CK-12 Advanced Probability and Statistics

Second Edition

CK-12 Probability and Statistics - Advanced (Second Edition)

Ellen Lawsky

Larry Ottman

Raja Almukkahal

Brenda Meery

Danielle DeLancey

cK-12 *f*lexbooK

next generation textbooks

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

AUTHORS
Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

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®**.

Printed: December 17, 2014

Contents

CHAPTER **1**

An Introduction to Analyzing Statistical Data

Chapter Outline

1.1 Definitions of Statistical Terminology

Learning Objectives

- Distinguish between quantitative and categorical variables.
- Understand the concept of a population and the reason for using a sample.
- Distinguish between a statistic and a parameter.

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, Scologia 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 character-istics 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

(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. {{Problem | question=In each of the following situations, identify the population, the units, and each variable, and tell if the variable is categorical or quantitative.
   1. A quality control worker with Sweet-Tooth Candy weighs every 100$^{\text{th}}$ candy bar to make sure it is very close to the published weight.
2.

<div align="center">

POPULATION:

UNITS:

VARIABLE:

TYPE:

</div>

   1. Doris decides to clean her sock drawer out and sorts her socks into piles by color.
3.

<div align="center">

POPULATION:

UNITS:

VARIABLE:

TYPE:

</div>

   1. 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.

4.

POPULATION:

UNITS:

VARIABLE:

TYPE:

5. In Physical Education class, the teacher has the students count off by two's to divide them into teams. Is this a categorical or quantitative variable?

6. 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.

On the Web

http://www.onlinestatbook.com/

http://www.en.wikipedia.org/wiki/Gal%C3%A1pagos_tortoise

Galapagos Conservancy:

Charles Darwin Research Center and Foundation: http://www.darwinfoundation.org

1.2 An Overview of Data

Learning Objective

- Understand the difference between the levels of measurement: nominal, ordinal, interval, and 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.

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.

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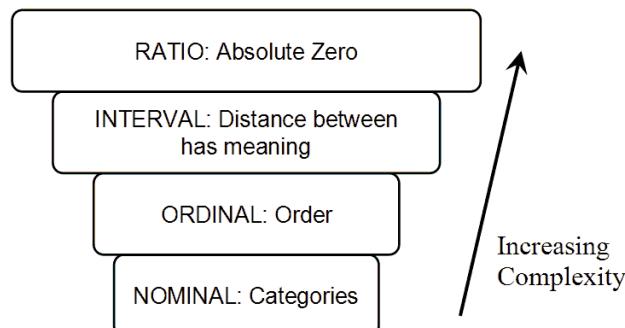
If we collect information about the students' gender, race, political opinions, or the town or sub-division in which they live, we have a nominal measurement.

If we collect data about the students' year in school, we are now ordering that data numerically (9th, 10th, 11th, or 12th 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 an absolute 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 collects the same data, only this time using degrees Kelvin.

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.

3. Look at Table 1.1 in Section 1. What is the highest level of measurement that could be correctly applied to the variable 'Population Density'?

   a. Nominal
   b. Ordinal
   c. Interval
   d. Ratio

Note: If you are curious about the "does not apply" in the last row of Table 1.1, read on! There is only one known individual Pinta tortoise, and he lives at the Charles Darwin Research station. He is affectionately known as Lonesome George. He is probably well over 100 years old and will most likely signal the end of the species, as attempts to breed have been unsuccessful.

On the Web

Levels of Measurement:

http://en.wikipedia.org/wiki/Level_of_measurement

http://www.socialresearchmethods.net/kb/measlevl.php

Peter and Rosemary Grant: http://en.wikipedia.org/wiki/Peter_and_Rosemary_Grant

1.3 Measures of Center

Learning Objectives

- Calculate the mode, median, and mean for a set of data, and understand the differences between each measure of center.
- Identify the symbols and know the formulas for sample and population means.
- Determine the values in a data set that are outliers.
- Identify the values to be removed from a data set for an *n%* trimmed mean.
- Calculate the midrange, 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.
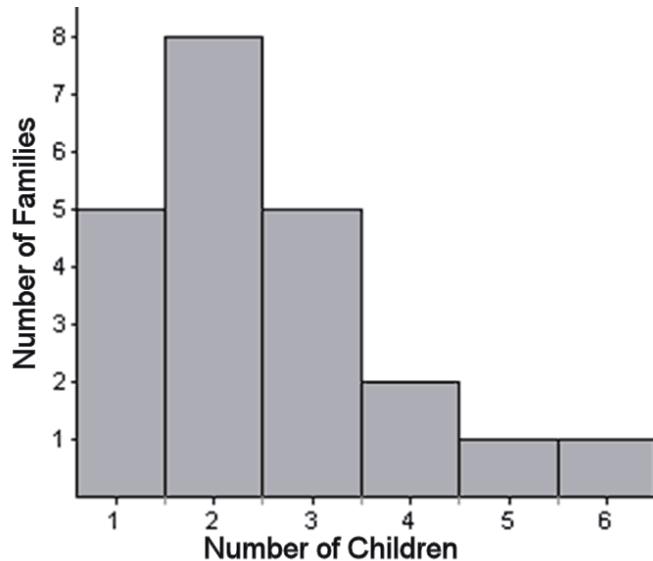
If there were seven 3-child households and seven 2-child households, we would say the data set has two modes. In other words, the data would be *bimodal*. When a data set is described as being bimodal, it is clustered about two different modes. Technically, if there were more than two, they would all 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.
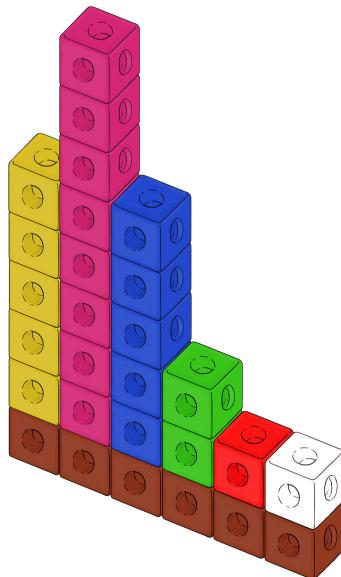
Mean

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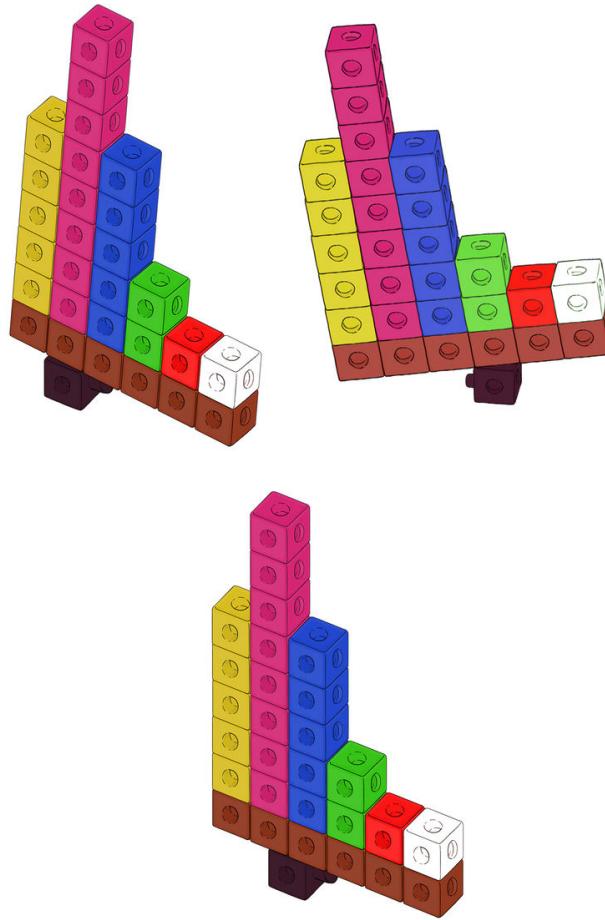
If you have snap cubes like you used to use in elementary school, you can make a physical model of the graph, using one cube to represent each student's family and a row of six cubes at the bottom to hold them together, like this:

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:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where:

x_i is the ith data value of the sample.

n is the sample size.

The mean of the population is denoted by the Greek letter, μ.

$\bar{x}$ 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 grades:

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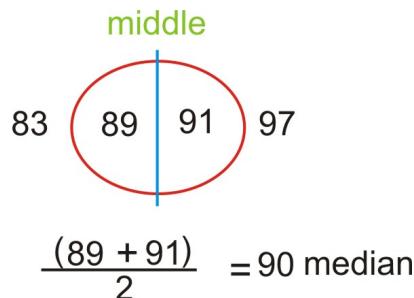
The middle number in this case is the third grade, or 90, so the median of this data is 90.

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.

middle

83 (89 | 91) 97

$$\frac{(89 + 91)}{2} = 90 \text{ median}$$

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 (see http://www.forbes.com/static/execpay2005/rank.html for more). 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

Midrange

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

Example: Consider the following quiz grades: 75, 80, 90, 94, and 96. The midrange would be:

$$\frac{75+96}{2} = \frac{171}{2} = 85.5$$

Since it is based on only the two most extreme values, the midrange is not commonly used as a measure of central tendency.

Trimmed Mean

Recall that the mean is not resistant to the effects of outliers. Many students ask their teacher to "drop the lowest grade." The argument is that everyone has a bad day, and one extreme grade that is not typical of the rest of their work should not have such a strong influence on their mean grade. The problem is that this can work both ways; it could also be true that a student who is performing poorly most of the time could have a really good day (or even get lucky) and get one extremely high grade. We wouldn't blame this student for not asking the teacher to drop the highest grade! Attempting to more accurately describe a data set by removing the extreme values is referred to as trimming the data. To be fair, though, a valid trimmed statistic must remove both the extreme maximum and minimum values. So, while some students might disapprove, to calculate a *trimmed mean*, you remove the maximum and minimum values and divide by the number of values that remain.

Example: Consider the following quiz grades: 75, 80, 90, 94, 96.

A trimmed mean would remove the largest and smallest values, 75 and 96, and divide by 3.

$$\cancel{75}, 80, 90, 94, \cancel{96}$$
$$\frac{80+90+94}{3} = 88$$

Instead of removing just the minimum and maximums in a larger data set, a statistician may choose to remove a certain percentage of the extreme values. This is called an *n% trimmed mean*. To perform this calculation, remove the specified percent of the number of values from the data, half on each end. For example, in a data set that contains 100 numbers, to calculate a 10% trimmed mean, remove 10% of the data, 5% from each end. In this simplified example, the five smallest and the five largest values would be discarded, and the sum of the remaining numbers would be divided by 90.

Example: In real data, it is not always so straightforward. To illustrate this, let's return to our data from the number of children in a household and calculate a 10% trimmed mean. Here is the data set:

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

Placing the data in order yields the following:

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

Ten percent of 22 values is 2.2, so we could remove 2 numbers, one from each end (2 total, or approximately 9% trimmed), or we could remove 2 numbers from each end (4 total, or approximately 18% trimmed). Some statisticians would calculate both of these and then use proportions to find an approximation for 10%. Others might argue that 9% is closer, so we should use that value. For our purposes, and to stay consistent with the way we handle similar situations in later chapters, we will always opt to remove more numbers than necessary. The logic behind this is simple. You are claiming to remove 10% of the numbers. If you cannot remove exactly 10%, then you either have to remove more or fewer. We would prefer to err on the side of caution and remove at least the percentage reported. This is not a hard and fast rule and is a good illustration of how many concepts in statistics are open to individual interpretation. Some statisticians even say that the only correct answer to every question asked in statistics is, "It depends!"

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:

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

where:

x_i is the ith data point.

f_i is the number of times that data point occurs.

n is the number of data points.

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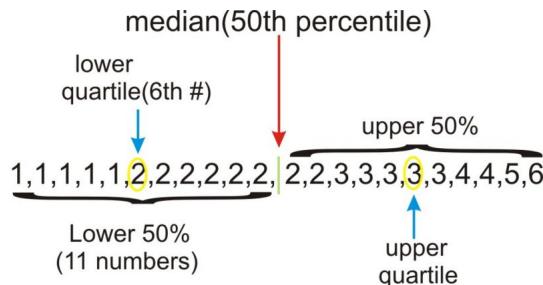
Example: To check a child's physical development, pediatricians use height and weight charts that help them to know how the child compares to children of the same age. A child whose height is in the 70th percentile is taller than 70% of children of the same age.

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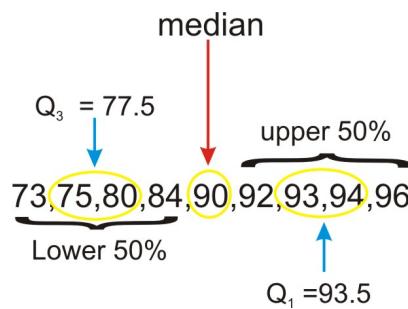
Example: Let's return to the previous data set, which is as follows:

1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 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. To read more about some alternate methods for calculating quartiles in certain situations, click on the subsequent link.

On the Web

http://mathforum.org/library/drmath/view/60969.html

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.

Another measure of center is the midrange, which is the mean of the maximum and minimum values. In an $n\%$ trimmed mean, you remove a certain n percentage of the data (half from each end) before calculating the mean. 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).
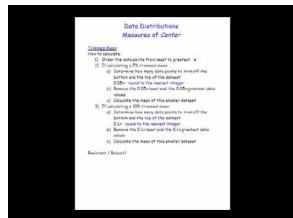
A percentile is a data value for which the specified percentage of the data is below that value. The median is the 50th percentile. Two well-known percentiles are the 25th percentile, which is called the lower quartile, Q_1, and the 75th percentile, which is called the upper quartile, Q_3.

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?

Multimedia Links

For a discussion of four measures of central tendency **(5.0)**, see American Public University, Data Distributions - Measures of a Center (6:24).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1053
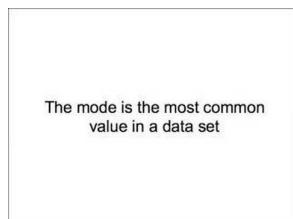
For an explanation and examples of mean, median and mode **(10.0)**, see keithpeterb, Mean, Mode and Median from Frequency Tables (7:06).

The mode is the most common value in a data set

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/53382

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. How many data points should be removed from each end of a sample of 300 values in order to calculate a 10% trimmed mean?

    a. 5
    b. 10
    c. 15
    d. 20
    e. 30

4. In the last example, after removing the correct numbers and summing those remaining, what would you divide by to calculate the mean?

5. 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
Española	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, Scologia Aplicada, Vol. 3, Num. 1,2, pp. 98-11).

For this data, calculate each of the following:

(a) mode

(b) median

(c) mean

(d) a 10% trimmed mean

(e) midrange

(f) upper and lower quartiles

(g) the percentile for the number of Santiago tortoises reintroduced

6. In the previous question, why is the answer to (c) significantly higher than the answer to (b)?

On the Web

http://edhelper.com/statistics.htm

http://en.wikipedia.org/wiki/Arithmetic_mean

Java Applets helpful to understand the relationship between the mean and the median:

http://www.ruf.rice.edu/~lane/stat_sim/descriptive/index.html

http://www.shodor.org/interactivate/activities/PlopIt/

Technology Notes: Calculating the Mean on the TI-83/84 Graphing Calculator

Step 1: Entering the data

On the home screen, press **[2ND][{]**, and then enter the following data separated by commas. When you have entered all the data, press **[2ND][}][STO][2ND][L1][ENTER]**. You will see the screen on the left below:

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

Step 2: Computing the mean

On the home screen, press **[2ND][LIST]** to enter the **LIST** menu, press the right arrow twice to go to the **MATH** menu (the middle screen above), and either arrow down and press **[ENTER]** or press **[3]** for the mean. Finally, press **[2ND][L1][)]** to insert **L1** and press **[ENTER]** (see the screen on the right above).

Calculating Weighted Means on the TI-83/84 Graphing Calculator

Use the data of the number of children in a family. In list **L1**, enter the number of children, and in list **L2**, enter the frequencies, or weights.

