu_d = 0$ (4)(b) Matched-pairs t test (4)(c) $t = 1.12$ and p -value = .299; do not reject H_0 (5) $H_0: p_B = p_I$ and $H_a: p_B \neq p_I$ $z = 1.02$ and p -value = .309; do not reject H_0

Keywords α α is called the **level of significance**.

$$\alpha = P(\text{rejecting a true null hypothesis}) = P(\text{making a Type I error})$$

Alpha level

The general approach to hypothesis testing focuses on the *type I error*: rejecting the null hypothesis when it may be true. The level of significance, also known as the *alpha level*.

Alternative hypothesis

The alternative hypothesis to be accepted if the null hypothesis is rejected.

 β β is the probability of making a type II error.

$$\beta = P(\text{not rejecting a false null hypothesis}) = P(\text{making a Type II error})$$

Critical region

The values of the test statistic that allow us to reject the null hypothesis.

Critical values

To calculate the critical regions, we must first find the cut-offs, or the *critical values*, where the critical regions start.

Degrees of freedom

how the shape of Student's *t*-distribution corresponds to the sample size (which corresponds to a measure called the degrees of freedom).

Dependent samples

Two samples of data are dependent when each observation in one sample is paired with a specific observation in the other sample. Also called matched pairs.

Hypothesis testing

Testing the difference between a hypothesized value of a parameter and the test statistic.

Independent samples

When we work with two *independent samples*, we assume that if the samples are selected at random, the two samples will vary only by chance.

Level of significance

The strength of the sample evidence needed to reject the null hypothesis.

Null hypothesis (H_0)

The default hypothesis, a hypothesis about a parameter that is tested.

One-tailed test

When the alternative hypothesis is one-sided, then the rejection region is taken only on one side of the sampling distribution. It is called one-tailed test.

P-value

We can also evaluate a hypothesis by asking, “What is the probability of obtaining the value of the test statistic that we did if the null hypothesis is true?” This is called the *P*–value.

Pooled estimate of variance

Here, n_1 and n_2 are the sizes of the two samples, and s^2 , the *pooled estimate of variance*, is calculated with the formula $s^2 = \frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$.

Power of a test

The power of a test is defined as the probability of rejecting the null hypothesis when it is false. This is a correct decision, and it is calculated by the formula $\text{power} = (1 - \beta)$.

Standard error of the difference

When testing a hypothesis about two independent samples, we follow a similar process as when testing one random sample. However, when computing the test statistic, we need to calculate the estimated standard error

of the difference between sample means, $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$.

Student's *t*–distributions

Student's *t*– distributions are a family of distributions that, like the normal distribution, are symmetrical, bell-shaped, and centered on a mean. Each curve's shape is dependent on the degrees of freedom.

Test statistic

Before evaluating our hypotheses by determining the critical region and calculating the *test statistic*, we need to confirm that the distribution is normal and determine the hypothesized mean, μ , of the distribution. $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Two-tailed test

The **two-tailed test** is a statistical test used in inference, in which a given statistical hypothesis, H_0 (the null hypothesis), will be rejected when the value of the test statistic is either sufficiently small or sufficiently large.

Type I error

A type I error occurs when one rejects the null hypothesis when it is true. The probability of a Type I error is the level of significance, α .

Type II error

A type II error occurs when one should have rejected the null hypothesis, because it is false and should be rejected, but it wasn't. The probability of a Type II error is β .

8.6 References

1. CK-12 Foundation. . CCSA
2. . . CC BY-NC-SA
3. . . CC BY-NC-SA
4. CK-12 Foundation. . CCSA
5. CK-12 Foundation. . CCSA
6. . . CC BY-NC-SA
7. . . CC BY-NC-SA
8. CK-12 Foundation. . CCSA
9. CK-12 Foundation. . CCSA
10. CK-12 Foundation. . CCSA
11. CK-12 Foundation. . CCSA
12. CK-12 Foundation. . CCSA
13. CK-12 Foundation. . CCSA
14. CK-12 Foundation. . CCSA

CHAPTER

9**Regression and Correlation****Chapter Outline**

- 9.1 SCATTERPLOTS AND LINEAR CORRELATION**
 - 9.2 LEAST-SQUARES REGRESSION**
 - 9.3 INFERENCES ABOUT REGRESSION**
 - 9.4 REFERENCES**
-

9.1 Scatterplots and Linear Correlation

Learning Objectives

- Understand the concepts of bivariate data and correlation, and the use of scatterplots to display bivariate data.
- Understand when the terms 'positive', 'negative', 'strong', and 'perfect' apply to the correlation between two variables in a scatterplot graph.
- Calculate the linear correlation coefficient and coefficient of determination of bivariate data, using technology tools to assist in the calculations.
- Understand properties and common errors of correlation.

Introduction

So far we have learned how to describe distributions of a single variable and how to perform hypothesis tests concerning parameters of these distributions. But what if we notice that two variables seem to be related? We may notice that the values of two variables, such as verbal SAT score and GPA, behave in the same way and that students who have a high verbal SAT score also tend to have a high GPA (see table below). In this case, we would want to study the nature of the connection between the two variables.

TABLE 9.1: A table of verbal SAT values and GPAs for seven students.

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

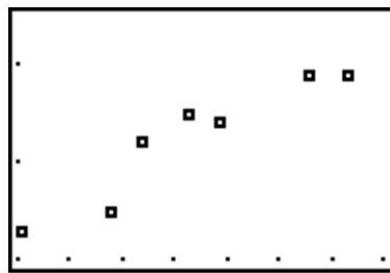
These types of studies are quite common, and we can use the concept of correlation to describe the relationship between the two variables.

Bivariate Data, Correlation Between Values, and the Use of Scatterplots

Correlation measures the relationship between bivariate data. *Bivariate data* are data sets in which each subject has two observations associated with it. In our example above, we notice that there are two observations (verbal SAT score and GPA) for each subject (in this case, a student). Can you think of other scenarios when we would use bivariate data?

If we carefully examine the data in the example above, we notice that those students with high SAT scores tend to have high GPAs, and those with low SAT scores tend to have low GPAs. In this case, there is a tendency for students to score similarly on both variables, and the performance between variables appears to be related.

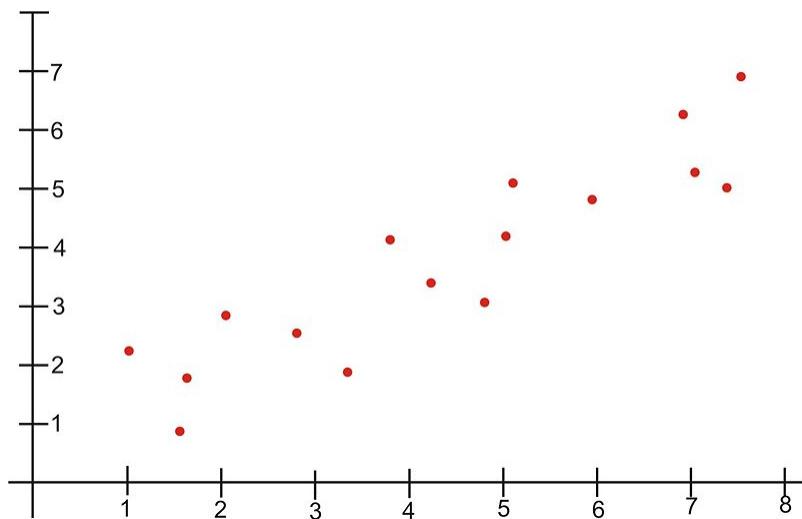
Scatterplots display these bivariate data sets and provide a visual representation of the relationship between variables. In a scatterplot, each point represents a paired measurement of two variables for a specific subject, and each subject is represented by one point on the scatterplot.



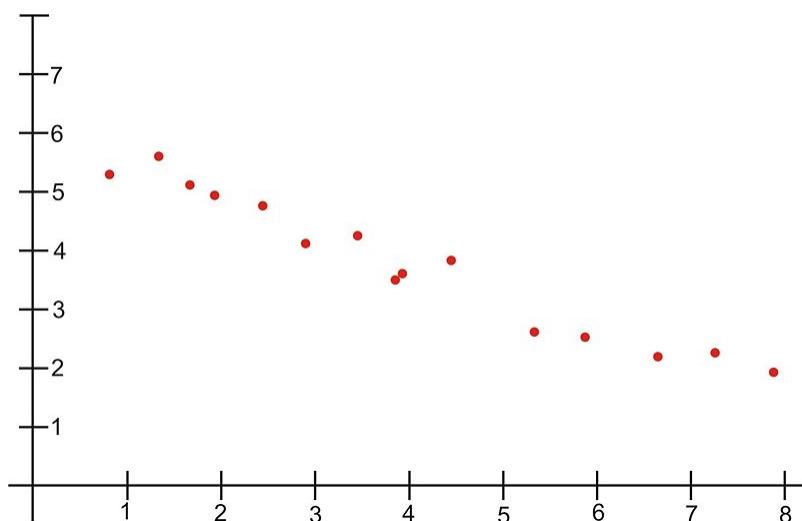
Correlation Patterns in Scatterplot Graphs

Examining a scatterplot graph allows us to obtain some idea about the relationship between two variables.