The data should be entered as shown in the left screen below:

Press **[2ND][STAT]** to enter the **LIST** menu, press the right arrow twice to go to the **MATH** menu (the middle screen above), and either arrow down and press **[ENTER]** or press **[3]** for the mean. Finally, press **[2ND][L1][,][2ND][L2][)][ENTER]**, and you will see the screen on the right above. Note that the mean is 2.5, as before.

Calculating Medians and Quartiles on the TI-83/84 Graphing Calculator

The median and quartiles can also be calculated using a graphing calculator. You may have noticed earlier that median is available in the **MATH** submenu of the **LIST** menu (see below).

While there is a way to access each quartile individually, we will usually want them both, so we will access them through the one-variable statistics in the **STAT** menu.

You should still have the data in **L1** and the frequencies, or weights, in **L2**, so press **[STAT]**, and then arrow over to **CALC** (the left screen below) and press **[ENTER]** or press **[1]** for '1-Var Stats', which returns you to the home screen (see the middle screen below). Press **[2ND][L1][,][2ND][L2][ENTER]** for the data and frequency lists (see third screen). When you press **[ENTER]**, look at the bottom left hand corner of the screen (fourth screen below). You will notice there is an arrow pointing downward to indicate that there is more information. Scroll down to reveal the quartiles and the median (final screen below).

Remember that Q_1 corresponds to the 25$^{\text{th}}$ percentile, and Q_3 corresponds to the 75$^{\text{th}}$ percentile.

1.4 Measures of Spread

Learning Objectives

- Calculate the range and interquartile range.
- Calculate the standard deviation for a population and a sample, and understand its meaning.
- Distinguish between the variance and the standard deviation.
- Calculate and apply Chebyshev's Theorem to any set of data.

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 largest value (maximum) and the smallest value (minimum) 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. *Source*: http://www.livescience.com/environment/070518_rainy_cities.html . The following table lists measurements of the approximate annual rainfall in Mobile over a 10 year period. Find the range and *IQR* for this data.

TABLE 1.4:

	Rainfall (inches)
1998	90

TABLE 1.4: (continued)

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

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

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 randomly selected individuals' shoe sizes 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

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.

Formulas

Sample Standard Deviation:

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

where:

x_i is the ith data value.

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

n is the sample size.

Variance of a sample:

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

where:

x_i is the ith data value.

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

n is the sample size.

Chebyshev's Theorem

Pafnuty Chebyshev was a 19th Century Russian mathematician. The theorem named for him gives us information about how many elements of a data set are within a certain number of standard deviations of the mean.

The formal statement for *Chebyshev's Theorem* is as follows:

The proportion of data points that lie within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2}, \, k > 1$$

Example: Given a group of data with mean 60 and standard deviation 15, at least what percent of the data will fall between 15 and 105?
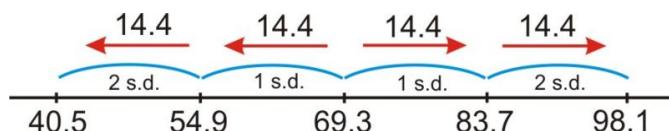
15 is three standard deviations below the mean of 60, and 105 is 3 standard deviations above the mean of 60. Chebyshev's Theorem tells us that at least $1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} \approx 0.89 = 89\%$ of the data will fall between 15 and 105.

Example: Return to the rainfall data from Mobile. The mean yearly rainfall amount is 69.3, and the sample standard deviation is about 14.4.

Chebyshev's Theorem tells us about the proportion of data within k standard deviations of the mean. If we replace k with 2, the result is as shown:

$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}$$

So the theorem predicts that at least 75% of the data is within 2 standard deviations of the mean.

According to the drawing above, Chebyshev's Theorem states that at least 75% of the data is between 40.5 and 98.1. This doesn't seem too significant in this example, because all of the data falls within that range. The advantage of Chebyshev's Theorem is that it applies to any sample or population, no matter how it is distributed.

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 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*. Chebyshev's Theorem gives us information about the minimum percentage of data that falls within a certain number of standard deviations of the mean, and it applies to any population or sample, regardless of how that data set is distributed.

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 from figure 1 to answer this question.

    a. Calculate and record the sample mean:
    b. Complete the chart to calculate the variance and the standard deviation.

TABLE 1.8:

Year	Rainfall (inches)	Deviation	Squared Deviations
1998	90		
1999	56		
2000	60		
2001	59		
2002	74		
2003	76		
2004	81		
2005	91		
2006	47		
2007	59		

Variance:

Standard Deviation:

Use the Galapagos Tortoise data below to answer questions 2 and 3.

TABLE 1.9:

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

2. Calculate the range and the *IQR* for this data.
3. Calculate the sample standard deviation for this data.
4. If $\sigma^2 = 9$, then the population standard deviation is:

   a. 3
   b. 8
   c. 9
   d. 81

5. 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

On the Web

The following links discuss various issues related to measures of spread, including 1) why the population standard deviation is calculated by dividing by the entire population N, while the standard deviation of a sample is calculated by dividing by the total sample N **minus 1**; and 2) Why the standard deviation is calculated by a rather complex process of summing **the squares** of the differences between the data points and the mean, averaging these differences, and then taking the square root of the average, rather than simply averaging the **non squared** absolute differences.

http://mathcentral.uregina.ca/QQ/database/QQ.09.99/freeman2.html

http://mathforum.org/library/drmath/view/52722.html

http://edhelper.com/statistics.htm

http://www.newton.dep.anl.gov/newton/askasci/1993/math/MATH014.HTM
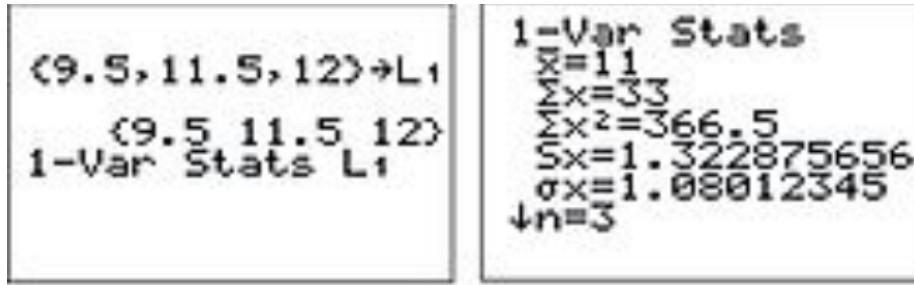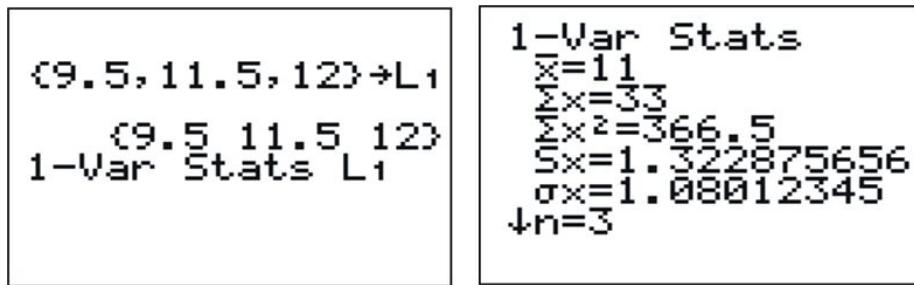
Technology Notes: Calculating Standard Deviation on the TI-83/84 Graphing Calculator

Enter the data 9.5, 11.5, 12 in list **L1** (see first screen below).

Then choose '1-Var Stats' from the **CALC** submenu of the **STAT** menu (second screen).

Enter **L1** (third screen) and press **[ENTER]** to see the fourth screen.

In the fourth screen, the symbol s_x is the sample standard deviation.

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 standard deviation can be greater than the range.
    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 in 2008:

TABLE 1.10: 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

Which month contains the median number of days of rain?

(a) January

(b) February

(c) June

(d) July

(e) September

3. 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

4. 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.

5. 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

6. 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

7. 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.

8. Sampling error is best described as:

   a. The unintentional mistakes a researcher makes when collecting information
   b. The natural variation that is present when you do not get data from the entire population
   c. A researcher intentionally asking a misleading question, hoping for a particular response
   d. When a drug company does its own experiment that proves its medication is the best
   e. When individuals in a sample answer a survey untruthfully

9. If the sum of the squared deviations for a sample of 20 individuals is 277, the standard deviation is closest to:

   a. 3.82
   b. 3.85
   c. 13.72
   d. 14.58
   e. 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

   a. 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:

      a. mode
      b. mean
      c. median
      d. upper and lower quartiles
      e. midrange
      f. range

   b. 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 ounces.

(a) Calculate the sample mean.

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

TABLE 1.11:

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.

3. 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 1.12:

Island	Approximate Area (sq. km)
Baltra	8
Darwin	1.1
Española	60
Fernandina	642
Floreana	173

TABLE 1.12: (continued)

Island	Approximate Area (sq. km)
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:

(i) mode

(ii) mean

(iii) median

(iv) upper quartile

(v) lower quartile

(vi) range

(vii) 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.

4. 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

Bimodal

Categorical variable

Census

Chebyshev's Theorem

Deviation

Interquartile range (*IQR*)

Interval

Interval estimate

Levels of measurement

Lower quartile

Mean

Mean absolute deviation

Median

Midrange

Mode

n% trimmed mean

Nominal

Numerical variable

Ordinal

Outliers

Parameter

Percentile

Point estimate

Population

Qualitative variable

Quantitative variable

Range

Ratio

Resistant

Sample

Sampling error

Standard deviation

Statistic

Trimmed mean

Unit

Upper quartile

Variables

Variance

Weighted mean

Weighted mean

CHAPTER 2

Visualizations of Data

Chapter Outline

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.

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

Here are the correct frequencies using the imaginary data presented above:

Figure: Imaginary Class Data on Water Bottle Usage

TABLE 2.1: 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

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 2.2: Frequency table using tally marks

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

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

TABLE 2.3:

Country	Liters of Bottled Water Consumed per Person per Year
Italy	183.6

TABLE 2.3: (continued)

Country	Liters of Bottled Water Consumed per Person per Year
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.

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 2.4:

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
$[160 - 170)$	2
$[170 - 180)$	0
$[180 - 190)$	1

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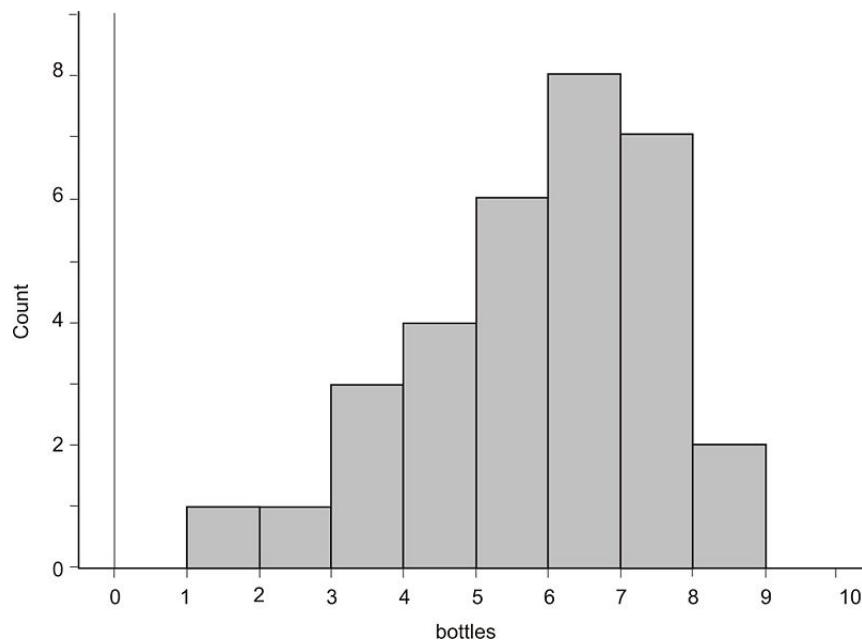
Histograms

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 2.5: 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
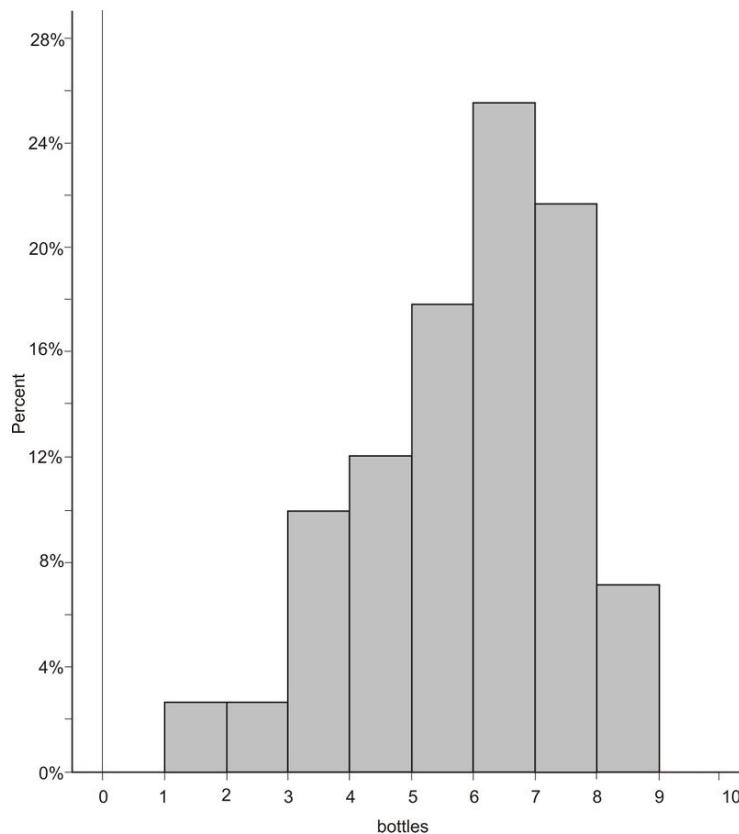
often, when you see histograms in newspapers, magazines, or online, they may instead label the midpoint of each bin. Some graphing software will also label the midpoint of each bin, unless you specify otherwise.

On the Web

http://illuminations.nctm.org/ActivityDetail.aspx?ID=78   Here you can change the bin width and explore how it effects the shape of the histogram.

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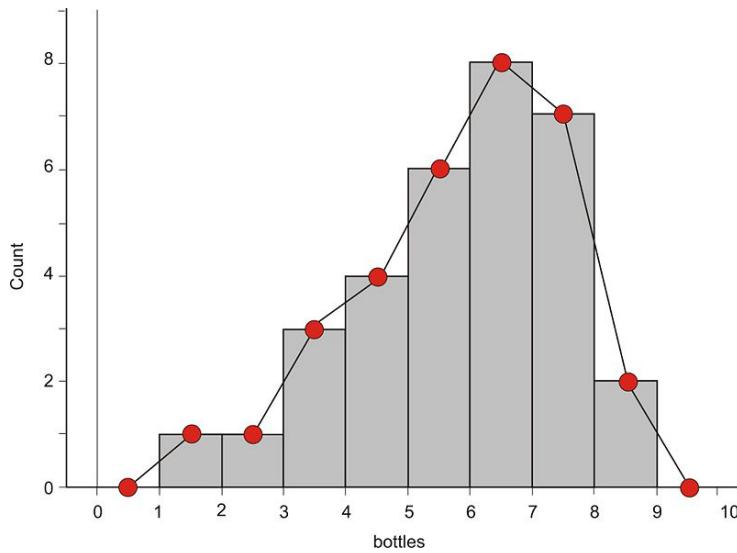
Frequency Polygons
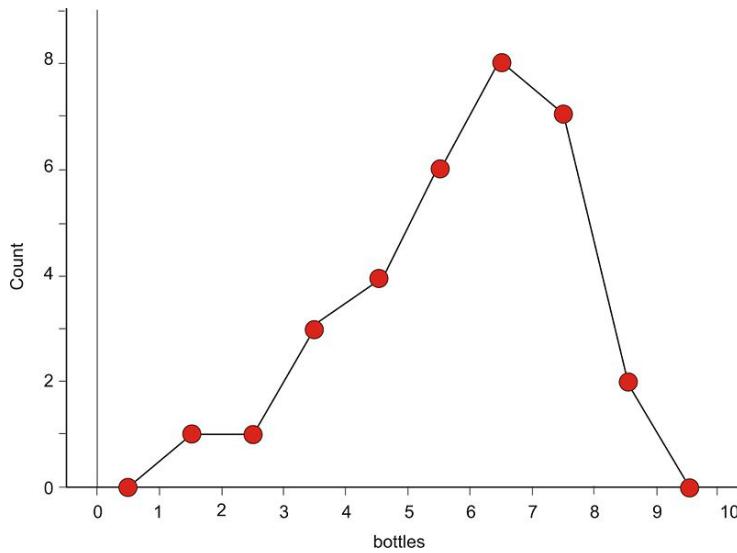
A *frequency polygon* is similar to a histogram, but instead of using bins, a polygon is created by plotting the frequencies and connecting those points with a series of line segments.