When the points on a scatterplot graph produce a lower-left-to-upper-right pattern (see below), we say that there is a *positive correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be high as well, and vice versa.

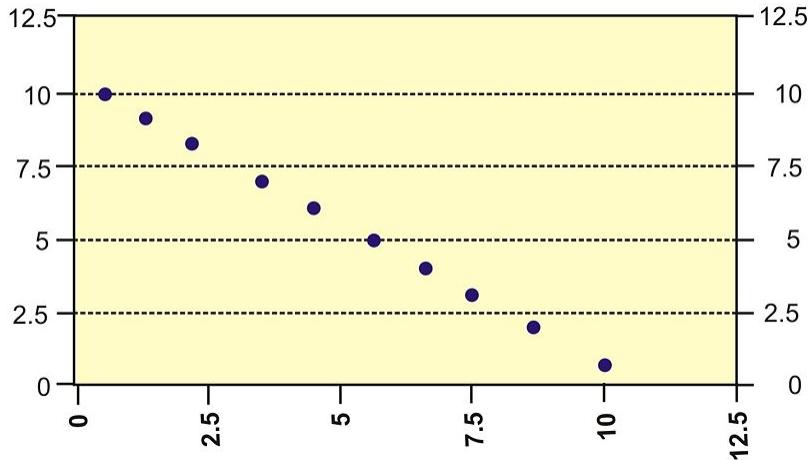


When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a *negative correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low, and vice versa.

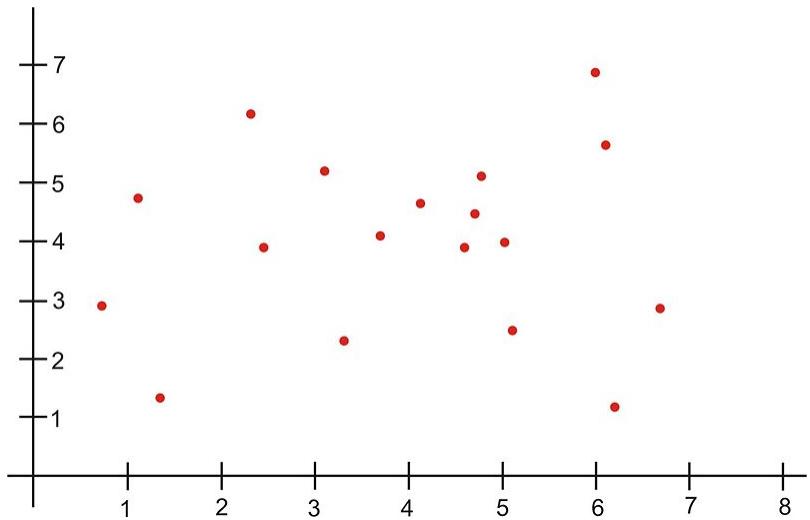


When all the points on a scatterplot lie on a straight line, you have what is called a *perfect correlation* between the two variables (see below).

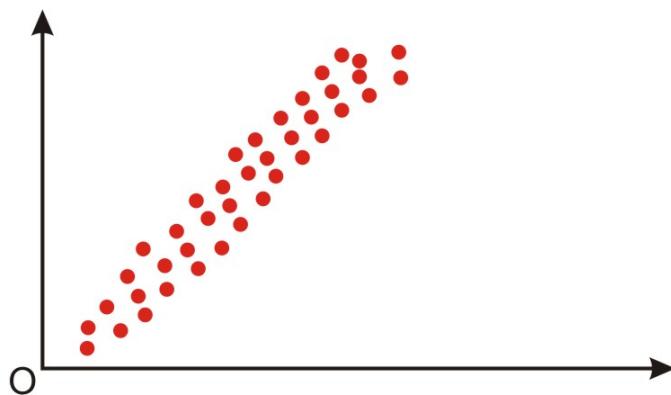
Perfect Negative Correlation



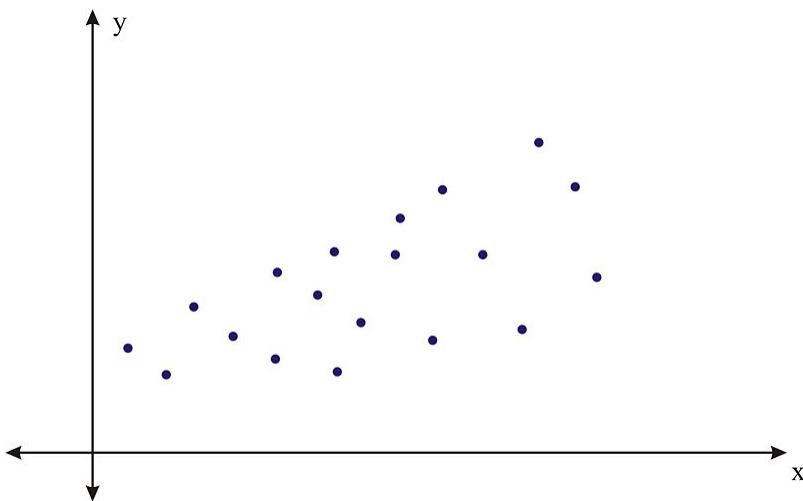
A scatterplot in which the points do not have a linear trend (either positive or negative) is called a *zero correlation* or a *near-zero correlation* (see below).



When examining scatterplots, we also want to look not only at the direction of the relationship (positive, negative, or zero), but also at the *magnitude* of the relationship. If we drew an imaginary oval around all of the points on the scatterplot, we would be able to see the extent, or the magnitude, of the relationship. If the points are close to one another and the width of the imaginary oval is small, this means that there is a strong correlation between the variables (see below).



However, if the points are far away from one another, and the imaginary oval is very wide, this means that there is a weak correlation between the variables (see below).



Correlation Coefficients

While examining scatterplots gives us some idea about the relationship between two variables, we use a statistic called the ***correlation coefficient*** to give us a more precise measurement of the relationship between the two variables. The correlation coefficient is an index that describes the relationship and can take on values between -1.0 and $+1.0$, with a positive correlation coefficient indicating a positive correlation and a negative correlation coefficient indicating a negative correlation.

The absolute value of the coefficient indicates the magnitude, or the strength, of the relationship. The closer the absolute value of the coefficient is to 1 , the stronger the relationship. For example, a correlation coefficient of 0.20 indicates that there is a weak linear relationship between the variables, while a coefficient of -0.90 indicates that there is a strong linear relationship.

The value of a perfect positive correlation is 1.0 , while the value of a perfect negative correlation is -1.0 .

When there is no linear relationship between two variables, the correlation coefficient is 0 . It is important to remember that a correlation coefficient of 0 indicates that there is no *linear* relationship, but there may still be a strong relationship between the two variables. For example, there could be a quadratic relationship between them.

The Pearson product-moment correlation coefficient is a statistic that is used to measure the strength and direction of a linear correlation. It is symbolized by the letter r . To understand how this coefficient is calculated, let's suppose that there is a positive relationship between two variables, X and Y . If a subject has a score on X that is above the mean, we expect the subject to have a score on Y that is also above the mean. Pearson developed his correlation

coefficient by computing the sum of cross products. He multiplied the two scores, X and Y , for each subject and then added these cross products across the individuals. Next, he divided this sum by the number of subjects minus one. This coefficient is, therefore, the mean of the cross products of scores.

Pearson used standard scores (z -scores, t -scores, etc.) when determining the coefficient.

Therefore, the formula for this coefficient is as follows:

$$r_{XY} = \frac{\sum z_X z_Y}{n - 1}$$

In other words, the coefficient is expressed as the sum of the cross products of the standard z -scores divided by the number of degrees of freedom.

An equivalent formula that uses the raw scores rather than the standard scores is called the raw score formula and is written as follows:

$$r_{XY} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2]} \sqrt{[n \sum y^2 - (\sum y)^2]}}$$

Again, this formula is most often used when calculating correlation coefficients from original data. Let's use our example from the introduction to demonstrate how to calculate the correlation coefficient using the raw score formula.

Example: What is the Pearson product-moment correlation coefficient for the two variables represented in the table below?

TABLE 9.2: The table of values for this example.

Student	SAT Score	GPA
1	595	3.4
2	520	3.2
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

In order to calculate the correlation coefficient, we need to calculate several pieces of information, including xy , x^2 , and y^2 . Therefore, the values of xy , x^2 , and y^2 have been added to the table.

TABLE 9.3:

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

TABLE 9.3: (continued)

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
---------	-------------------	-------------	------	-------	-------

Applying the formula to these data, we find the following:

$$r_{XY} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2]} \sqrt{[n \sum y^2 - (\sum y)^2]}} = \frac{(7)(13262) - (3970)(22.7)}{\sqrt{[(7)(2321400) - 3970^2][(7)(76.01) - 22.7^2]}}$$

$$= \frac{2715}{2864.22} \approx 0.95$$

The correlation coefficient not only provides a measure of the relationship between the variables, but it also gives us an idea about how much of the total variance of one variable can be associated with the variance of the other. For example, the correlation coefficient of 0.95 that we calculated above tells us that to a high degree, the variance in the scores on the verbal SAT is associated with the variance in the GPA, and vice versa. For example, we could say that factors that influence the verbal SAT, such as health, parent college level, etc., would also contribute to individual differences in the GPA. The higher the correlation we have between two variables, the larger the portion of the variance that can be explained by the independent variable.

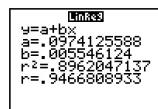
The calculation of this variance is called the **coefficient of determination** and is calculated by squaring the correlation coefficient. Therefore, the coefficient of determination is written as r^2 . The result of this calculation indicates the proportion of the variance in one variable that can be associated with the variance in the other variable.

Using Technology to Calculate the Correlation Coefficient

Your TI-83/84 calculator can be used to ease the burdensome calculations presented above. First, prepare the calculator for displaying the correlation coefficient by pressing the [2nd] and [0] buttons to display the calculator's Catalog, the entries in which are listed alphabetically. Scroll down to "Diagnostic On" and press [ENTER] twice.

Now press [STAT] and then [ENTER] to access the listmaking screen. Into List 1 enter the first set of data values, such as the SAT scores shown above. Then, into List 2, enter the GPA data.

Now press [STAT] and scroll right to CALC. Scroll down to Option 8 (Lin Reg ($a + bx$)). At the new screen, enter L₁ for XList and L₂ for YList. Scroll down to Calculate and press [ENTER]. Your screen will now display:



The values for the correlation coefficient r and the coefficient of determination r^2 are displayed. Note that these values agree with those obtained by using the raw score formula.

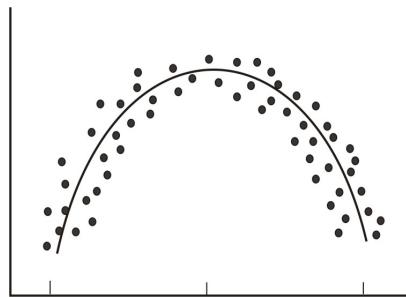
The Properties and Common Errors of Correlation

Correlation is a measure of the linear relationship between two variables—it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to remember that we are interpreting the variables and the variance not as causal, but instead as relational.

When examining correlation, there are three things that could affect our results: linearity, homogeneity of the group, and sample size.

Linearity

As mentioned, the correlation coefficient is the measure of the linear relationship between two variables. However, while many pairs of variables have a linear relationship, some do not. For example, let's consider performance anxiety. As a person's anxiety about performing increases, so does his or her performance up to a point. (We sometimes call this good stress.) However, at some point, the increase in anxiety may cause a person's performance to go down. We call these non-linear relationships *curvilinear relationships*. We can identify curvilinear relationships by examining scatterplots (see below). One may ask why curvilinear relationships pose a problem when calculating the correlation coefficient. The answer is that if we use the traditional formula to calculate these relationships, it will not be an accurate index, and we will be underestimating the relationship between the variables. If we graphed performance against anxiety, we would see that anxiety has a strong affect on performance. However, if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient is not always the best statistic to use to understand the relationship between variables.



Homogeneity of the Group

Another error we could encounter when calculating the correlation coefficient is homogeneity of the group. When a group is homogeneous, or possesses similar characteristics, the range of scores on either or both of the variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQs over 140) are sampled, we will most likely find a very low correlation between IQ and salary, since most members will have a consistently high IQ, but their salaries will still vary. This does not mean that there is not a relationship—it simply means that the restriction of the sample limited the magnitude of the correlation coefficient.

Sample Size

Finally, we should consider sample size. One may assume that the number of observations used in the calculation of the correlation coefficient may influence the magnitude of the coefficient itself. However, this is not the case. Yet while the sample size does not affect the correlation coefficient, it may affect the accuracy of the relationship. The larger the sample, the more accurate of a predictor the correlation coefficient will be of the relationship between the two variables.

Lesson Summary

Bivariate data are data sets with two observations that are assigned to the same subject. Correlation measures the direction and magnitude of the linear relationship between bivariate data. When examining scatterplot graphs, we can determine if correlations are positive, negative, perfect, or zero. A correlation is strong when the points in the scatterplot are close together.

The correlation coefficient is a precise measurement of the relationship between the two variables. This index can take on values between and including -1.0 and $+1.0$.

To calculate the correlation coefficient, we most often use the raw score formula, which allows us to calculate the coefficient by hand.

This formula is as follows: $r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]} \sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}.$

When calculating the correlation coefficient, there are several things that could affect our computation, including curvilinear relationships, homogeneity of the group, and the size of the group.

Review Questions

1. Give 2 scenarios or research questions where you would use bivariate data sets.
2. Sketch four scatterplot graphs showing:
 - a. a positive correlation
 - b. a negative correlation
 - c. a perfect correlation
 - d. a zero correlation
3. Sketch two scatterplot graphs showing:
 - a. a weak correlation
 - b. a strong correlation.
4. What does the correlation coefficient measure?
5. The following observations were taken for five students measuring grade and reading level.

TABLE 9.4: A table of grade and reading level for five students.

Student Number	Grade	Reading Level
1	2	6
2	6	14
3	5	12
4	4	10
5	1	4

- (a) Draw a scatterplot for these data. What type of relationship does this correlation have?
- (b) Use the raw score formula to compute the Pearson correlation coefficient. Double check your calculations using technology.
6. A teacher gives two quizzes to his class of 10 students. The following are the scores of the 10 students.

TABLE 9.5: Quiz results for ten students.

Student	Quiz 1	Quiz 2
1	15	20
2	12	15
3	10	12
4	14	18
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

- (a) Use technology to compute the Pearson correlation coefficient, r , between the scores on the two quizzes.
- (b) Find the percentage of the variance, r^2 , in the scores of Quiz 2 associated with the variance in the scores of Quiz 1.
- (c) Interpret both r and r^2 in words.
7. What are the three factors that we should be aware of that affect the magnitude and accuracy of the Pearson correlation coefficient?

Answers: (1)(2)(3)(4)(5a) see detailed answers (5)(b) $r = 1$ (6)(a) $r = .568$ (6)(b) $r^2 = .323$ (6c)(7) see detailed answers

9.2 Least-Squares Regression

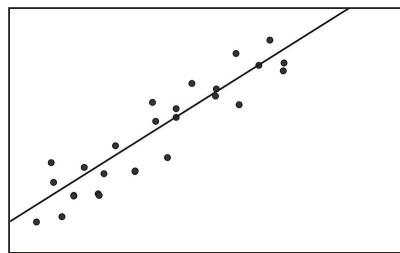
Learning Objectives

- Calculate and graph a regression line.
- Predict values using bivariate data plotted on a scatterplot.
- Understand outliers and influential points.
- Perform transformations to achieve linearity.
- Calculate residuals and understand the least-squares property and its relation to the regression equation.
- Plot residuals and test for linearity.

Introduction

In the last section, we learned about the concept of correlation, which we defined as the measure of the linear relationship between two variables. As a reminder, when we have a strong positive correlation, we can expect that if the score on one variable is high, the score on the other variable will also most likely be high. With correlation, we are able to roughly predict the score of one variable when we have the other. Prediction is simply the process of estimating scores of one variable based on the scores of another variable.

In the previous section, we illustrated the concept of correlation through scatterplot graphs. We saw that when variables were correlated, the points on a scatterplot graph tended to follow a straight line. If we could draw this straight line, it would, in theory, represent the change in one variable associated with the change in the other. This line is called the *least squares line*, or the *linear regression line* (see figure below).



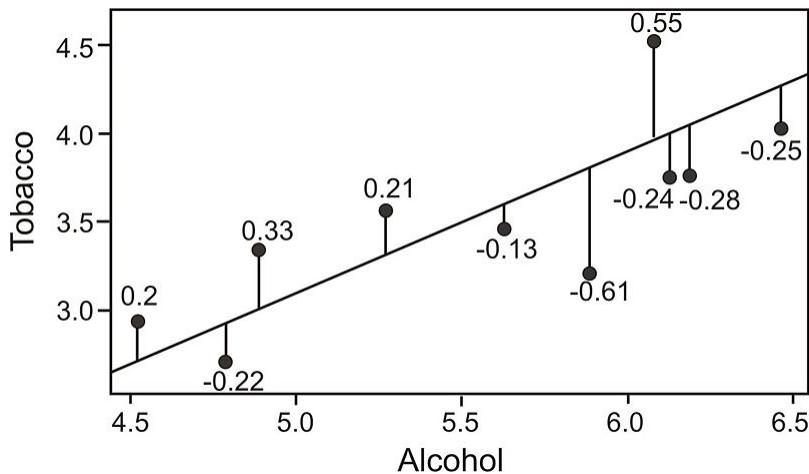
Calculating and Graphing the Regression Line

Linear regression involves using data to calculate a line that best fits that data and then using that line to predict scores. In linear regression, we use one variable (the *explanatory variable*) to predict the outcome of another (the *response variable*). To calculate this line, we analyze the patterns between the two variables.

We are looking for a line of best fit, and there are many ways one could define this "best" fit. Statisticians define this line to be the one which minimizes the sum of the squared distances from the observed data to the line.