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:
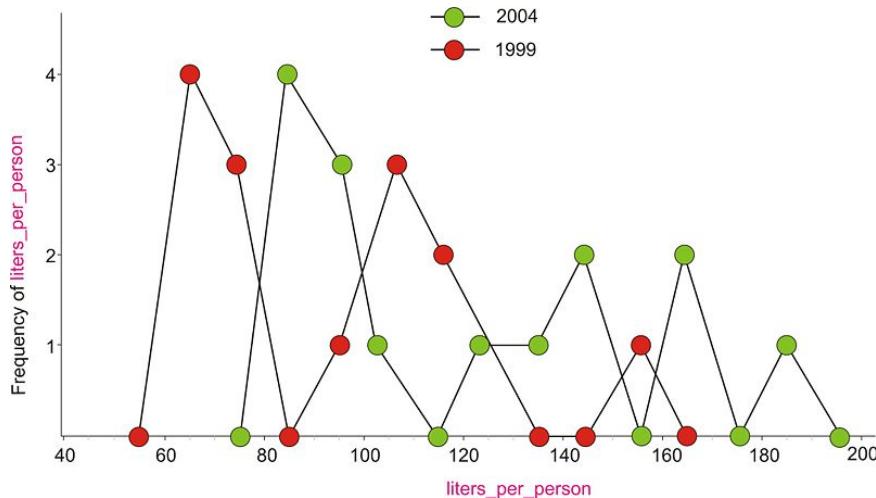
Here is the frequency polygon in finished form:

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.
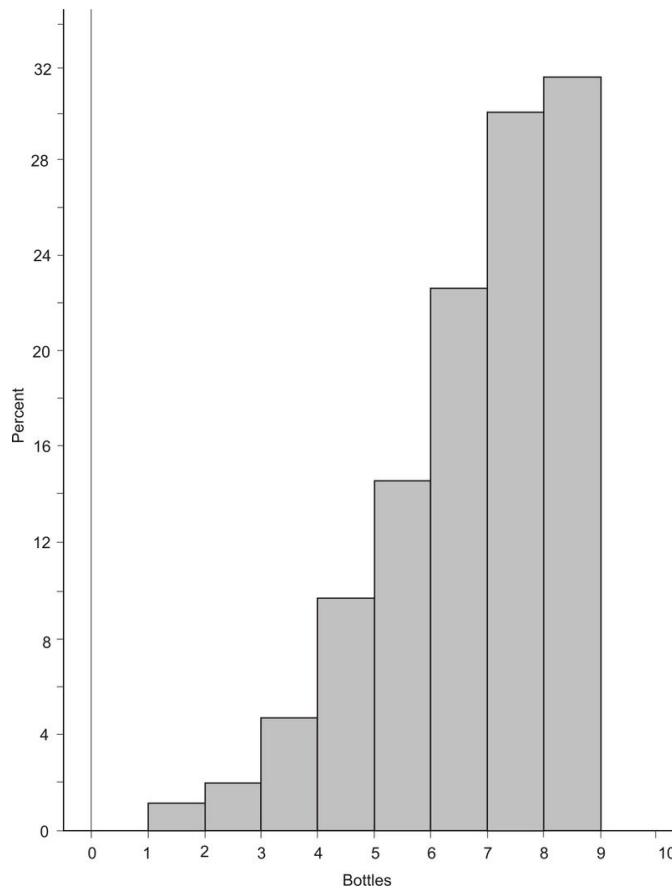
TABLE 2.6:

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:
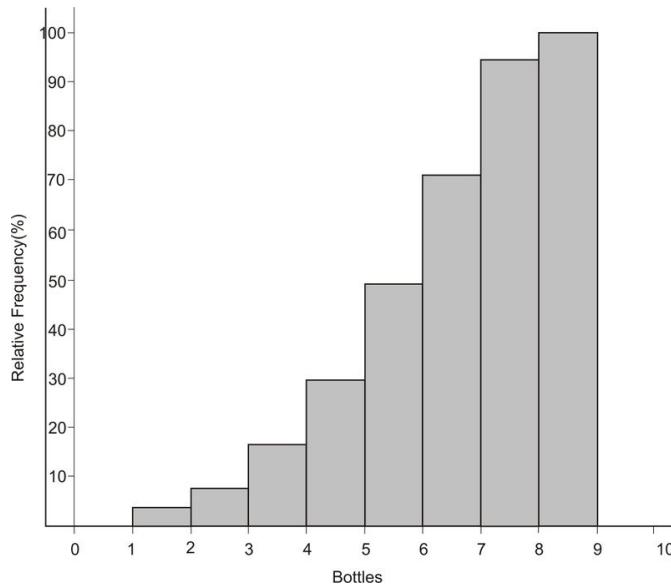
A *relative cumulative frequency histogram* would be the same, except that the vertical bars would represent the relative cumulative frequencies of the data:
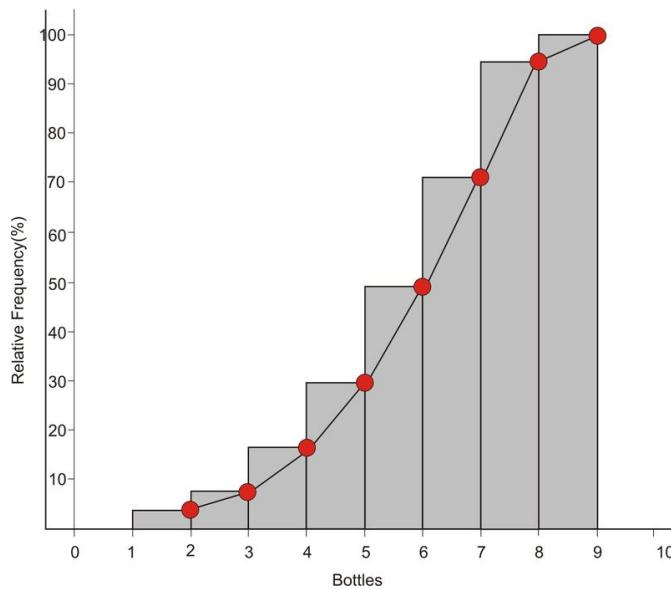
TABLE 2.7:

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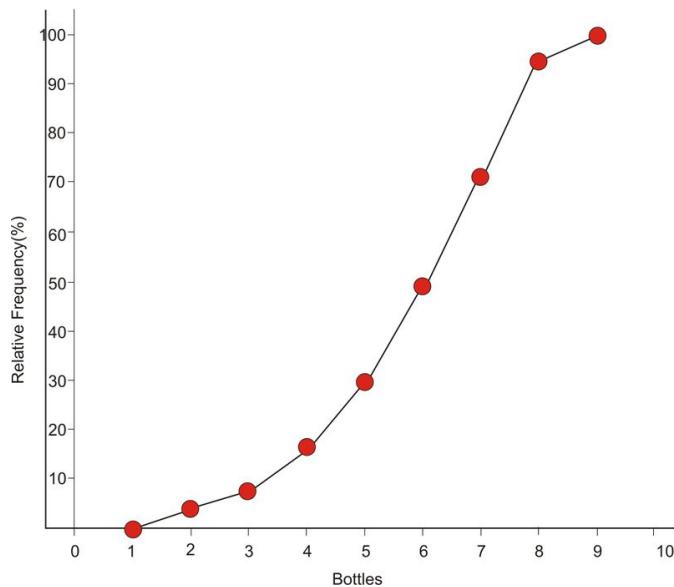
Figure: Relative Cumulative Frequency Table for Bottle Data

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.
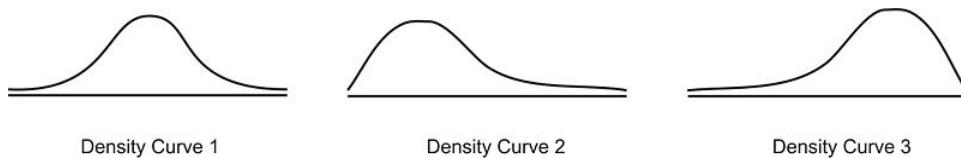
Referring to our imaginary student poll on using plastic beverage containers, we notice that the data are spread out from 0 to 9. The graph for the data illustrates this concept, and the range quantifies it. Look back at the graph and notice that there is a large concentration of students in the 5, 6, 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:

Density Curve 1   Density Curve 2   Density Curve 3

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, but she ran out of time as class was ending. When she arrived home, something had spilled in her backpack and smudged the data for the 2's. Fortunately, none of the other values was affected, and she knew there were 30 total students in the class. Complete her frequency table.

TABLE 2.8:

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

2. The following frequency table contains exactly one data value that is a positive multiple of ten. What must that value be?

TABLE 2.9:

Class	Frequency
$[0-5)$	4
$[5-10)$	0
$[10-15)$	2
$[15-20)$	1
$[20-25)$	0
$[25-30)$	3
$[30-35)$	0
$[35-40)$	1

(a) 10

(b) 20

(c) 30

(d) 40

(e) There is not enough information to determine the answer.

3. The following table includes the data from the same group of countries from the earlier bottled water consumption example, but is for the year 1999, instead.

TABLE 2.10:

Country	Liters of Bottled Water Consumed per Person per Year
Italy	154.8
Mexico	117.0
United Arab Emirates	109.8
Belgium and Luxembourg	121.9
France	117.3
Spain	101.8
Germany	100.7
Lebanon	67.8
Switzerland	90.1
Cyprus	67.4
United States	63.6
Saudi Arabia	75.3
Czech Republic	62.1
Austria	74.6
Portugal	70.4

Figure: Bottled Water Consumption per Person in Leading Countries in 1999. *Source:* http://www.earth-policy.org/Updates/2006/Update51_data.htm

(a) Create a frequency table for this data set.

(b) Create the histogram for this data set.

(c) How would you describe the shape of this data set?

4. The following table shows the potential energy that could be saved by manufacturing each type of material using the maximum percentage of recycled materials, as opposed to using all new materials.

TABLE 2.11:

Manufactured Material	Energy Saved (millions of BTU's per ton)
Aluminum Cans	206
Copper Wire	83
Steel Cans	20
LDPE Plastics (e.g., trash bags)	56
PET Plastics (e.g., beverage bottles)	53
HDPE Plastics (e.g., household cleaner bottles)	51
Personal Computers	43
Carpet	106
Glass	2
Corrugated Cardboard	15
Newspaper	16
Phone Books	11
Magazines	11
Office Paper	10

Amount of energy saved by manufacturing different materials using the maximum percentage of recycled material as opposed to using all new material. *Source:* National Geographic, January 2008. Volume 213 No., pg 82-83.

(a) Construct a frequency table, including the actual frequency, the relative frequency (round to the nearest tenth of a percent), and the relative cumulative frequency. Assume a bin width of 25 million BTUs.

(b) Create a relative frequency histogram from your table in part (a).

(c) Draw the corresponding frequency polygon.

(d) Create the ogive plot.

(e) Comment on the shape, center, and spread of this distribution as it relates to the original data. (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) There is a portion of your ogive plot that should be horizontal. Explain what is happening with the data in this area that creates this horizontal section.

(h) What does the steepest part of an ogive plot tell you about the distribution?

On the Web

http://en.wikipedia.org/wiki/Ogive

Technology Notes: Histograms on the TI-83/84 Graphing Calculator

To draw a histogram on your TI-83/84 graphing calculator, you must first enter the data in a list. In the home screen, press **[2ND][{]**, and then enter the data separated by commas (see the screen below). When all the data have been entered, press **[2ND][}][STO]**, and then press **[2ND][L1][ENTER]**.
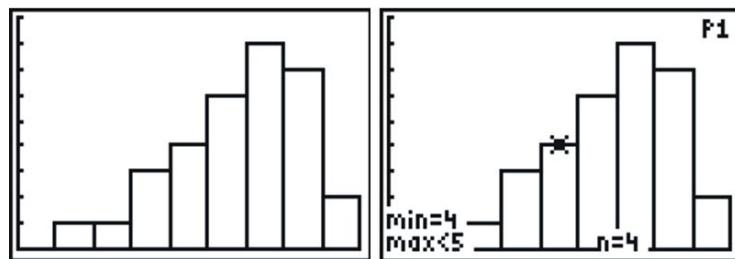
Now you are ready to plot the histogram. Press **[2ND][STAT PLOT]** to enter the **STAT-PLOTS** menu. You can plot up to three statistical plots at one time. Choose **Plot1**. Turn the plot on, change the type of plot to a histogram (see sample screen below), and choose **L1**. Enter '1' for the Freq by pressing **[2ND][A-LOCK]** to turn off alpha lock, which is normally on in this menu, because most of the time you would want to enter a variable here. An alternative would be to enter the values of the variables in **L1** and the frequencies in **L2** as we did in Chapter 1.

Finally, we need to set a window. Press **[WINDOW]** and enter an appropriate window to display the plot. In this case, 'XSCL' is what determines the bin width. Also notice that the maximum *x* value needs to go up to 9 to show the last bin, even though the data values stop at 8. Enter all of the values shown below.

```
WINDOW
 Xmin=0
 Xmax=9
 Xscl=1
 Ymin=0
 Ymax=9
 Yscl=1
 Xres=1
```

Press **[GRAPH]** to display the histogram. If you press **[TRACE]** and then use the left or right arrows to trace along the graph, notice how the calculator uses the notation to properly represent the values in each bin.

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 and a line graph.
- 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
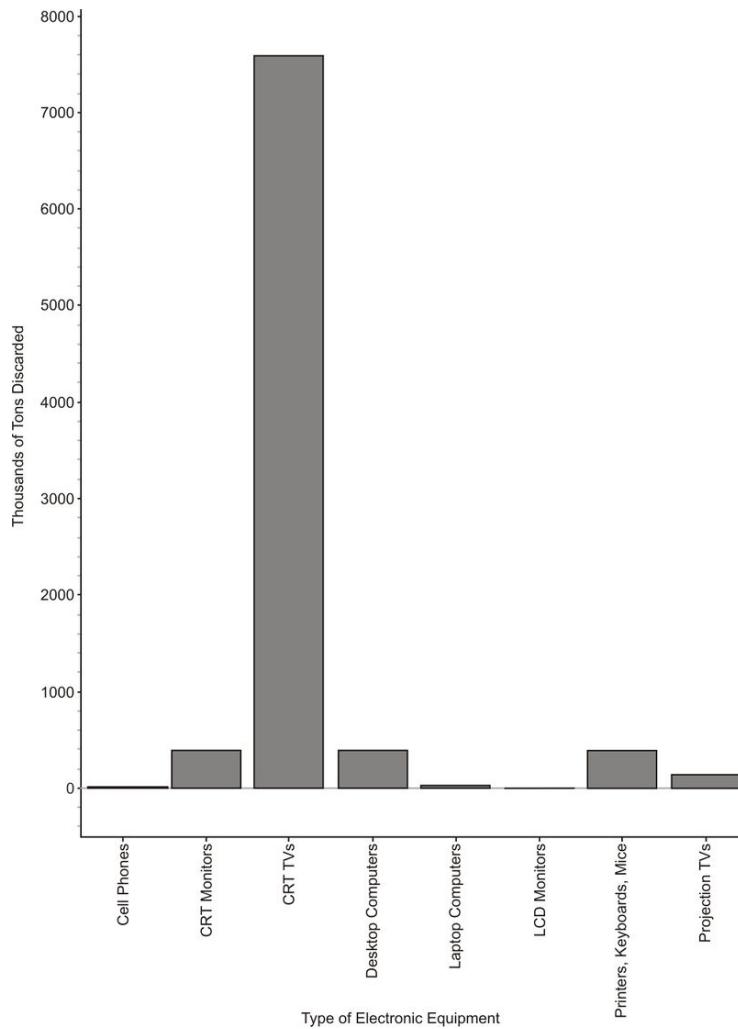
We live in an age of unprecedented access to increasingly sophisticated and affordable personal technology. Cell phones, computers, and televisions now improve so rapidly that, while they may still be in working condition, the drive to make use of the latest technological breakthroughs leads many to 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 2.12:

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
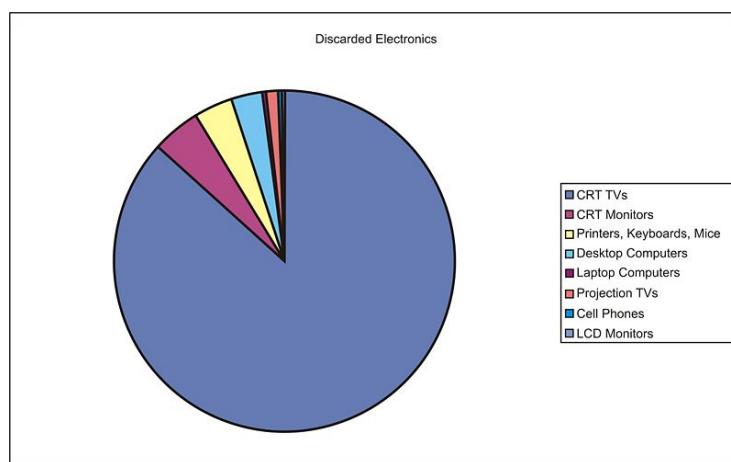
Usually, data that can be represented in a bar graph can also be shown using a *pie graph* (also commonly called a circle graph or pie chart). In this representation, we convert the count into a percentage so we can show each category relative to the total. Each percentage is then converted into a proportionate sector of the circle. To make this conversion, simply multiply the percentage by 3.6, which represents 360 (the total number of degrees in a circle) divided by 100% (the total percentage available.)

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

<center>TABLE 2.13:</center>

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.
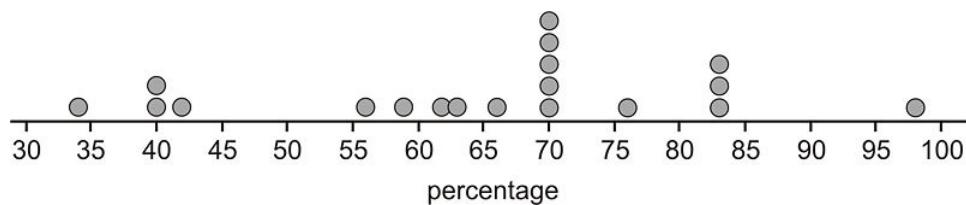
TABLE 2.14: Percentage of the paper packaging used in a country that is recycled. Source: National Geographic, January 2008. Volume 213 No.1, pg 86-87.

Country	% of Paper Packaging Recycled
Estonia	34
New Zealand	40
Poland	40
Cyprus	42
Portugal	56

Country	% of Paper Packaging Recycled
United States	59
Italy	62
Spain	63
Australia	66
Greece	70
Finland	70
Ireland	70
Netherlands	70
Sweden	70
France	76
Germany	83
Austria	83
Belgium	83
Japan	98

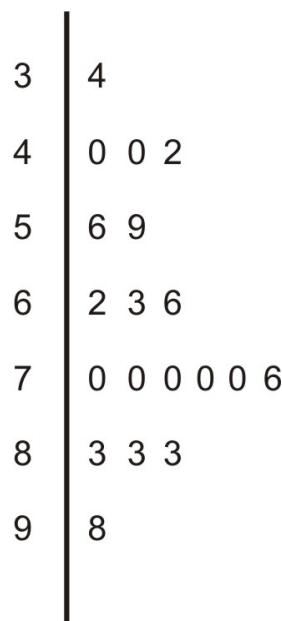The dot plot for this data would look like this:

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:

```
        3
        4
        5
        6
        7
        8
        9
```
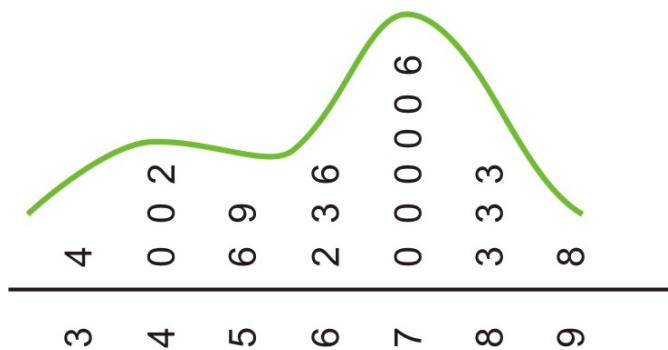
Once the stems are decided, the leaves representing the one's digits are listed in numerical order from left to right:

```
3 | 4
4 | 0 0 2
5 | 6 9
6 | 2 3 6
7 | 0 0 0 0 0 6
8 | 3 3 3
9 | 8
```

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.

Example: Consider the following populations of counties in California.

Butte - 220,748

Calaveras - 45,987

Del Norte - 29,547

Fresno - 942,298

Humboldt - 132,755

Imperial - 179,254

San Francisco - 845,999

Santa Barbara - 431,312

To construct a stem and leaf plot, we need to first make sure each piece of data has the same number of digits. In our data, we will add a 0 at the beginning of our 5 digit data points so that all data points have six digits. Then, we can either round or truncate all data points to two digits.

TABLE 2.15:

Value	Value Rounded	Value Truncated
149	15	14
657	66	65
188	19	18

$2|2$ represents $220,000 - 229,999$ when data has been truncated

$2|2$ represents $215,000 - 224,999$ when data has been rounded.

If we decide to round the above data, we have:

Butte - 220,000

Calaveras - 050,000

Del Norte - 030,000

Fresno - 940,000

Humboldt - 130,000

Imperial - 180,000

San Francisco - 850,000

Santa Barbara - 430,000

And the stem and leaf will be as follows:

```
0 | 3  5
1 | 3  8
2 | 2
4 | 5
9 | 4
```

where:

$2|2$ represents $215,000 - 224,999$.

Source: California State Association of Counties http://www.counties.org/default.asp?id=399
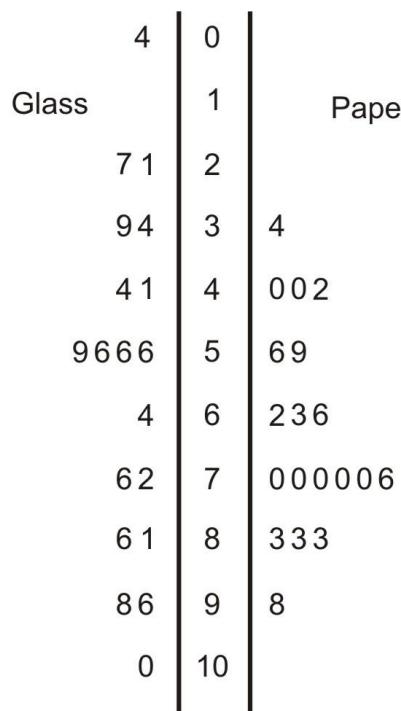
Back-to-Back Stem Plots

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 a 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 2.16: Percentage of the glass packaging used in a country that is recycled. Source: National Geographic, January 2008. Volume 213 No.1, pg 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.

```
              4 | 0  |
      Glass     | 1  |     Paper
            7 1 | 2  |
            9 4 | 3  | 4
            4 1 | 4  | 0 0 2
         9 6 6 6 | 5  | 6 9
              4 | 6  | 2 3 6
            6 2 | 7  | 0 0 0 0 0 6
            6 1 | 8  | 3 3 3
            8 6 | 9  | 8
              0 | 10 |
```

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 2.17:

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
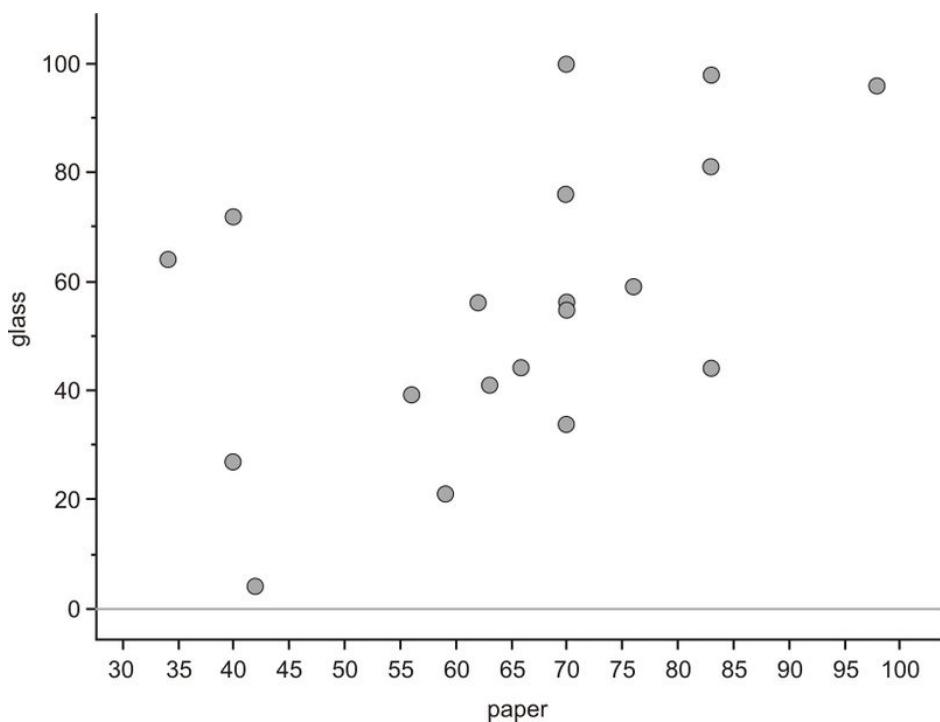
TABLE 2.17: (continued)

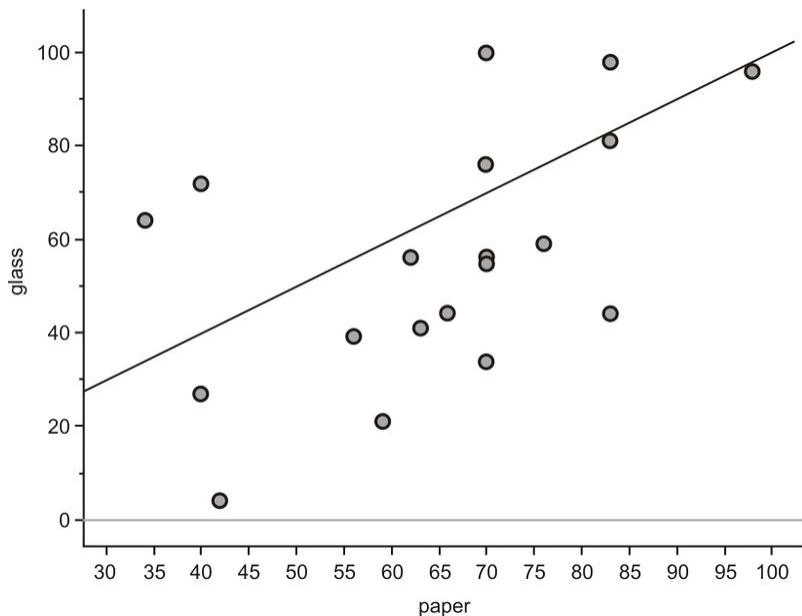| Country	% of Paper Packaging Recycled	% of Glass Packaging Recycled
Greece	70	34
Finland	70	56
Ireland	70	55
Netherlands	70	76
Sweden	70	100
France	76	59
Germany	83	81
Austria	83	44
Belgium	83	98
Japan	98	96

Figure: Paper and Glass Packaging Recycling Rates for 19 countries

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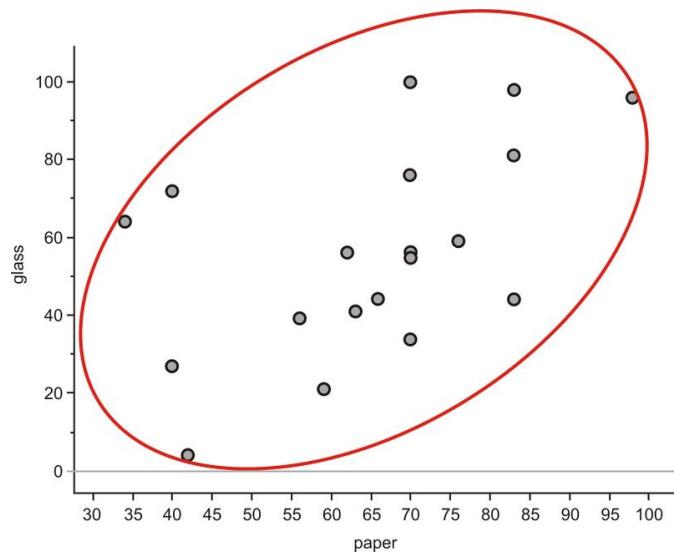
We will place the paper recycling rates on the horizontal axis and those for glass on the vertical axis. Next, we will plot a point that shows each country's rate of recycling for the two materials. This series of disconnected points is referred to as a *scatterplot*.

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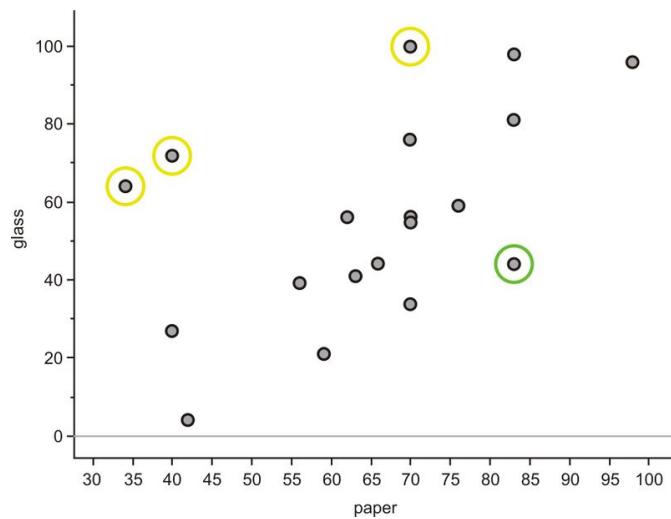, we will also discuss three important characteristics: shape, direction, and strength. These characteristics will inform us about the association between the two variables. The easiest way to describe these traits for this scatterplot is to think of the data as a cloud. If you draw an ellipse around 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.

On the Web

http://tinyurl.com/y8vcm5y Guess the correlation.

Line Plots

Example: The following data set shows the change in the total amount of municipal waste generated in the United States during the 1990's:

TABLE 2.18:
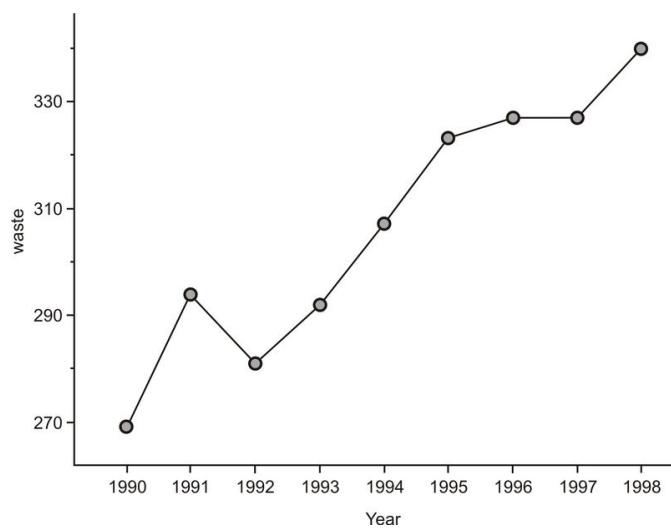
Year	Municipal Waste Generated (Millions of Tons)
1990	269
1991	294
1992	281
1993	292
1994	307
1995	323
1996	327
1997	327
1998	340

Figure: Total Municipal Waste Generated in the US by Year in Millions of Tons. *Source:* http://www.zerowaste america.org/MunicipalWasteManagementReport1998.htm

In this example, the time in years is considered the *explanatory variable*, or independent variable, and the amount of municipal waste is the *response variable*, or dependent variable. It is not only the passage of time that causes our waste to increase. Other factors, such as population growth, economic conditions, and societal habits and attitudes also contribute as causes. However, it would not make sense to view the relationship between time and municipal

waste in the opposite direction.