To determine this line, we want to find the change in X that will be reflected by the average change in Y . After we calculate this average change, we can apply it to any value of X to get an approximation of Y . Since the regression line is used to predict the value of Y for any given value of X , all predicted values will be located on the regression line, itself. Therefore, we try to fit the regression line to the data by having the smallest sum of squared distances possible from each of the data points to the line. In the example below, you can see the calculated distances, or residual values, from each of the observations to the regression line. This method of fitting the data line so that there

is minimal difference between the observations and the line is called the *method of least squares*, which we will discuss further in the following sections.



As you can see, the regression line is a straight line that expresses the relationship between two variables. When predicting one score by using another, we use an equation such as the following, which is equivalent to the slope-intercept form of the equation for a straight line:

$$\hat{y} = a + bx$$

where:

\hat{y} is the score that we are trying to predict. We pronounce this as "y hat".

b is the slope of the line, also called the regression coefficient.

a is the y -intercept, or the value of Y when the value of X is 0.

This formula will use the value of x (the explanatory, or independent, variable) to **predict** the value of y (the response variable, or dependent variable.) For example, if we are discussing the linear relationship of years of education and annual income, we would use x , years of education (the explanatory variable), to predict the annual income (response variable). In the formula, we use the symbol \hat{y} for the y -value of income, because it is **predicted** from the linear equation.

To calculate the line itself, we need to find the values for b (the slope) and a (the y -*intercept*). The regression coefficient explains the nature of the relationship between the two variables. It is the slope of the regression line, and slope measures the change in y with respect to a unit change in x . Essentially, the regression coefficient tells us that a certain change in the explanatory variable (x) is associated with a certain change in the response variable (y). For example, if we had a regression coefficient (slope) of 10.76, rewritten as $\frac{10.76}{1}$ we would say that a change of 1 unit in X is associated with a change of 10.76 units of Y . To calculate this regression coefficient, we can use the following formulas:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

or

$$b = (r) \frac{s_y}{s_x}$$

where:

r is the correlation between the variables X and Y .

s_Y is the standard deviation of the Y scores.

s_X is the standard deviation of the X scores.

In addition to calculating the slope b , we also need to calculate the y -intercept, which is the place where the line crosses the y -axis. We use the following formula to calculate the y -intercept:

$$a = \frac{\sum y - b \sum x}{n} = \bar{y} - b \bar{x}$$

Example: Find the least squares line (also known as the linear regression line or the *line of best fit*) for the example measuring the verbal SAT scores and GPAs of students that was used in the previous section. Here are sample data collected from 7 students:

TABLE 9.6: SAT and GPA data including intermediate computations for computing a linear regression.

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
1	595	3.4	2023	354025	11.56
2	520	3.2	1664	270400	10.24
3	715	3.9	2789	511225	15.21
4	405	2.3	932	164025	5.29
5	680	3.9	2652	462400	15.21
6	490	2.5	1225	240100	6.25
7	565	3.5	1978	319225	12.25
Sum	3970	22.7	13262	2321400	76.01

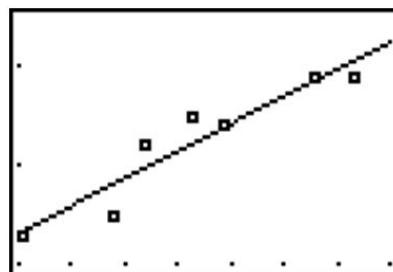
Using these data points, we first calculate the slope b and the y -intercept as follows:

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{(7)(13,262) - (3,970)(22.7)}{(7)(2,321,400) - 3,970^2} = \frac{2715}{488900} \approx 0.0055$$

$$a = \frac{\sum y - b \sum x}{n} \approx 0.097$$

Note: If you performed the calculations yourself and did not get exactly the same answers, it is probably due to rounding in the table for xy .

Now that we have the equation of this line, it is easy to plot on a scatterplot. To plot this line, we simply substitute two values of X and calculate the corresponding Y values to get two pairs of coordinates. Let's say that we wanted to plot this example on a scatterplot. We would choose two hypothetical values for X (say, 400 and 500) and then solve for Y in order to identify the coordinates (400, 2.1214) and (500, 2.6761). From these pairs of coordinates, we can draw the regression line on the scatterplot.



Using the TI-84 Calculator to Determine the Slope and y-intercept

The calculations presented are unwieldy, and you can instead use your graphing calculator to determine the values of the slope b and the y -intercept a , as follows:

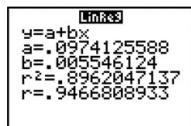
1. Press [STAT][EDIT] and then [ENTER] to get to the List function. Enter the SAT explanatory variable (x) data into L₁ and the GPA response variable (y) data into L₂. Your calculator display will look like this:

L1	L2	L3	3
585	3.4		
520	2.2		
715	3.2		
405	2.0		
680	3.0		
480	2.6		
565	3.0		

Now press [STAT] and scroll to [CALC]. Scroll down to Option 8, LinReg (a + bx) and press [ENTER]. For XList, you want to enter L₁, and to enter this into the calculator, press [2nd][1], and L₁ will be entered. For YList, press [2nd][2], and L₂ will appear in your display window. Skip the next two options. Your display will look like this:



Now scroll down to Calculate and press [ENTER]. Your screen will display the values of a and b , and also it will show the value of the correlation coefficient r . (We don't need the value of r for the regression equation, but it is nice to know that it can be calculated if needed later.)



The values of a and b are 0.097 and 0.0055, respectively, which are the same values obtained by using the formulas.

Predicting Values Using the Regression Equation

One of the uses of a regression line is to predict values. After calculating this line, we are able to predict values by simply substituting a value of an explanatory variable, X , into the regression equation and solving the equation for the predicted value of the response variable, \hat{y} . In our example above, we can predict the students' GPA's from their SAT scores by plugging in the desired values into our regression equation, $\hat{y} = 0.097 + 0.0055x$.

For example, say that we wanted to predict the GPA for a student who had an SAT score of 600. We plug the SAT score (explanatory variable) of $x = 600$ into the regression equation, as follows:

$$\hat{y} = 0.097 + .0055(600)$$

$$\hat{y} = 3.397$$

Thus, the predicted GPA for a student with an SAT score of 600 is 3.397, which rounds to 3.4.

Here are some other predicted values of GPA for selected values of SAT. We used the regression equation for these predictions.

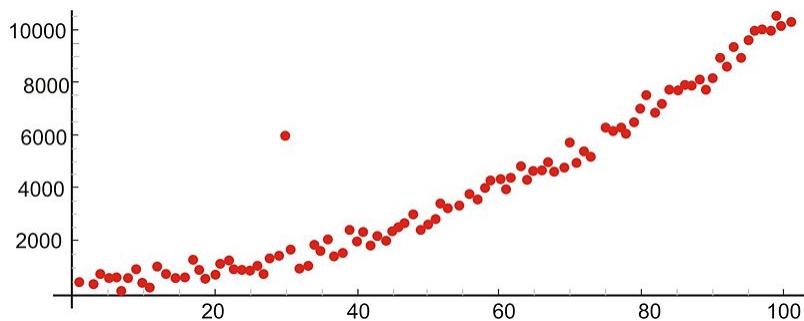
TABLE 9.7: GPA/SAT data, including predicted GPA values from the linear regression.

Student	SAT Score (X)	GPA (Y)	Predicted GPA (\hat{Y})
1	595	3.4	3.4
2	520	3.2	3.0
3	715	3.9	4.1
4	405	2.3	2.3
5	680	3.9	3.9
6	490	2.5	2.8
7	565	3.5	3.2
Hypothetical	600		3.4
Hypothetical	500		2.9

As you can see, we are able to predict the value for Y for any value of X within a specified range.

Outliers and Influential Points

An *outlier* is an extreme observation that does not fit the general correlation or regression pattern (see figure below). Since it is an unusual observation, the inclusion of an outlier may affect the slope and the y -intercept of the regression line. When examining a scatterplot graph and calculating the regression equation, it is worth considering whether extreme observations should be included or not. In the following scatterplot, the outlier has approximate coordinates of (30, 6,000).



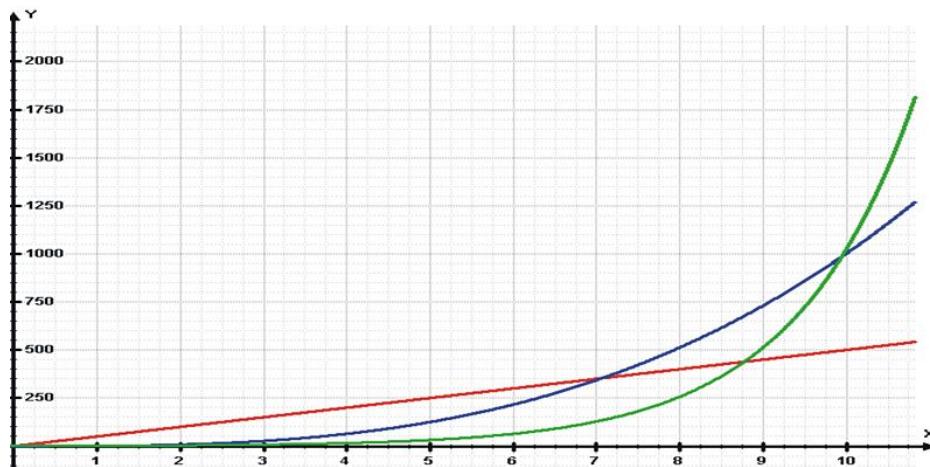
Let's use our example above to illustrate the effect of a single outlier. Say that we have a student who has a high GPA but who suffered from test anxiety the morning of the SAT verbal test and scored a 410. Using our original regression equation, we would expect the student to have a GPA of 2.2. But, in reality, the student has a GPA equal to 3.9. The inclusion of this value would change the slope of the regression equation from 0.0055 to 0.0032, which is quite a large difference.

There is no set rule when trying to decide whether or not to include an outlier in regression analysis. This decision depends on the sample size, how extreme the outlier is, and the normality of the distribution. As a general rule of thumb, we should consider values that are 1.5 times the inter-quartile range below the first quartile or above the third quartile as outliers. Extreme outliers are values that are 3.0 times the inter-quartile range below the first quartile or above the third quartile.

Transformations to Achieve Linearity

Sometimes we find that there is a relationship between X and Y , but it is not best summarized by a straight line. When looking at the scatterplot graphs of correlation patterns, these relationships would be shown to be curvilinear. While many relationships are linear, there are quite a number that are not, including learning curves (learning more

quickly at the beginning, followed by a leveling out) and exponential growth (doubling in size, for example, with each unit of growth). Below is an example of a growth curve describing the growth of a complex society:

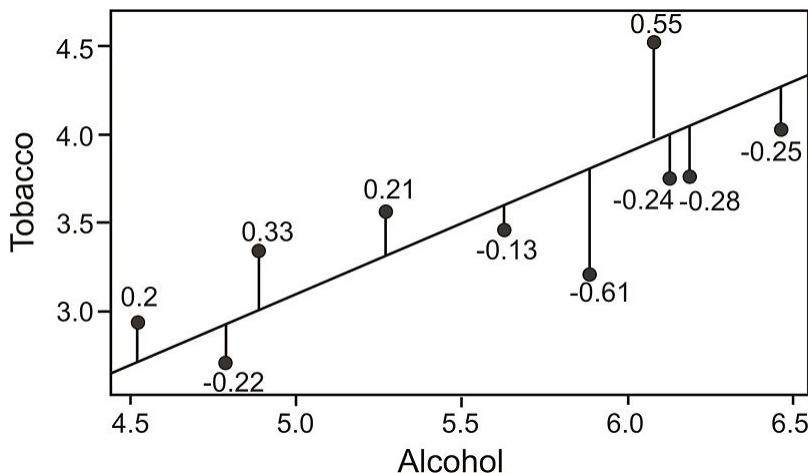


Since this is not a linear relationship, we cannot immediately fit a regression line to this data. However, we can perform a *transformation* to achieve a linear relationship. We commonly use transformations in everyday life. For example, the Richter scale, which measures earthquake intensity, and the idea of describing pay raises in terms of percentages are both examples of making transformations of non-linear data.

A common transformation that would change curvy graphs into near perfect lines would involve taking the logarithm of the x values and the y values. However, logarithmic transformations are beyond the scope of our course. You might want to consult an advanced statistics textbook for more information.

Calculating Residuals and Understanding their Relation to the Regression Equation

Recall that the linear regression line is the line that best fits the given data. Ideally, we would like to minimize the distances of all data points to the regression line. These distances are called the error, e , and are also known as the *residual values*. As mentioned, we fit the regression line to the data points in a scatterplot using the least-squares method. A good line will have small residuals. Notice in the figure below that the residuals are the vertical distances between the observations and the predicted values on the regression line:



To find the residual values, we subtract the predicted values from the actual values, so $e = y - \hat{y}$. Theoretically, the sum of all residual values is zero, since we are finding the line of best fit, with the predicted values as close as

possible to the actual value. It does not make sense to use the sum of the residuals as an indicator of the fit, since, again, the negative and positive residuals always cancel each other out to give a sum of zero. Therefore, we try to minimize the sum of the squared residuals, or $\sum(y - \hat{y})^2$.

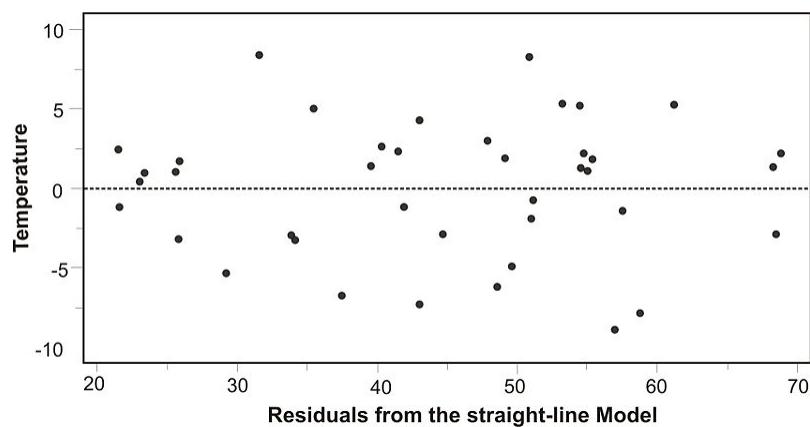
Example: Calculate the residuals for the predicted and the actual GPA's from our sample above.

TABLE 9.8: SAT/GPA data, including residuals.

Student	SAT Score (X)	GPA (Y)	Predicted GPA (\hat{Y})	Residual Value	Residual Value Squared
1	595	3.4	3.4	0	0
2	520	3.2	3.0	0.2	0.04
3	715	3.9	4.1	-0.2	0.04
4	405	2.3	2.3	0	0
5	680	3.9	3.9	0	0
6	490	2.5	2.8	-0.3	0.09
7	565	3.5	3.2	0.3	0.09
$\Sigma(y - \hat{y})^2$					0.26

Plotting Residuals and Testing for Linearity

To test for linearity and to determine if we should drop extreme observations (or outliers) from our analysis, it is helpful to plot the residuals. When plotting, we simply plot the x -value for each observation on the x -axis and then plot the residual score on the y -axis. When examining this scatterplot, the data points should appear to have no correlation, with approximately half of the points above 0 and the other half below 0. In addition, the points should be evenly distributed along the x -axis. Below is an example of what a residual scatterplot should look like if there are no outliers and a linear relationship.



If the scatterplot of the residuals does not look similar to the one shown, we should look at the situation a bit more closely. For example, if more observations are below 0, we may have a positive outlying residual score that is skewing the distribution, and if more of the observations are above 0, we may have a negative outlying residual score. If the points are clustered close to the y -axis, we could have an x -value that is an outlier. If this occurs, we may want to consider dropping the observation to see if this would impact the plot of the residuals. If we do decide to drop the observation, we will need to recalculate the original regression line. After this recalculation, we will have a regression line that better fits a majority of the data.

Lesson Summary

Prediction is simply the process of estimating scores of one variable based on the scores of another variable. We use the least-squares regression line, or linear regression line, to predict the value of a variable.

Using this regression line, we are able to use the slope b and the y -intercept a to predict the scores of a response variable. The predictions are represented by the variable \hat{y} . The general form of the least-squares regression line is $\hat{y} = a + bx$.

The differences between the actual and the predicted values are called **residual** values. We can construct scatterplots of these residual values to examine outliers and to test for linearity.

Review Questions

1. A school nurse is interested in predicting scores on a memory test from the number of times that a student exercises per week. Below are her observations:

TABLE 9.9: A table of memory test scores compared to the number of times a student exercises per week.

Student	Exercise Per Week	Memory Test Score
1	0	15
2	2	3
3	2	12
4	1	11
5	3	5
6	1	8
7	2	15
8	0	13
9	3	2
10	3	4
11	4	2
12	1	8
13	1	10
14	1	12
15	2	8

- (a) Plot this data on a scatterplot, with the x -axis representing the number of times exercising per week and the y -axis representing memory test score.
- (b) Does this appear to be a linear relationship? Why or why not?
- (c) What regression equation would you use to construct a linear regression model?
- (d) What is the slope in this linear regression model and what does this mean in words?
- (e) Calculate the regression equation for these data.
- (f) Draw the regression line on the scatterplot.
- (g) What is the predicted memory test score of a student who exercises 3 times per week?
- (h) What is the residual for a student who exercises 3 times per week?
- (i) Calculate the residuals for each of the observations and plot these residuals on a scatterplot.
- (j) Examine this scatterplot of the residuals. Does a linear model appear to be appropriate for these data?

Answers: See detailed answers for this checkpoint.