When one of the variables is time, it will almost always be the explanatory variable. Because time is a continuous variable, and we are very often interested in the change a variable exhibits over a period of time, there is some meaning to the connection between the points in a plot involving time as an explanatory variable. In this case, we use a line plot. A line plot is simply a scatterplot in which we connect successive chronological observations with a line segment to give more information about how the data values are changing over a period of time. Here is the line plot for the US Municipal Waste data:

It is easy to see general trends from this type of plot. For example, we can spot the year in which the most dramatic increase occurred (1990) by looking at the steepest line. We can also spot the years in which the waste output decreased and/or remained about the same (1991 and 1996). It would be interesting to investigate some possible reasons for the behaviors of these individual years.

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. In a line graph, there is significance to the change between consecutive points, so these points are connected. Line graphs are often used when the explanatory variable is time.

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?

Multimedia Links

For a description of how to draw a stem-and-leaf plot, as well as how to derive information from one **(14.0)**, see APUS07, Stem-and-Leaf Plot (8:08).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1059

Review Questions

1. 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 2.19:

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

(a) Create a bar graph for this data.

(b) Complete the chart below to show the approximate percentage of the total weight for each material.

TABLE 2.20:

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	

(c) Create a circle graph for this data.

2. 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 2.21:

State	Percentage
Alabama	23
Alaska	7
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
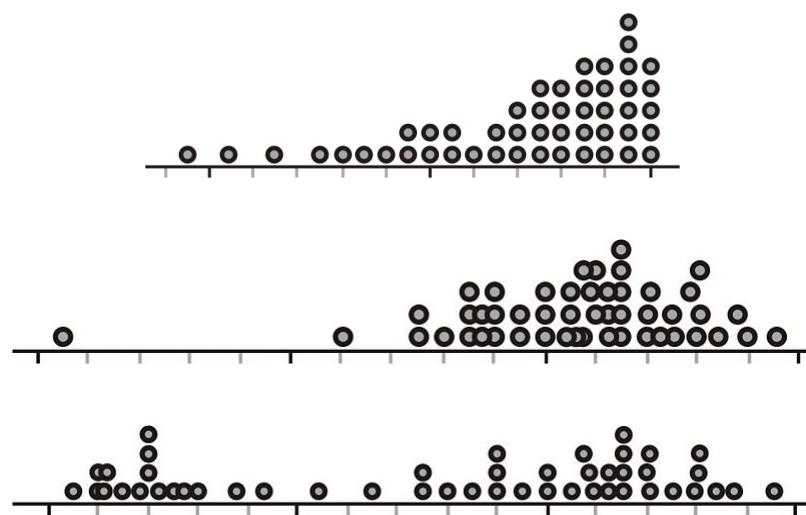
TABLE 2.21: (continued)
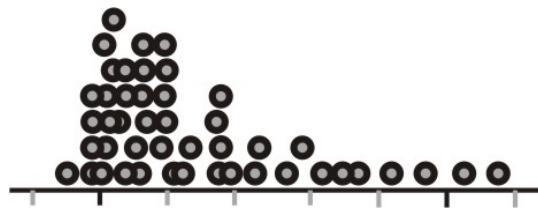
State	Percentage
Nevada	15
New Hampshire	25
New Jersey	45
New Mexico	12
New York	39
North Carolina	26
North Dakota	21
Ohio	19
Oklahoma	12
Oregon	28
Pennsylvania	26
Rhode Island	23
South Carolina	34
South Dakota	42
Tennessee	40
Utah	19
Vermont	30
Virginia	35
Washington	48
West Virginia	20
Wisconsin	36
Wyoming	5

Source: http://www.zerowasteamerica.org/MunicipalWasteManagementReport1998.htm
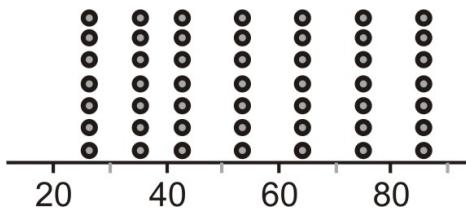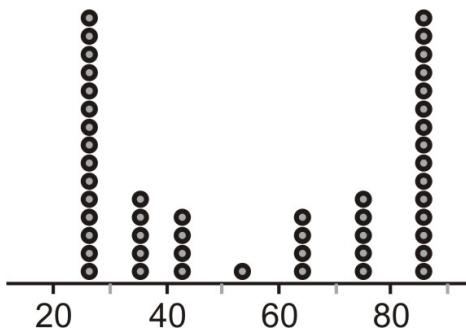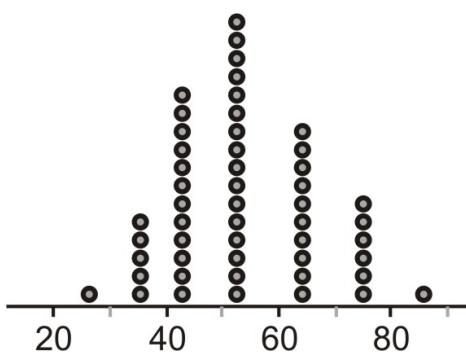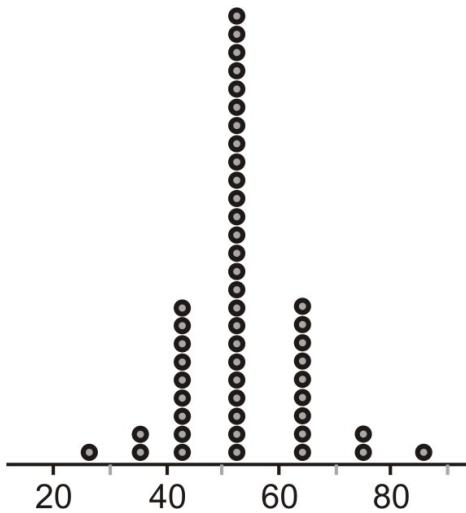
(a) Create a dot plot for this data.

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

(c) Create a stem-and-leaf plot for the data.

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

3. Identify the important features of the shape of each of the following distributions.

Questions 4-7 refer to the following dot plots:

4. Identify the overall shape of each distribution.
5. How would you characterize the center(s) of these distributions?
6. Which of these distributions has the smallest standard deviation?
7. Which of these distributions has the largest standard deviation?
8. In question 2, you looked at the percentage of waste recycled in each state. Do you think there is a relationship between the percentage recycled and the total amount of waste that a state generates? Here are the data, including both variables.

TABLE 2.22:
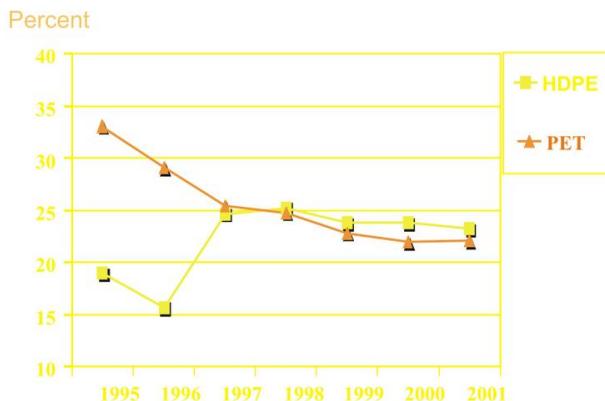
State	Percentage	Total Amount of Municipal Waste in Thousands of Tons
Alabama	23	5549
Alaska	7	560
Arizona	18	5700
Arkansas	36	4287
California	30	45000
Colorado	18	3084
Connecticut	23	2950
Delaware	31	1189
District of Columbia	8	246
Florida	40	23617
Georgia	33	14645
Hawaii	25	2125
Illinois	28	13386
Indiana	23	7171
Iowa	32	3462
Kansas	11	4250
Kentucky	28	4418
Louisiana	14	3894
Maine	41	1339
Maryland	29	5329
Massachusetts	33	7160
Michigan	25	13500
Minnesota	42	4780
Mississippi	13	2360
Missouri	33	7896
Montana	5	1039
Nebraska	27	2000
Nevada	15	3955
New Hampshire	25	1200
New Jersey	45	8200
New Mexico	12	1400
New York	39	28800
North Carolina	26	9843
North Dakota	21	510
Ohio	19	12339
Oklahoma	12	2500
Oregon	28	3836
Pennsylvania	26	9440
Rhode Island	23	477

TABLE 2.22: (continued)

State	Percentage	Total Amount of Municipal Waste in Thousands of Tons
South Carolina	34	8361
South Dakota	42	510
Tennessee	40	9496
Utah	19	3760
Vermont	30	600
Virginia	35	9000
Washington	48	6527
West Virginia	20	2000
Wisconsin	36	3622
Wyoming	5	530

(a) Identify the variables in this example, and specify which one is the explanatory variable and which one is the response variable.

(b) How much municipal waste was created in Illinois?

(c) Draw a scatterplot for this data.

(d) Describe the direction and strength of the association between the two variables.

9. The following line graph shows the recycling rates of two different types of plastic bottles in the US from 1995 to 2001.

a. Explain the general trends for both types of plastics over these years.
b. What was the total change in PET bottle recycling from 1995 to 2001?
c. Can you think of a reason to explain this change?
d. During what years was this change the most rapid?

References

National Geographic, January 2008. Volume 213 No.1

http://e-stewards.org/the-e-waste-crisis/

Technology Notes: Scatterplots on the TI-83/84 Graphing Calculator
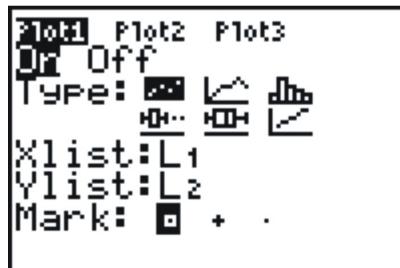
Press **[STAT][ENTER]**, and enter your data, with the explanatory variable in **L1** and the response variable in **L2**. (Note that this data set contains 18 points- not all are visible on the screen at once). Next, press **[2ND][STAT-PLOT]** to enter the **STAT-PLOTS** menu, and choose the first plot.
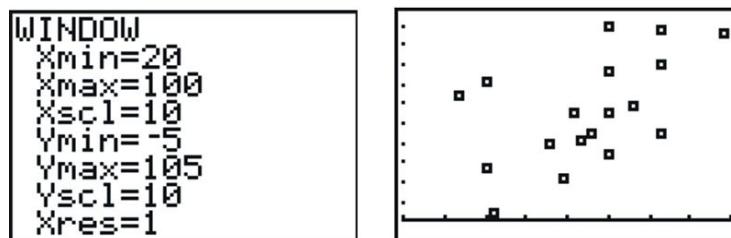
Change the settings to match the following screenshot:

This selects a scatterplot with the explanatory variable in **L1** and the response variable in **L2**. In order to see the points better, you should choose either the square or the plus sign for the mark. The square has been chosen in the screenshot. Finally, set the window as shown below to match the data. In this case, we looked at our lowest and highest data values in each variable and added a bit of room to create a pleasant window. Press **[GRAPH]** to see the result, shown below.

Line Plots on the TI-83/84 Graphing Calculator

Your graphing calculator will also draw a line plot, and the process is almost identical to that for creating a scatterplot. Enter the data into your lists, and choose a line plot in the **Plot1** menu, as in the following screenshot.

Next, set an appropriate window (not necessarily the one shown below), and graph the resulting plot.

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.
- Compare distributions of univariate data (shape, center, spread, and outliers).
- Describe the effects of changing units on summary measures.

Introduction

In this section, the box-and-whisker plot will be introduced, and the basic ideas of shape, center, spread, and outliers will be studied in this context.

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 amount of water held by each major reservoir in Arizona stated as a percentage of that reservoir's total capacity.

TABLE 2.23:

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].

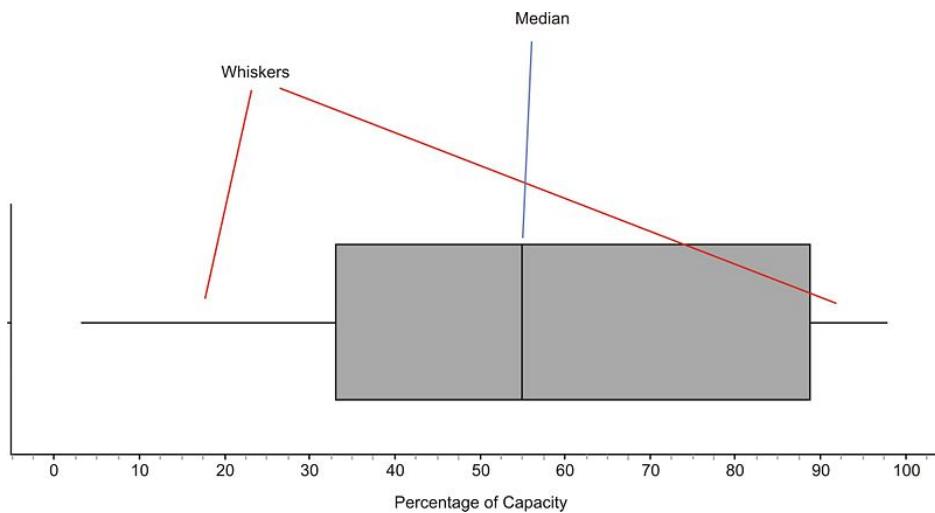Placing the data in order from smallest to largest gives the following:

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 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 is 89. Therefore, the five-number summary is as shown:
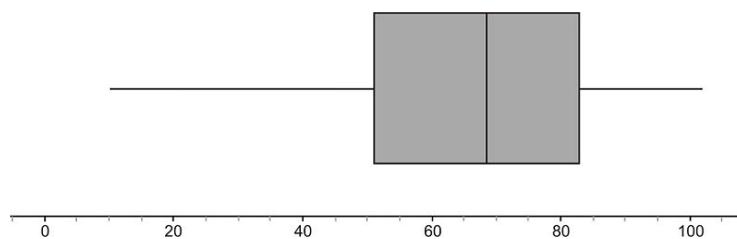
$$\{3, 33, 55, 89, 98\}$$

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 to the upper quartile. 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:
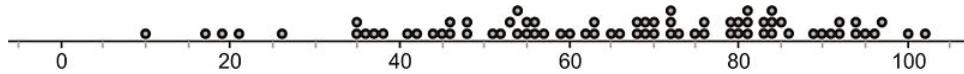
The plot divides the data into quarters. If the number of data points is divisible by 4, then there will be exactly the same number of values in each of the two whiskers, as well as the two sections in the box. In this example, because there are 10 data points, the number of values in each section will only be approximately the same, but 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).