9.3 Inferences about Regression

Learning Objectives

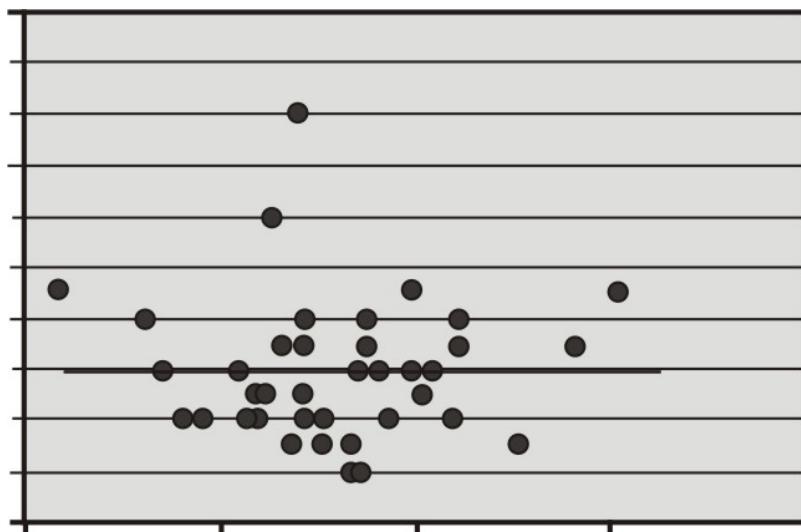
- Make inferences about regression models, including hypothesis testing for linear relationships.
- Check regression assumptions.

Introduction

In the previous section, we learned about the least-squares model, or the linear regression model. The linear regression model uses the concept of correlation to help us predict the score of a variable based on our knowledge of the score of another variable. In this section, we will investigate several inferences and assumptions that we can make about the linear regression model.

Hypothesis Testing for Linear Relationships

Let's think for a minute about the relationship between correlation and the linear regression model. As we learned, if there is no correlation between the two variables X and Y , then it would be nearly impossible to fit a meaningful regression line to the points on a scatterplot graph. If there was no correlation, and our correlation value, or r -value, was 0, we would always come up with the same predicted value, which would be the mean of all the y -values, which is \bar{y} . The figure below shows an example of what a regression line fit to variables with no correlation ($r = 0$) would look like. As you can see, for any value of X , we always get the same predicted value of Y .



Using this knowledge, we can determine that if there is no relationship between X and Y , constructing a regression line doesn't help us very much, because, again, the predicted score would always be the same. Therefore, when we estimate a linear regression model, we want to ensure that the slope, β , for the population does not equal zero. Furthermore, it is beneficial to test how strong (or far away) from zero the slope must be to strengthen our prediction of the Y scores.

In hypothesis testing of linear regression models, the null hypothesis to be tested is that the slope, β , equals zero. Our alternative hypothesis is that our slope does not equal zero.

$$\begin{aligned} H_0 &: \beta = 0 \\ H_a &: \beta \neq 0 \end{aligned}$$

The null hypothesis is saying, essentially, that there is no linear relationship between the explanatory and response variables. The alternative hypothesis is saying that the slope is significantly different from 0, and therefore, the linear relationship exists and can be useful for prediction.

The test statistic for this hypothesis test is calculated as follows:

$$t = \frac{b - \beta}{s_b}$$

where $s_b = \frac{s}{\sqrt{\sum(x - \bar{x})^2}} = \frac{s}{\sqrt{SS_X}}$,

$$s = \sqrt{\frac{SSE}{n-2}}, \text{ and}$$

$SSE = \text{sum of residual error squared}$

Example: Let's say that a football coach is using the results from a short physical fitness test to predict the results of a longer, more comprehensive one. He developed the regression equation $\hat{y} = 1.22 + 0.635x$, and the standard error of estimate is 0.56. The summary statistics are as follows:

Summary statistics for two foot ball fitness tests.

$n = 24$	$\sum xy = 591.50$
$\sum x = 118$	$\sum y = 104.3$
$\bar{x} = 4.92$	$\bar{y} = 4.35$
$\sum x^2 = 704$	$\sum y^2 = 510.01$
$SS_X = 123.83$	$SS_Y = 56.74$

Using $\alpha = 0.05$, test the null hypothesis that, in the population, the regression coefficient is zero, or $H_0 : \beta = 0$.

We use the t -distribution to calculate the test statistic and find that the critical values in the t -distribution at 22 degrees of freedom are 2.074 standard scores above and below the mean. Also, the test statistic can be calculated as follows:

$$\begin{aligned} s_b &= \frac{0.56}{\sqrt{123.83}} = 0.05 \\ t &= \frac{0.635 - 0}{0.05} = 12.70 \end{aligned}$$

Since the observed value of the test statistic exceeds the critical value, the null hypothesis would be rejected, and we can conclude that if the null hypothesis were true, we would observe a slope of 0.635 by chance less than 5% of the time.

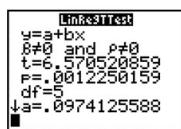
Using the Graphing Calculator to Test a Hypothesis about the Slope

Let's use the SAT versus GPA data from the previous section as an example. The raw data are already entered into L₁ and L₂. We now want to conduct the hypothesis test about the slope of the regression line, which we previously calculated as 0.0055.

Press [STAT], then scroll right to [TESTS]. Scroll down to Option F LinReg T Test and press [ENTER]. Make appropriate entries into your calculator so that it looks like this:



Scroll down to Calculate and press [ENTER]. Your new screen will look like this:



We see that for the alternative hypothesis $H_a : \beta \neq 0$ our t test statistic is 6.57. Our p-value is .001, which is highly significant because it is less than .05 and even less than .01. Therefore we reject the null hypothesis that the slope is equal to 0. There is a significant linear relationship between the explanatory variable (SAT score) and response variable (GPA).

Regression Assumptions

We make several assumptions under a linear regression model, including:

At each value of X , there is a distribution of Y . These distributions have a mean centered at the predicted value and a standard error that is calculated using the sum of squares.

Using a regression model to predict scores only works if the regression line is a good fit to the data. If this relationship is non-linear, we could either transform the data (i.e., a logarithmic transformation) or try one of the other regression equations that are available with Excel or a graphing calculator.

The standard deviations and the variances of each of these distributions for each of the predicted values are equal. This is called *homoscedasticity*.

Finally, for each given value of X , the values of Y are independent of each other.

Lesson Summary

When we estimate a linear regression model, we want to ensure that the regression coefficient (slope) for the population, β does not equal zero. To do this, we perform a hypothesis test, where we set the regression coefficient equal to zero in the null hypothesis and then test for significance.

We make several assumptions when dealing with a linear regression model. If they are not met, a transformation of the data might be indicated, or an alternate regression method might be performed.

Review Questions

1. The following data refer to the hardness of an alloy when 8 samples were treated at 500 degrees C. for different periods of time (in hours).

TABLE 9.10:

Time (in hours)	Hardness (in Rockwell units)
6	50
9	52
12	60
15	59
15	63
18	64
21	71

-
- Sketch a scatterplot of the data.
 - What is the equation of the least-squares regression equation?
 - Predict the hardness for a sample that is treated for 20 hours.
 - Perform the hypothesis test of no linear relationship between time and hardness. Use alpha of 0.05.
 - Based on the results of your hypothesis test, what can be concluded?

Answers: (1)(a) see detailed answers (1)(b) $\hat{y} = 41.38 + 1.35 x$ (1)(c) 68.38 (1)(d) $t = 8.17$ $p\text{-value} = 0.00045$ (1)(e) significant liner relationship

9.4 References

1. . . CC BY-NC-SA
2. CK-12 Foundation. . CCSA
3. CK-12 Foundation. . CCSA
4. CK-12 Foundation. . CCSA
5. CK-12 Foundation. . CCSA
6. CK-12 Foundation. . CCSA

CHAPTER

10

Chi-Square

Chapter Outline

- 10.1 THE GOODNESS-OF-FIT TEST**
 - 10.2 TEST OF INDEPENDENCE**
-

10.1 The Goodness-of-Fit Test

Learning Objectives

- Know that the chi-square distribution is a family of curves defined by the degrees of freedom.
- Identify the conditions which must be satisfied when using the chi-square test.
- Evaluate a hypothesis using the goodness-of-fit test.

Introduction

In previous lessons, we learned that there are several different tests that we can use to analyze data and test hypotheses. The type of test that we choose depends on the data available and what question we are trying to answer. We analyze simple descriptive statistics, such as the mean, median, mode, and standard deviation to give us an idea of the distribution and to remove outliers, if necessary. We calculate probabilities to determine the likelihood of something happening. Finally, we use regression analysis to examine the relationship between two or more continuous variables.

However, there is another test that we have yet to cover. To analyze patterns between distinct categories, such as genders, political candidates, locations, or preferences, we use the chi-square test. Up to this point, every hypothesis test has used numerical data, but the chi-square test gives us a powerful tool for analyzing **categorical data**.

This test is used when estimating how closely a sample matches the expected distribution (also known as the goodness-of-fit test) and when estimating if two random variables are independent of one another (also known as the test of independence).

In this lesson, we will learn more about the **goodness-of-fit test** and how to create and evaluate hypotheses using this test.

The Chi-Square Distribution

The *chi-square distribution* can be used to perform the *goodness-of-fit test*, which compares the observed values of a categorical variable with the expected values of that same variable.

Example: We would use the chi-square goodness-of-fit test to evaluate if there was a preference in the type of lunch that 11th grade students bought in the cafeteria. For this type of comparison, it helps to make a table to visualize the problem. We could construct the following table, known as a *contingency table*, to compare the observed and expected values.

Research Question: Do 11th grade students prefer a certain type of lunch? Notice that the categories are now categorical (salad, sub, daily special, homemade), and we are counting only how many students fit into each of these categories.

Using a sample of 11th grade students, we recorded the following information:

TABLE 10.1: Frequency of Type of School Lunch Chosen by Students

Type of Lunch	Observed Frequency (O)	Expected Frequency (E)
Salad	21	25

TABLE 10.1: (continued)

Type of Lunch	Observed Frequency (O)	Expected Frequency (E)
Sub Sandwich	29	25
Daily Special	14	25
Homemade	36	25

If there is no difference in which type of lunch is preferred, we would expect the students to prefer each type of lunch equally. To calculate the expected frequency of each category when assuming school lunch preferences are distributed equally, we divide the number of observations by the number of categories. Since there are 100 observations and 4 categories, the expected frequency of each category is $\frac{100}{4}$, or 25.

The value that indicates the comparison between the observed and expected frequency is called the **chi-square statistic**. The idea is that if the observed frequency is close to the expected frequency, then the chi-square statistic will be small. On the other hand, if there is a substantial difference between the two frequencies, then we would expect the chi-square statistic to be large.

To calculate the chi-square statistic, χ^2 , we use the following formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where:

χ^2 is the chi-square test statistic.

O is the observed frequency value for each event.

E is the expected frequency value for each event.

We compare the value of the test statistic to a tabled chi-square value to determine the probability that a sample fits an expected pattern.

Features of the Goodness-of-Fit Test

As mentioned, the goodness-of-fit test is used to determine patterns of distinct categorical variables. The test requires that the data are obtained through a random sample. The number of *degrees of freedom* associated with a particular chi-square test is equal to the number of categories minus one. That is, $df = c - 1$.

Example: Using our example about the preferences for types of school lunches, we calculate the degrees of freedom as follows:

$$df = \text{number of categories} - 1$$

$$3 = 4 - 1$$

There are many situations that use the goodness-of-fit test, including surveys, taste tests, genetics, and analysis of behaviors. Interestingly, goodness-of-fit tests are also used in casinos to determine if there is cheating in games of chance, such as cards or dice. For example, if a certain card or number on a die shows up more than expected (a high observed frequency compared to the expected frequency), officials use the goodness-of-fit test to determine the likelihood that the player may be cheating or that the game may not be fair.

Evaluating Hypotheses Using the Goodness-of-Fit Test

Let's use our original example to create and test a hypothesis using the goodness-of-fit chi-square test. First, we will need to state the null and alternative hypotheses for our research question. Since our research question asks, "Do

11th grade students prefer a certain type of lunch?" our null hypothesis for the chi-square test would state that there is no difference between the observed and the expected frequencies. Therefore, our alternative hypothesis would state that there is a significant difference between the observed and expected frequencies.

Null Hypothesis

$H_0 : O = E$ (There is no statistically significant difference between observed and expected frequencies.)

Alternative Hypothesis

$H_a : O \neq E$ (There is a statistically significant difference between observed and expected frequencies.)

Also, the number of degrees of freedom for this test is 3.

Using an alpha level of 0.05, we look under the column for 0.05 and the row for degrees of freedom, which, again, is 3. According to the standard chi-square distribution table, we see that the critical value for chi-square is 7.815. Therefore, we would reject the null hypothesis if the chi-square statistic is greater than 7.815.

Note that we can calculate the chi-square statistic with relative ease.

TABLE 10.2: Frequency Which Student Select Type of School Lunch

Type of Lunch	Observed Frequency	Expected Frequency	$\frac{(O-E)^2}{E}$
Salad	21	25	0.64
Sub Sandwich	29	25	0.64
Daily Special	14	25	4.84
Brought Own Lunch	36	25	4.84
Total (chi-square)			10.96

Since our chi-square statistic of 10.96 is greater than 7.815, we reject the null hypotheses and accept the alternative hypothesis. Therefore, we can conclude that there is a significant difference between the types of lunches that 11th grade students prefer.

Lesson Summary

We use the chi-square test to examine patterns between **categorical** variables, such as genders, political candidates, locations, or preferences.

There are two types of chi-square tests: the goodness-of-fit test and the test for independence. We use the **goodness-of-fit** test to estimate how closely a sample matches the expected distribution.

To test for significance, it helps to make a table detailing the observed and expected frequencies of the data sample. Using the standard chi-square distribution table, we are able to create criteria for accepting the null or alternative hypotheses for our research questions.

To test the null hypothesis, it is necessary to calculate the chi-square statistic, χ^2 . To calculate the chi-square statistic, we use the following formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

χ^2 is the chi-square test statistic.

O is the observed frequency value for each event.

E is the expected frequency value for each event.

Using the chi-square statistic and the level of significance, we are able to determine whether to reject or fail to reject the null hypothesis and write a summary statement based on these results.

Review Questions

1. What is the name of the statistical test used to analyze the patterns between two categorical variables?
 - a. Student's t -test
 - b. the ANOVA test
 - c. the chi-square test
 - d. the z -score
2. There are three types of chi-square tests. Which type of chi-square test estimates how closely a sample matches an expected distribution?
 - a. the goodness-of-fit test
 - b. the test for independence
 - c. the test for homogeneity
3. Which of the following is considered a categorical variable?
 - a. income
 - b. gender
 - c. height
 - d. weight
4. If there were 250 observations in a data set and 2 uniformly distributed categories that were being measured, the expected frequency for each category would be:
 - a. 125
 - b. 500
 - c. 250
 - d. 5
5. What is the formula for calculating the chi-square statistic?
6. A principal is planning a field trip. She samples a group of 99 students to see if they prefer a sporting event, a play at the local college, or a science museum. She records the following results:

TABLE 10.3:

Type of Field Trip	Number Preferring
Sporting Event	42
Play	26
Science Museum	31

- (a) What is the observed frequency value for the Science Museum category?
- (b) What is the expected frequency value for the Sporting Event category?
- (c) State the null and alternative hypotheses for the scenario. Use both symbols and words.
- (d) Create a chart for the observed and expected counts for each category. Then calculate the chi-square statistic for the research question above.
- (e) What is your decision (reject or not reject)? State your conclusion in terms of the original problem.

Answers: (1) A (2) A (3) B (4) A (5) see detailed answers (6)(a) 31 (6)(b) 33 (6)(c) see detailed

answers (6)(d) test statistic is 4.06 (6)(e) do not reject

10.2 Test of Independence

Learning Objectives

- Identify the data items needed to perform calculations when performing the chi-square test from contingency tables.
- Conduct the test of independence to determine whether two variables are independent or not.
- Conduct the test of homogeneity to examine the proportions of a variable attributed to different populations.

Introduction

As mentioned in the previous lesson, the chi-square test can be used to both estimate how closely an observed distribution matches an expected distribution (the goodness-of-fit test) and to estimate whether two random variables are independent of one another (the test of independence). In this lesson, we will examine the test of independence in greater detail.

The chi-square test of independence is used to assess if two factors are related. This test is often used in social science research to determine if factors are independent of each other. For example, we would use this test to determine relationships between voting patterns and race, income and gender, and behavior and education.

In general, when running the test of independence, we ask, “Is Variable X independent of Variable Y ?” It is important to note that this test does not test how the variables are related, just simply whether or not they are independent of one another. For example, while the test of independence can help us determine if income and gender are independent, it cannot help us assess how one category might affect the other.

Drawing Data from Contingency Tables Needed to Perform Calculations when Running a Chi-Square Test

Contingency tables can help us frame our hypotheses and solve problems. Often, we use contingency tables to list the variables and observational patterns that will help us to run a chi-square test. For example, we could use a contingency table to record the answers to phone surveys or observed behavioral patterns.

Example: We would use a contingency table to record the data when analyzing whether women are more likely to vote for a Republican or Democratic candidate when compared to men. In this example, we want to know if voting patterns are independent of gender. Hypothetical data for 76 females and 62 males from the state of California are in the contingency table below.

TABLE 10.4: Frequency of California Citizens voting for a Republican or Democratic Candidate

	Democratic	Republican	Total
Female	48	28	76
Male	36	26	62
Total	84	54	138

Similar to the chi-square goodness-of-fit test, the *test of independence* is a comparison of the differences between observed and expected values. However, in this test, we need to calculate the expected value using the row and column totals from the table. The expected value for each of the potential outcomes in the table can be calculated

using the following formula:

$$\text{Expected Frequency} = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$$

In the table above, we calculated the row totals to be 76 females and 62 males, while the column totals are 84 Democrats and 54 Republicans. Using the formula, we find the following expected frequencies for the potential outcomes:

The expected frequency for female Democratic outcome is $\frac{(76)(84)}{138} = 46.26$

The expected frequency for female Republican outcome is $\frac{(76)(54)}{138} = 29.74$.

The expected frequency for male Democratic outcome is $\frac{(62)(84)}{138} = 37.74$.

The expected frequency for male Republican outcome is $\frac{(62)(54)}{138} = 24.26$.