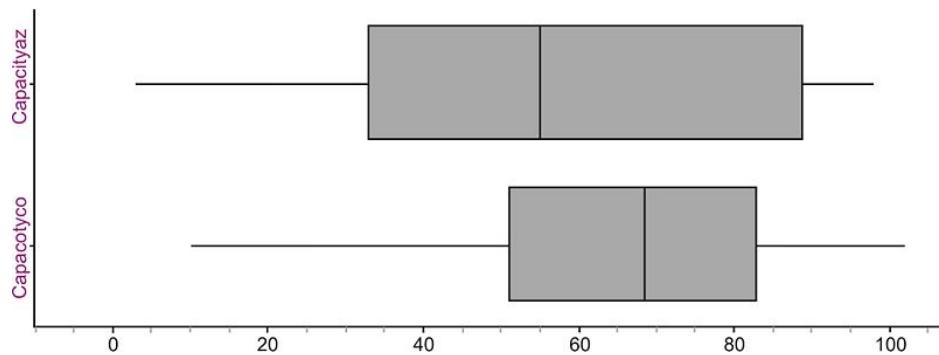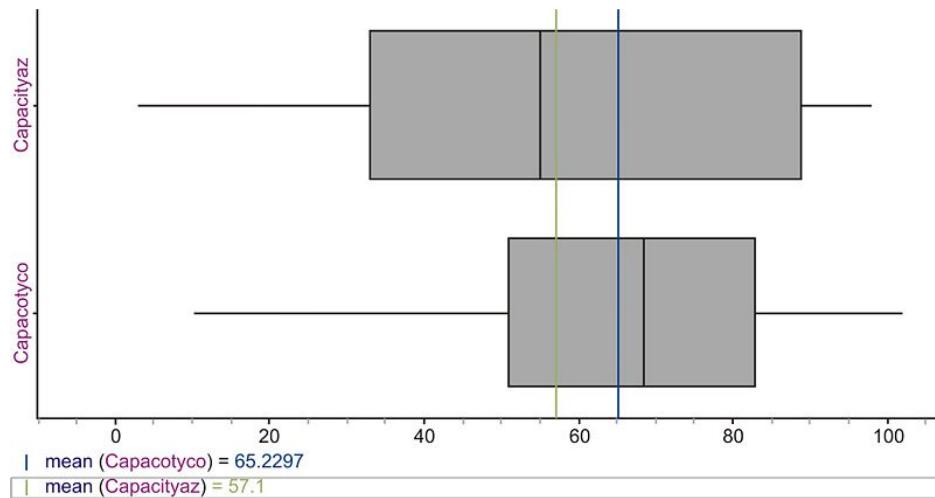## Comparing Multiple Box Plots

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 reservoirs are less full, as a percentage of their individual capacities, than Colorado's. 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 affect is still visible. Here are both distributions with the means plotted for each.

| mean (Capacotyco) = 65.2297
| mean (Capacityaz) = 57.1
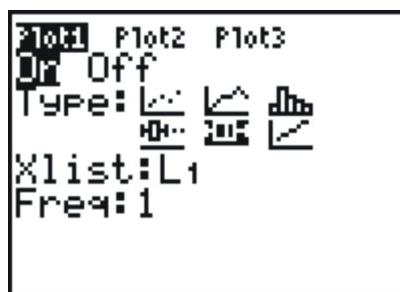
Notice that the long left whisker in the Colorado data causes the mean to be pulled toward the left, making it lower than the median. In the Arizona plot, you can see that the mean is slightly higher than the median, due to the slightly elongated right side of the box. If these data sets were perfectly symmetric, the mean would be equal to the median in each case.

Outliers in Box-and-Whisker Plots
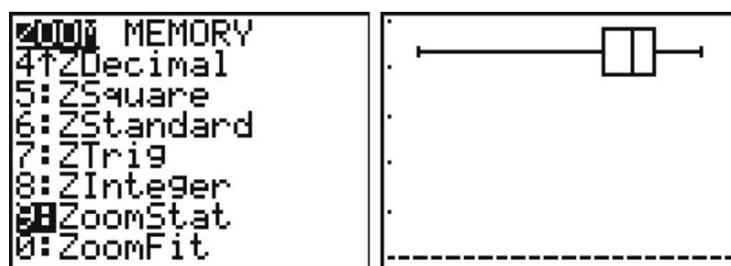
Here are the reservoir data for California (the names of the lakes and reservoirs have been omitted):

80, 83, 77, 95, 85, 74, 34, 68, 90, 82, 75
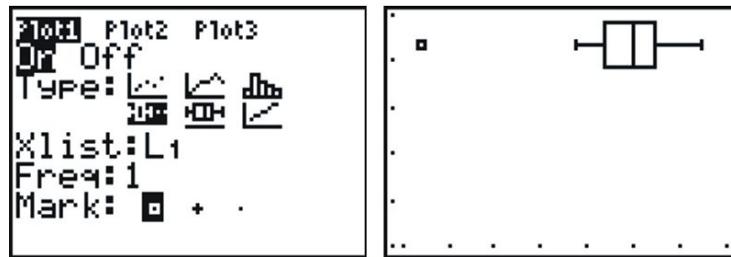
At first glance, the 34 should stand out. It appears as if this point is significantly different from the rest of the data. Let's use a graphing calculator to investigate this plot. Enter your data into a list as we have done before, and then choose a plot. Under 'Type', you will notice what looks like 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'.
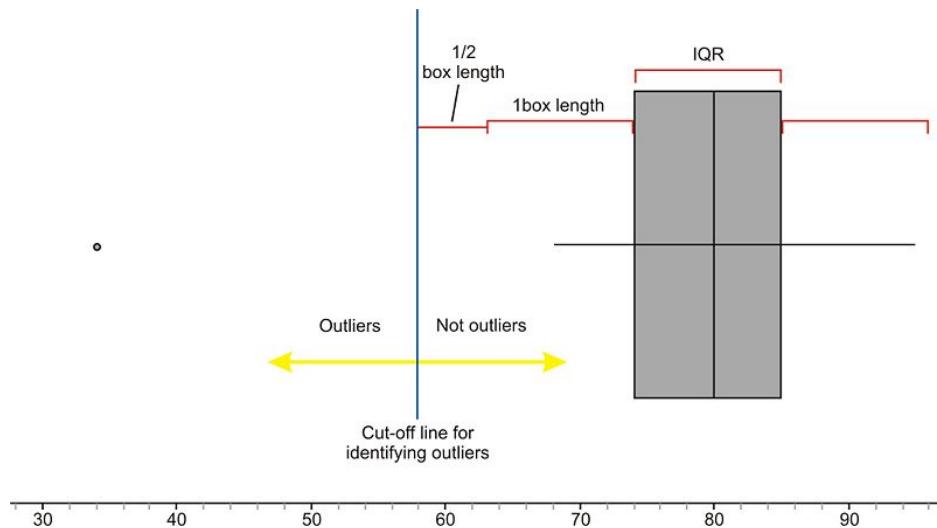
While box plots give us a nice summary of the important features of a distribution, we lose the ability to identify individual points. The left whisker is elongated, but if we did not have the data, we would not know if all the points in that section of the data were spread out, or if it were just the 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: 74

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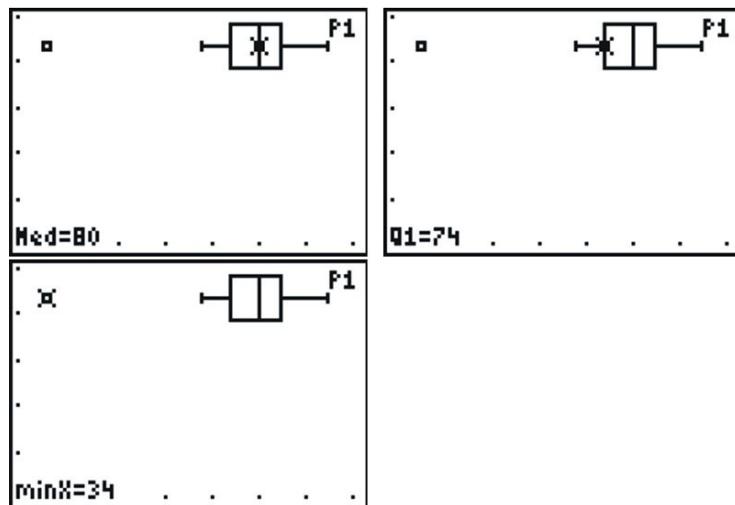
Interquartile range $(IQR) : 85 - 74 = 11$

$1.5 * IQR = 16.5$

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

Notice that we did not even bother to test the calculation on the right whisker, because it should be obvious from a quick visual inspection that there are no points that are farther than even one box length away from the upper quartile.

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 2.24:

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.
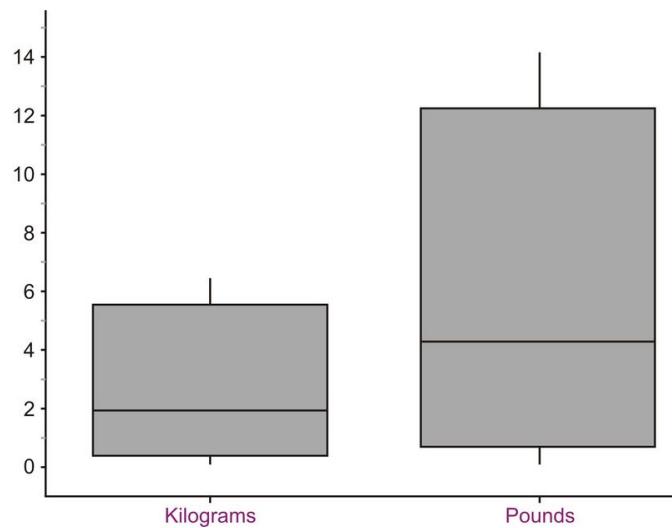
TABLE 2.25:

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.

Example: This is easier to think of with numbers. Suppose that your mean is 20, and that two of the data values in your distribution are 21 and 23. If you multiply 21 and 23 by 2, you get 42 and 46, and your mean also changes by a factor of 2 and is now 40. Before your deviations were $21 - 20 = 1$ and $23 - 20 = 3$, but now, your deviations are $42 - 40 = 2$ and $46 - 40 = 6$, so your deviations are getting twice as big as well.

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.

On the Web

http://tinyurl.com/34s6sm Investigate the mean, median and box plots.

http://tinyurl.com/3ao9px More investigation of boxplots.

Lesson Summary
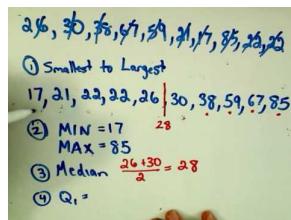
The five-number summary is a useful collection of statistical measures consisting of the following in ascending order: minimum, lower quartile, median, upper quartile, maximum. A box-and-whisker plot is a graphical representation of the five-number summary showing a box bounded by the lower and upper 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 changing the units of a distribution, the center and spread will be affected, but the shape will stay the same.

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?

Multimedia Links

For a description of how to draw a box-and-whisker plot from given data **(14.0)**, see patrickJMT, Box and Whisker Plot (5:53).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1060

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. 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. The following table contains recent data on the average price of a gallon of gasoline for states that share a border crossing into Canada.

    a. Find the five-number summary for this data.

    b. Show all work to test for outliers.

    c. Graph the box-and-whisker plot for this data.

    d. Canadian gasoline is sold in liters. Suppose a Canadian crossed the border into one of these states and wanted to compare the cost of gasoline. There are 3.7854 liters in a gallon. If we were to convert the distribution to liters, describe the resulting shape, center, and spread of the new distribution.

    e. Complete the following table. Convert to cost per liter by dividing by 3.7854, and then graph the resulting box plot.

As an interesting extension to this problem, you could look up the current data and compare that distribution with the data presented here. You could also find the exchange rate for Canadian dollars and convert the prices into the other currency.

TABLE 2.26:

State	Average Price of a Gallon of Gasoline (US$)	Average Price of a Liter of Gasoline (US$)
Alaska	3.458	
Washington	3.528	
Idaho	3.26	
Montana	3.22	
North Dakota	3.282	
Minnesota	3.12	
Michigan	3.352	
New York	3.393	
Vermont	3.252	
New Hampshire	3.152	
Maine	3.309	

Average Prices of a Gallon of Gasoline on March 16, 2008
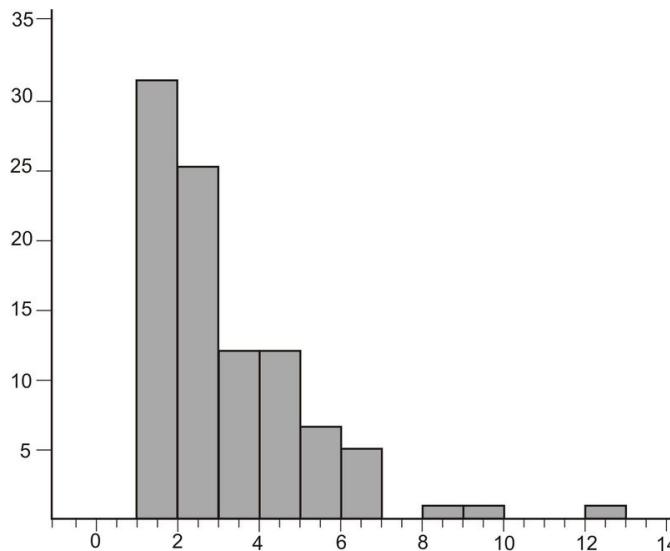
Figure: Average prices of a gallon of gasoline on March 16, 2008. *Source:* AAA, http://fuelgaugereport.opisnet.com/sbsavg.html

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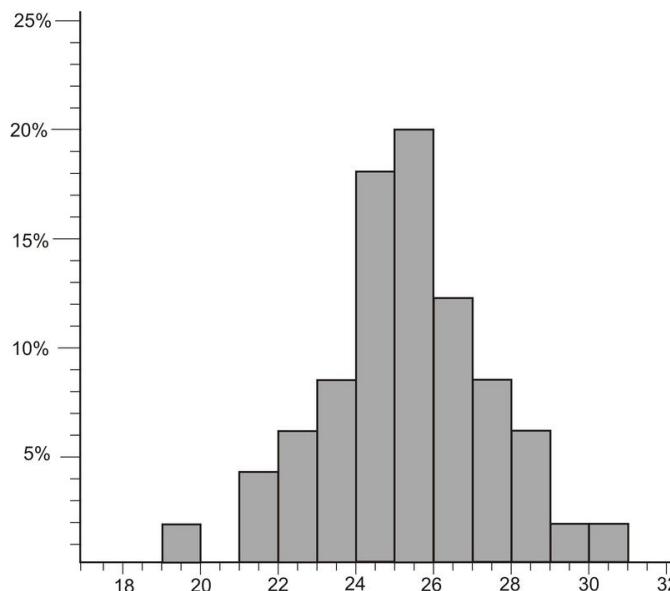
[1] Kunzig, Robert. Drying of the West. National Geographic, February 2008, Vol. 213, No. 2, Page 94.

http://en.wikipedia.org/wiki/Box_plot

Part One: Questions

1. Which of the following can be inferred from this histogram?
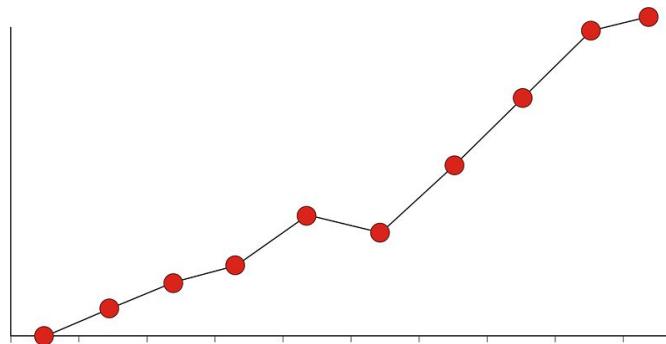
    a. The mode is 1.
    b. mean <median
    c. median <mean
    d. The distribution is skewed left.
    e. None of the above can be inferred from this histogram.

2. Sean was given the following relative frequency histogram to read.
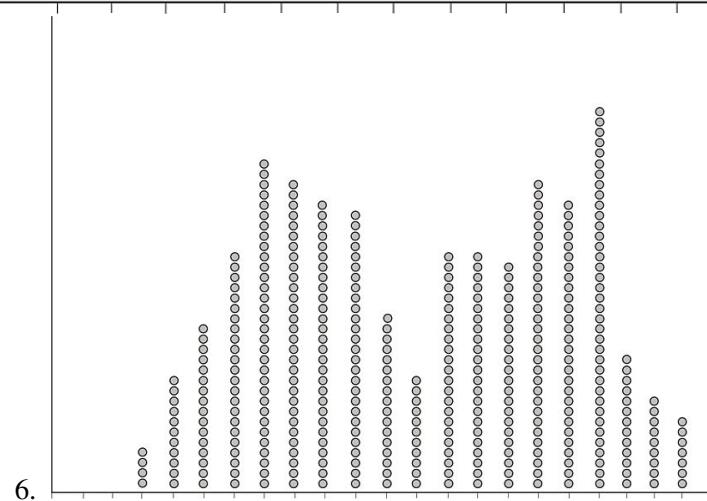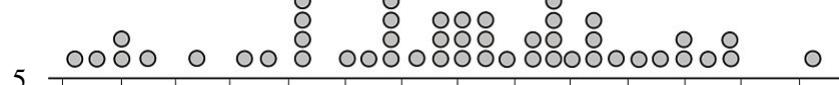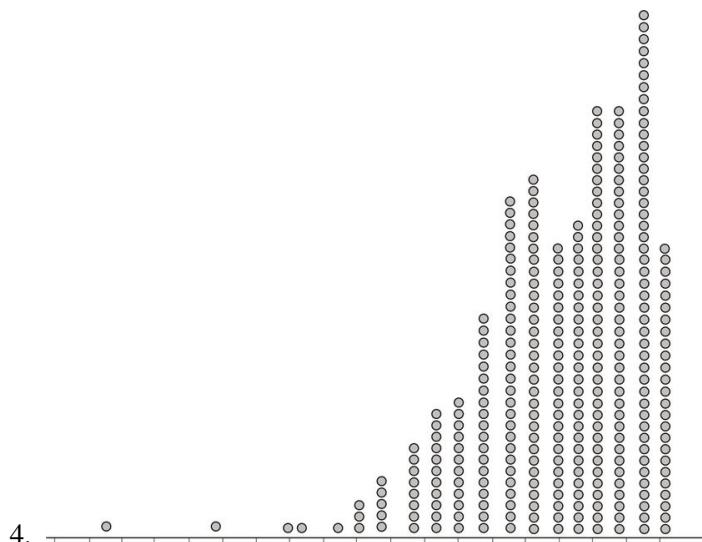
Unfortunately, the copier cut off the bin with the highest frequency. Which of the following could possibly be the relative frequency of the cut-off bin?
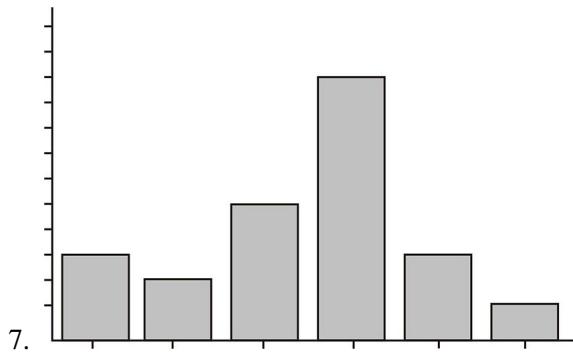
    a. 16
    b. 24
    c. 30
    d. 68

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:
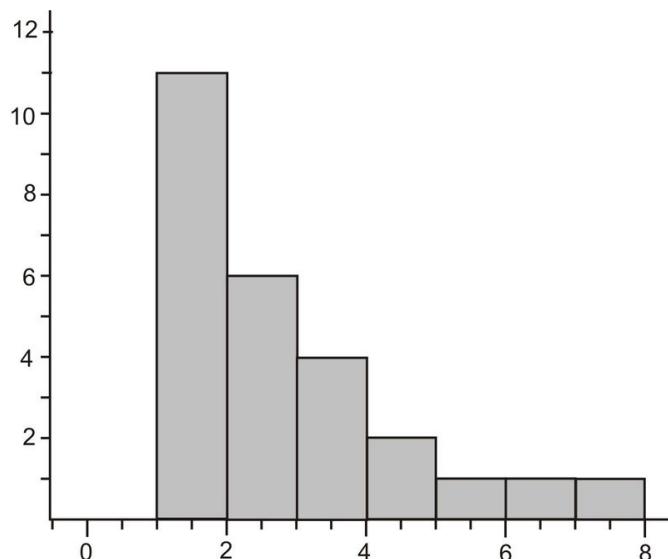
Identify which of the two graphs she has and briefly explain why.
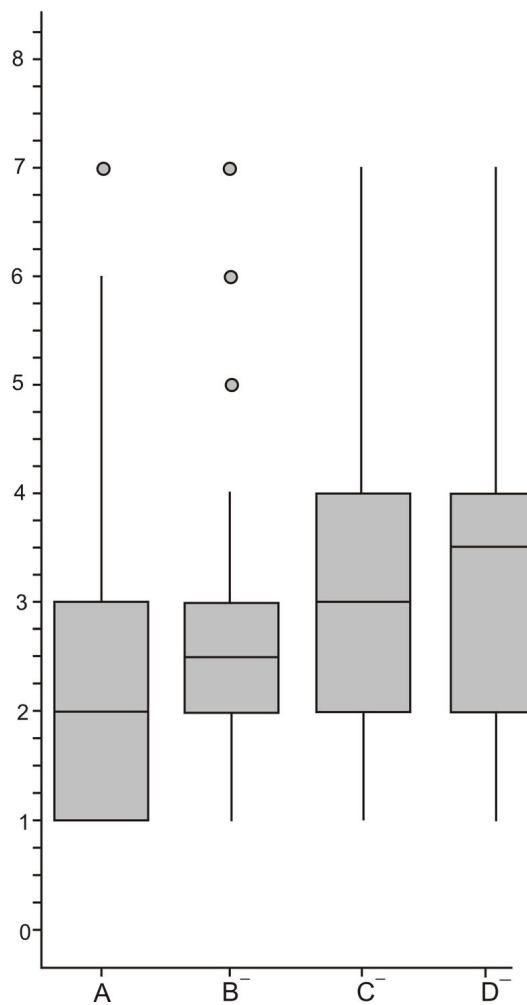
In questions 4-7, match the distribution with the choice of the correct real-world situation that best fits the graph.

4.

5.

6.

7.
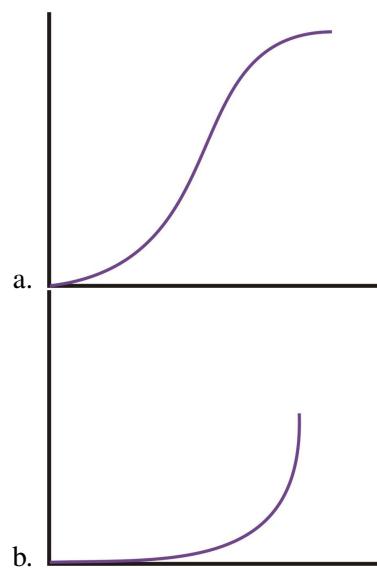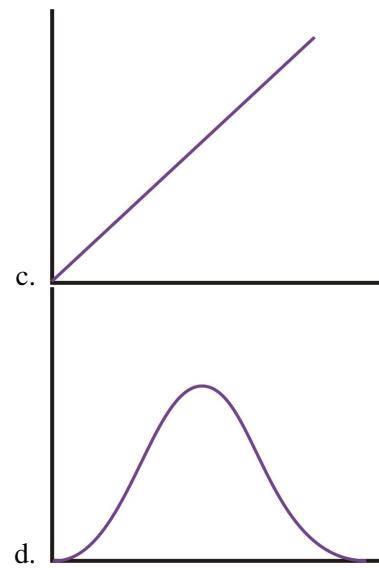
a. Endy collected and graphed the heights of all the 12th grade students in his high school.
b. Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or piggy bank. She created a plot of the dates of the pennies.
c. Thamar asked her friends 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?
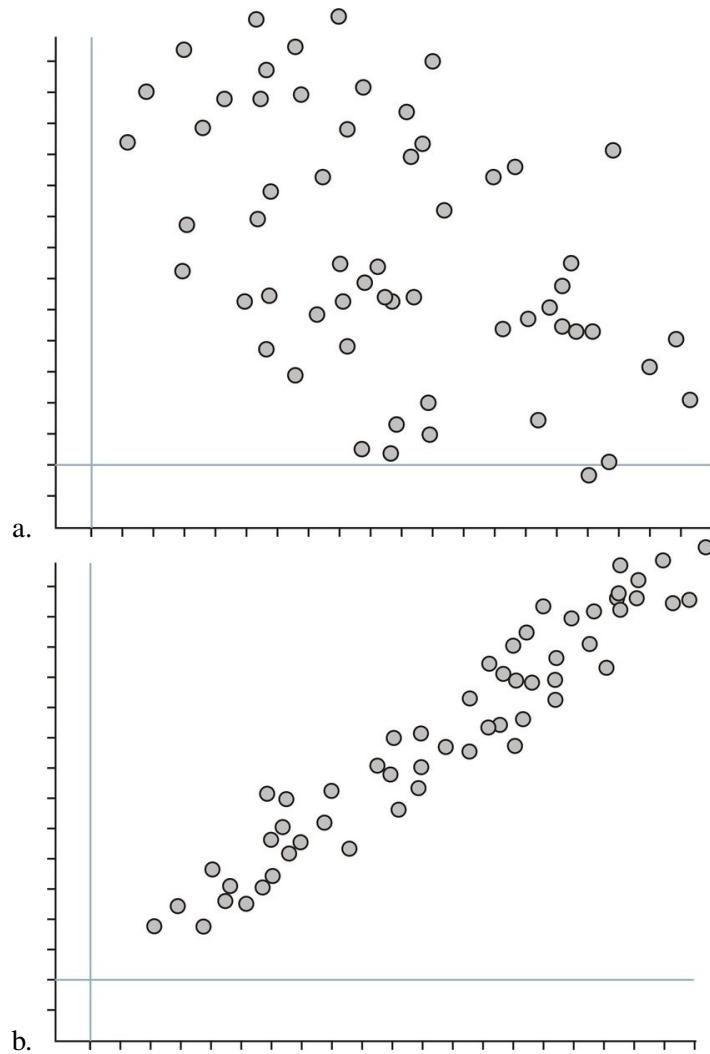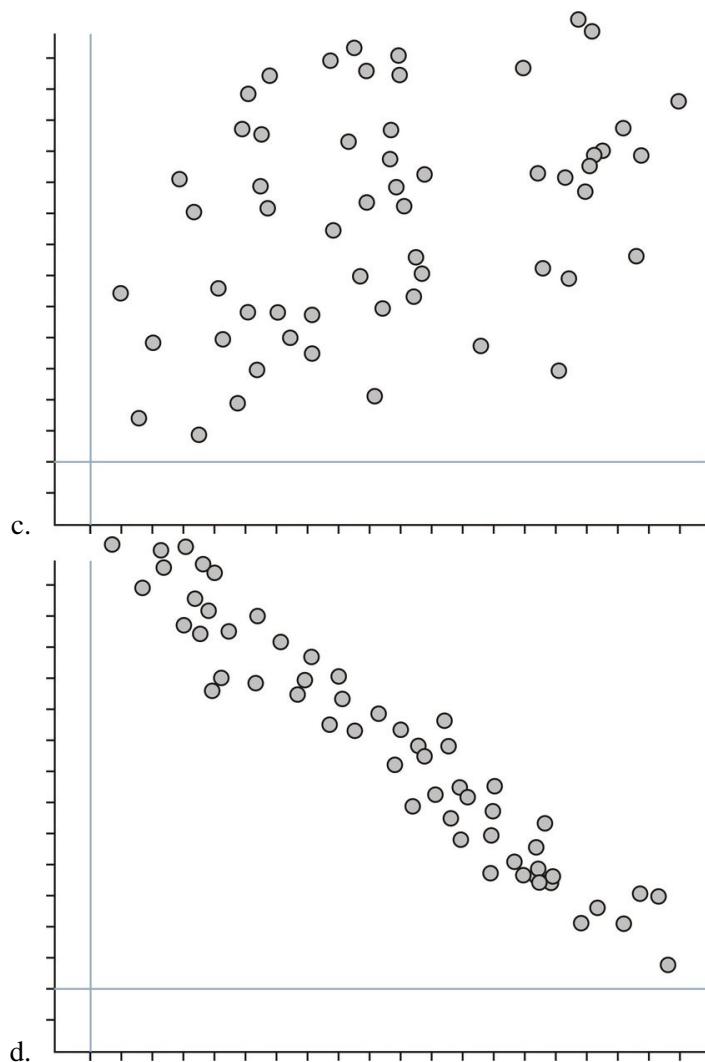
9. If a data set is roughly symmetric with no skewing or outliers, which of the following would be an appropriate sketch of the shape of the corresponding ogive plot?

c.

d.

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

a.

b.

c.

d.

Part Two: Open-Ended Questions

1. The chart below lists the 14 tallest buildings in the world (as of 12/2007).

TABLE 2.27:

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
Bank of America Tower	New York	1200

<p style="text-align:center">**TABLE 2.27:** (continued)</p>

Building	City	Height (ft)
Emirates Office Tower	Dubai	1163
Tuntex Sky Tower	Kaohsiung	1140

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

<p style="text-align:center">**TABLE 2.28:**</p>

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.

<p style="text-align:center">**TABLE 2.29:**</p>

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
1998	395,942	37,567
1999	416,789	21,994
2000	546,056	33,439
2001	775,499	46,526
2002	521,636	29,806
2003	283,554	67,660
2004	394,007	18,115
2005	267,908	8.048
2006	87,966	1,897

Figure: Total Fall Salmon Escapement in the Sacramento River. *Source:* http://www.pcouncil.org/bb/2008/1108/D1a_ATT2_1108.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 2.30:

Island	Approximate Area (sq. km)
Baltra	8
Darwin	1.1
Española	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.

   a. 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.)
   b. 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

Bar graph

Bias

Bivariate data

Box-and-whisker plot

Cumulative frequency histogram

Density curves

Dot plot

Explanatory variable

Five-number summary

Frequency polygon

Frequency tables

Histogram

Modified box plot

Mound-shaped

Negative linear association

Ogive plot

Pie graph

Positive linear association

Relative cumulative frequency histogram

Relative cumulative frequency plot

Relative frequency histogram

Response variable

Scatterplot

Skewed left

Skewed right

Stem-and-leaf plot

Symmetric

Tail

CHAPTER **3**

An Introduction to Probability

Chapter Outline

3.1 Events, Sample Spaces, and Probability

Learning Objectives

- Know basic statistical terminology.
- List simple events and sample spaces.
- Know the basic rules of probability.

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.

	H	T
H	HH	HT
T	TH	TT

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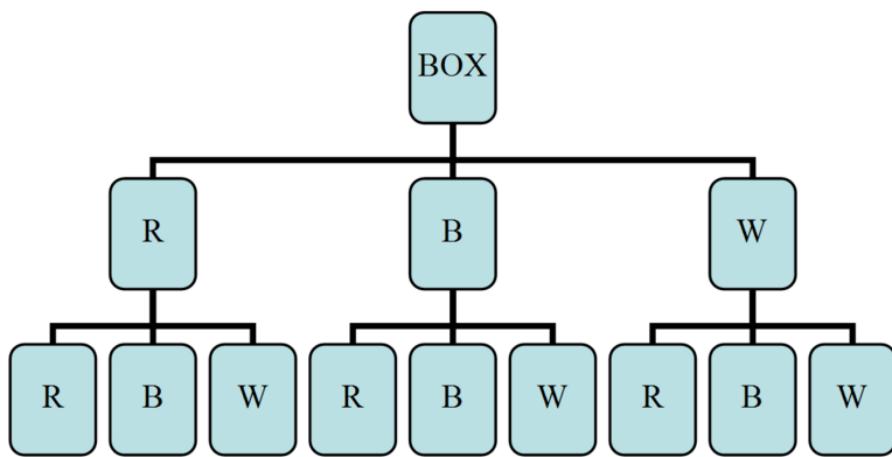
Solution: As we know, there are 6 possible outcomes for throwing a die. We may get 1, 2, 3, 4, 5, or 6, so we write the sample space as the set of all possible outcomes:

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

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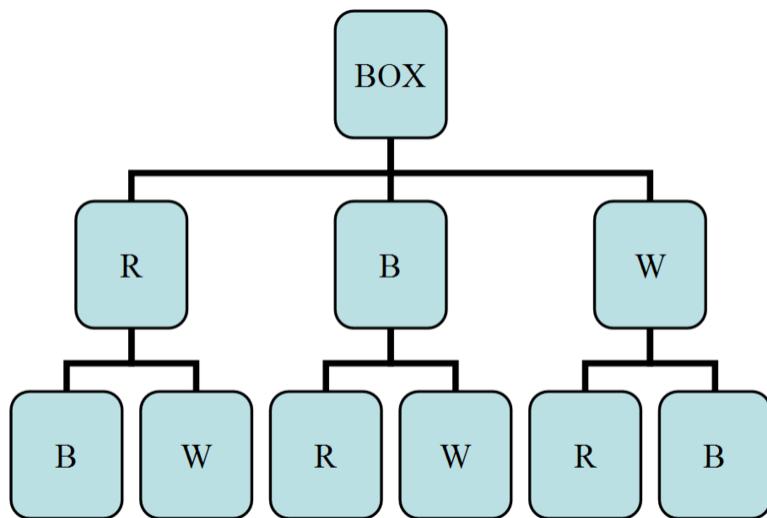
We can also represent all the possible drawings by a table or a matrix:

	R	B	W
R	RR	RB	RW
B	BR	BB	BW
W	WR	WB	WW

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 while 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.