Using these calculated expected frequencies, we can modify the table above to look something like this:

TABLE 10.5:

	Democratic Observed	Democratic Expected	Republican Observed	Republican Expected	Total
Female	48	46.26	28	29.74	76
Male	36	37.74	26	24.26	62
Total	84		54		138

With the figures above, we are able to calculate the chi-square statistic with relative ease.

The Chi-Square Test of Independence

When running the test of independence, we use similar steps as when running the goodness-of-fit test described earlier. First, we need to establish a hypothesis based on our research question. Using our scenario of gender and voting patterns, our null hypothesis is that there is not a significant difference in the frequencies with which females vote for a Republican or Democratic candidate when compared to males. Therefore, our hypotheses can be stated as follows:

Null Hypothesis

$H_0 : O = E$ (There is no statistically significant difference between the observed and expected frequencies.)

Alternative Hypothesis

$H_a : O \neq E$ (There is a statistically significant difference between the observed and expected frequencies.)

Using the table above, we can calculate the degrees of freedom and the chi-square statistic. The formula for calculating the chi-square statistic is the same as before:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

χ^2 is the chi-square test statistic.

O is the observed frequency value for each event.

E is the expected frequency value for each event.

Using this formula and the example above, we get the following expected frequencies and chi-square statistic:

TABLE 10.6:

	Democratic	Democratic	Democratic	Republican	Republican	Republican
	Obs. Freq.	Exp. Freq.	$\frac{(O-E)^2}{E}$	Obs. Freq.	Exp. Freq.	$\frac{(O-E)^2}{E}$
Female	48	46.26	0.07	28	29.74	0.10
Male	36	37.74	0.08	26	24.26	0.12
Totals	84			54		

$$\chi^2 = 0.07 + 0.08 + 0.10 + 0.12 = 0.37$$

Also, the degrees of freedom can be calculated as follows:

$$\begin{aligned} df &= (C - 1)(R - 1) \\ &= (2 - 1)(2 - 1) = 1 \end{aligned}$$

With an alpha level of 0.05, we look under the column for 0.05 and the row for degrees of freedom, which, again, is 1, in the standard chi-square distribution table (<http://tinyurl.com/3ypvj2h>). According to the table, we see that the critical value for chi-square is 3.841. Therefore, we would reject the null hypothesis if the chi-square statistic is greater than 3.841.

Since our calculated chi-square value of 0.37 is less than 3.841, we **fail to reject** the null hypothesis. Therefore, we can conclude that females are not significantly more likely to vote for a Republican or Democratic candidate than males. In other words, these two factors appear to be **independent** of one another.

Test of Homogeneity

The chi-square goodness-of-fit test and the test of independence are two ways to examine the relationships between categorical variables. To determine whether or not the assignment of categorical variables is random (that is, to examine the randomness of a sample), we perform the **test of homogeneity**. In other words, the test of homogeneity tests whether samples from populations have the same proportion of observations with a common characteristic. For example, we found in our last test of independence that the factors of gender and voting patterns were independent of one another. However, our original question was if females were more likely to vote for a Republican or Democratic candidate when compared to males. We would use the test of homogeneity to examine the probability that choosing a Republican or Democratic candidate was the same for females and males. Here we are focusing on **gender differences** instead of focusing on the independence of gender and voting.

Another commonly used example of the test of homogeneity is comparing dice to see if they all work the same way.

Example: The manager of a casino has two potentially loaded dice that he wants to examine. (Loaded dice are ones that are weighted on one side so that certain numbers have greater probabilities of showing up.) The manager rolls each of the dice exactly 20 times and comes up with the following results:

TABLE 10.7: Number Rolled with the Potentially Loaded Dice

	1	2	3	4	5	6	Totals
Die 1	6	1	2	2	3	6	20
Die 2	4	1	3	3	1	8	20
Totals	10	2	5	5	4	14	40

Like the other chi-square tests, we first need to establish a null hypothesis based on a research question. In this case, our research question would be something like, “Is the probability of rolling a specific number the same for Die 1 and Die 2?” This would give us the following hypotheses:

Null Hypothesis

$$H_0 : O = E \text{ (The probabilities are the same for both dice.)}$$

Alternative Hypothesis

$$H_a : O \neq E \text{ (The probabilities differ for both dice.)}$$

Similar to the test of independence, we need to calculate the expected frequency for each potential outcome and the total number of degrees of freedom. To get the expected frequency for each potential outcome, we use the same formula as we used for the test of independence, which is as follows:

$$\text{Expected Frequency} = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$$

The following table includes the expected frequency (in parentheses) for each outcome, along with the chi-square statistic, $\chi^2 = \frac{(O-E)^2}{E}$, in a separate column:

Number Rolled on the Potentially Loaded Dice

TABLE 10.8:

	1	χ^2	2	χ^2	3	χ^2	4	χ^2	5	χ^2	6	χ^2	χ^2 Total
Die 1	6(5)	0.2	1(1)	0	2(2.5)	0.1	2(2.5)	0.1	3(2)	0.5	6(7)	0.14	1.04
Die 2	4(5)	0.2	1(1)	0	3(2.5)	0.1	3(2.5)	0.1	1(2)	0.5	8(7)	0.14	1.04
Totals	10		2		5		5		4		14		2.08

$$df = (C - 1)(R - 1) \\ = (6 - 1)(2 - 1) = 5$$

From the table above, we can see that the value of the test statistic is 2.08.

Using an alpha level of 0.05, we look under the column for 0.05 and the row for degrees of freedom, which, again, is 5, in the standard chi-square distribution table. According to the table, we see that the critical value for chi-square is 11.070. Therefore, we would reject the null hypothesis if the chi-square statistic is greater than 11.070.

Since our calculated chi-square value of 2.08 is less than 11.070, we fail to reject the null hypothesis. Therefore, we can conclude that each number is just as likely to be rolled on one die as on the other. This means that if the dice are loaded, they are probably loaded in the same way or were made by the same manufacturer.

Lesson Summary

The chi-square test of independence is used to assess if two **categorical** factors are related. It is commonly used in social science research to examine behaviors, preferences, measurements, etc.

As with the chi-square goodness-of-fit test, contingency tables help capture and display relevant information. For each of the possible outcomes in the table constructed to run a chi-square test, we need to calculate the expected frequency. The formula used for this calculation is as follows:

$$\text{Expected Frequency} = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$$

To calculate the chi-square statistic for the test of independence, we use the same formula as for the goodness-of-fit test. If the calculated chi-square value is greater than the critical value, we reject the null hypothesis.

We perform the test of homogeneity to examine the randomness of a sample. The test of homogeneity tests whether various populations are homogeneous or different with respect to certain characteristics.

Review Questions

1. What is the chi-square test of independence used for?
2. True or False: In the test of independence, you can test if two variables are related, but you cannot test the nature of the relationship itself.
3. When calculating the expected frequency for a possible outcome in a contingency table, you use the formula:
 - a. Expected Frequency = $\frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$
 - b. Expected Frequency = $\frac{(\text{Total Observations})(\text{Column Total})}{\text{Row Total}}$
 - c. Expected Frequency = $\frac{(\text{Total Observations})(\text{Row Total})}{\text{Column Total}}$
4. Use the table below to answer the following review questions.

TABLE 10.9: Research Question: Is Studying Abroad Independent of Gender?

	Studied Abroad	Did Not Study Abroad
Females	322	460
Males	128	152

- (a) What is the total number of females in the sample?
- (b) What is the total number of observations in the sample?
- (c) What is the expected frequency for the number of males who did not study abroad?
- (d) How many degrees of freedom are in this example?
- (e) What are the null and alternative hypotheses, in words?
- (f) What is the value of the chi-square test statistic for this example? What is your statistical decision? What is your conclusion?
5. If data were collected and analyzed, and the chi-square critical value at a significance level of 0.05 and 1 degree of freedom were given as 3.81, and we calculated a chi-square test statistic of 6.13, we would:
 - A. reject the null hypothesis
 - B. fail to reject the null hypothesis
6. The test of homogeneity is conducted the same way as
 - A. the goodness-of-fit test
 - B. the test of independence
7. Researchers at two hospitals, one in the U.S. and the other in Iceland, randomly select 100 patient records, noting the blood type of each patient. They summarize their results in the following 2-way table.

TABLE 10.10:

Type O	Type A	Type B	Type AB	Totals
--------	--------	--------	---------	--------

TABLE 10.10: (continued)

U.S.	45	40	11	4
Iceland	56	32	9	3
Totals				

The researchers want to determine if the distribution of blood types is the same for citizens of these 2 countries. Therefore, they will conduct a chi-square test of homogeneity.

- a. State the hypotheses using both symbols and words.
- b. Calculate each expected frequency, and then calculate the chi-square test statistic.
- c. How many degrees of freedom will be used for this procedure?
- d. Using the chi-square table of critical values, what is the critical value for this test? What is your statistical decision? What do you conclude about the distribution of blood types in these 2 countries?

Answers to selected problems: (see detailed answers for complete solutions). (2) True (3) A (4)(a) 782 (4)(b) 1062 (4)(c) test statistic = 1.76 (5) A (6) B (7)(b) test statistic = 2.36 (7)(c) 3 (7)(d) do not reject