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.

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

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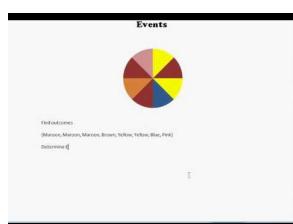
An experiment is the process of taking a measurement or making an observation.

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.

Multimedia Links

For a description of how to find an event given a sample space **(1.0)**, see teachertubemath, Probability Events (2:23).

| MEDIA |
| Click image to the left or use the URL below. |
| URL: http://www.ck12.org/flx/render/embeddedobject/1054 |
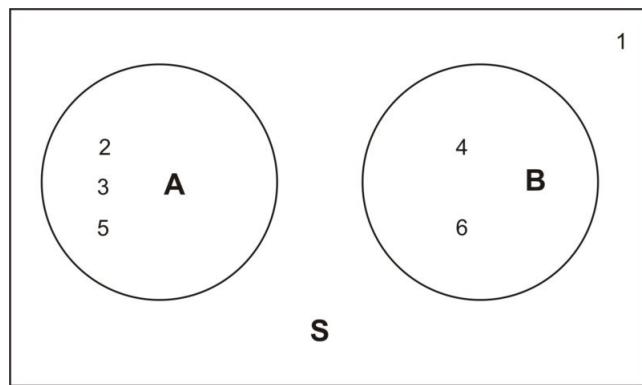
Review Questions

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. The Venn diagram below shows an experiment with six simple events. Events A and B are also shown. The probabilities of the simple events are:

$$P(1) = P(2) = P(4) = \frac{2}{9}$$
$$P(3) = P(5) = P(6) = \frac{1}{9}$$

a. Find $P(A)$
b. Find $P(B)$

3. 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

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:

$$A : \text{observe an even number}$$

$$B : \text{observe a number less than or equal to 3}$$

1. Describe $A \cup B$ for this experiment.
2. Describe $A \cap B$ for this experiment.
3. 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.

$$C : \text{observe a number that is greater than } 5$$

$$D : \text{observe a number that is exactly } 5$$

1. Find the simple events in $A \cup B \cup C$.
2. Find the simple events in $A \cap D$.
3. 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 $D = \{5\}, A \cap D = \{2,3,6\} \cap \{5\} = \varnothing$,

where $\varnothing$ 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.

$$(A \cap B) \cap C = (\{2,4,6\} \cap \{1,2,3\}) \cap \{6\} = \{2\} \cap \{6\} = \varnothing$$

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.

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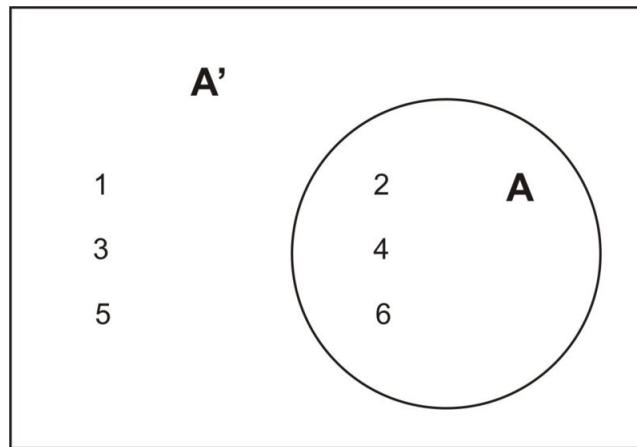
In this lesson, you will learn what is meant by the complement of an event, and you will be introduced to the Complement Rule. You will also learn how to calculate probabilities when the complement of an event is involved.

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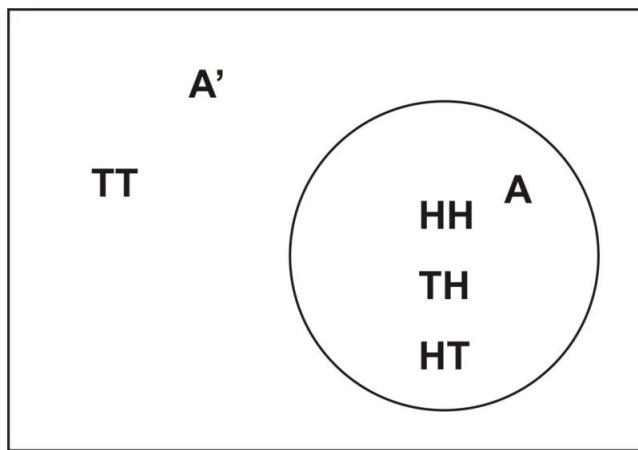
Let the event A be getting the flu this winter. We are given $P(A) = 0.43$. The event not getting the flu is A'. Thus, $P(A') = 1 - P(A) = 1 - 0.43 = 0.57$.

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.

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

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 $HTTHTHHTTH$, 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: $A' = \{TTTTTTTTTT\}$. 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, $P(A) + P(A') = 1$.

Multimedia Links

For an explanation of complements and using them to calculate probabilities (**1.0**), see jsnider3675, An Event's Complement (9:40).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1062
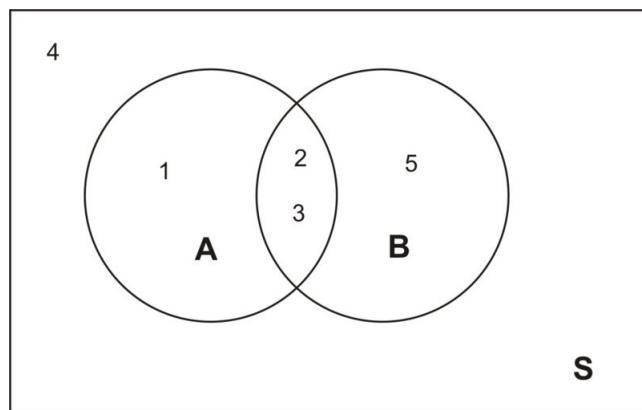
Review Questions

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: $P(A'), P(B'), P(A' \cap B), P(A \cap B), P(A \cup B'), P(A \cup B), P(A \cap B'), P[(A \cup B)']$.

3.4 Conditional Probability

Learning Objective
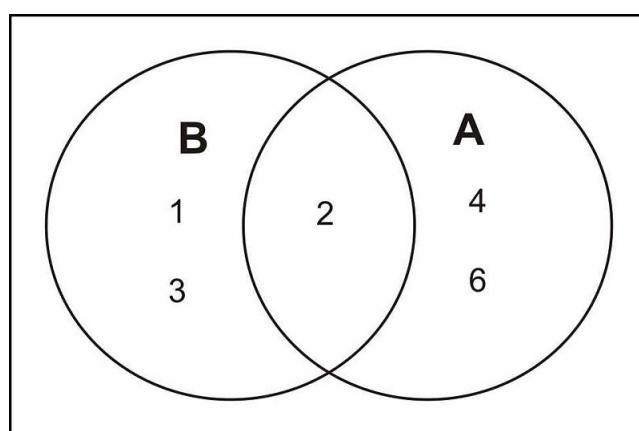
- Calculate the conditional probability that event *A* occurs, given that event *B* has occurred.

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:

$$A : \text{observe an even number}$$

$$B : \text{observe a number less than or equal to 3}$$

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:

$$A : \text{individual smokes}$$

$$C : \text{individual develops cancer}$$

$$A' : \text{individual does not smoke}$$

$$C' : \text{individual does not develop cancer}$$

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.

There is another and interesting way to analyze this problem, which has been called the *natural frequencies approach* (see *G. Gigerenzer, "Calculated Risks" Simon and Schuster, 2002*).

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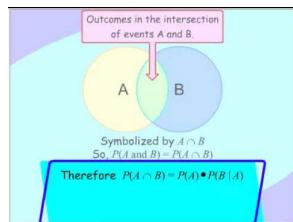
If A and B are two events, then the probability of event A occurring, given that event B has occurred, is called conditional probability. We write it with the notation $P(A|B)$, which reads "the probability of A, given B."

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.

Multimedia Links

For an introduction to conditional probability **(2.0)**, see SomaliNew, Conditonal Probability Venn Diagram (4:25).
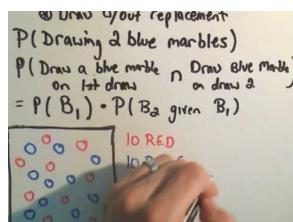
> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1063

For an explanation of how to find the probability of "And" statements and dependent events **(2.0)**, see patrickJMT, Calculating Probability - "And" Statements, Dependent Events (5:36).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1064

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:

$$A : \text{At least one head appears}$$

$$B : \text{Only one head appears}$$

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

3. A box of six marbles contains two white marbles, two red marbles, and two blue marbles. Two marbles are randomly selected without replacement, and their colors are recorded.

    a. List all the possible outcomes in the sample space.

    b. Suppose three events are defined as follows:

$$A : \text{Both marbles have the same color}$$

$$B : \text{Both marbles are red}$$

$$C : \text{At least one marble is red or white}$$

Find $P(B|A), P(B|A'), P(B|C), P(A|C), P(C|A')$, and $P(C|A')$.

3.5 Additive and Multiplicative Rules

Learning Objectives

- Calculate probabilities using the Additive Rule.
- Calculate probabilities using the Multiplicative Rule.
- Identify events that are not mutually exclusive and explain how to represent them in a Venn diagram.
- Understand the condition of independence.

Introduction

In this lesson, you will learn how to combine probabilities with the Additive Rule and the Multiplicative 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 Additive Rule and the Multiplicative Rule, to find these probabilities. The examples that follow will illustrate how we can do this.
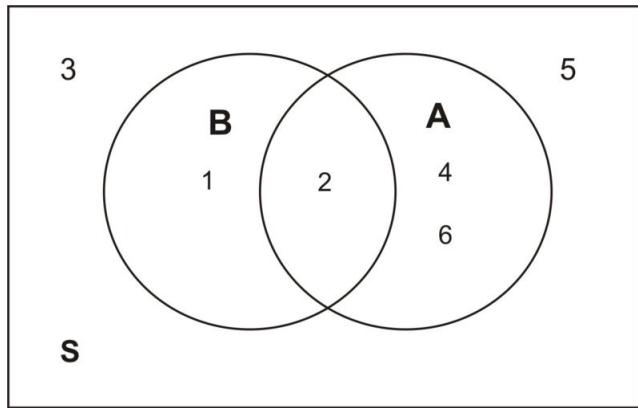
Example: Suppose we have a loaded (unfair) die, and we toss it several times and record the outcomes. We will define the following events:

$$A : \text{observe an even number}$$

$$B : \text{observe a number less than 3}$$

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:

$$P(A \cup B) = P(1) + P(2) + P(4) + P(6)$$
$$P(A) = P(2) + P(4) + P(6)$$
$$P(B) = P(1) + P(2)$$
$$P(A \cap B) = P(2)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
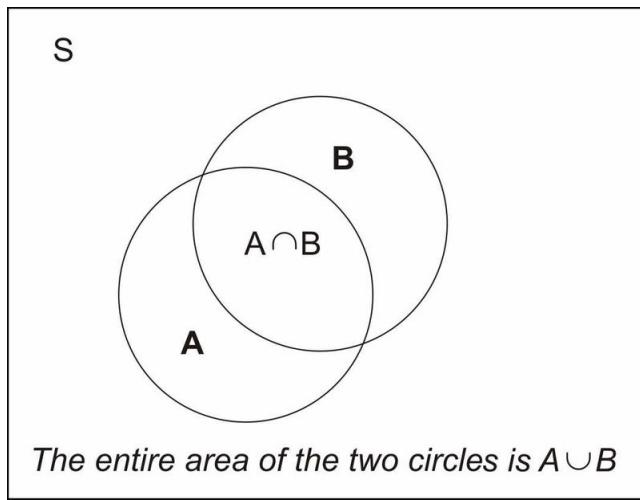
This is the *Additive Rule of Probability*, which is demonstrated below:

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

The entire area of the two circles is $A \cup B$

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.

Additive 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 : 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 Additive Rule of Probability. First, define two events as follows:

C : card selected is a spade

D : 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 Additive Rule of Probability formula:

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

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:

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

$$B : \text{person selected in under 18}$$

Also, keep in mind that the following probabilities have been given to us:
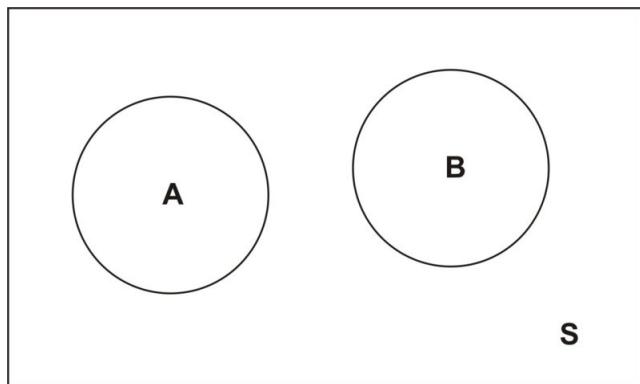
$$P(A) = 0.842 \qquad\qquad P(B) = 0.183 \qquad\qquad P(A \cap B) = 0.141$$

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.842 + 0.183 - 0.141$$
$$= 0.884$$
$$= 88.4\%$$

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 $(A \cap B = \varnothing)$, 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:

$$A : \text{observe only one head}$$

$$B : \text{observe two heads}$$

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

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

Multiplicative 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 *Multiplicative Rule of Probability*.

The Multiplicative 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:

$$F : \text{randomly selected worker who is female}$$

$$W : \text{randomly selected white-collar worker}$$

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 Multiplicative 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:

$$F1 : \text{first student selected is female}$$

$$M2 : \text{second student selected is male}$$

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 Multiplicative 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:

$$P(M2|F1) = P(M2) = \frac{17}{41} = 0.415$$

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:

$$A : \text{first toss is a head}$$

$$B : \text{second toss is a head}$$

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:

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

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

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

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 3.3:

	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
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 outcomes of one event are mutually exclusive (they have no overlap), they are not independent. If you know that outcomes 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 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, or $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
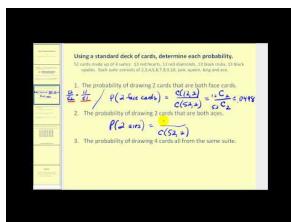
If $A \cap B$ contains no simple events, then A and B are mutually exclusive. Mathematically, this means $P(A \cup B) = P(A) + P(B)$.

The Multiplicative 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 $P(A \cap B) = P(A) \bullet P(B)$.

Multimedia Links

For an explanation of how to find probabilities using the Multiplicative and Additive Rules with combination notation **(1.0)**, see bullcleo1, Determining Probability (9:42).
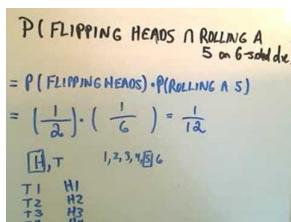
| MEDIA |
| Click image to the left or use the URL below. |
| URL: http://www.ck12.org/flx/render/embeddedobject/1065 |

For an explanation of how to find the probability of 'and' statements and independent events **(1.0)**, see patrickJMT, Calculating Probability - "And" Statements, Independent Events (8:04).

| MEDIA |
| Click image to the left or use the URL below. |
| URL: http://www.ck12.org/flx/render/embeddedobject/1066 |

Review Questions

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

$$A : \text{sum of the numbers is odd}$$
$$B : \text{sum of the numbers is 9, 11, or 12}$$

    a. Are events A and B independent? Why or why not?
    b. Are events A and B mutually exclusive? Why or why not?

2. 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?

3.6 Basic Counting Rules

Learning Objectives

- Understand the definition of simple random sample.
- Calculate ordered arrangements using factorials.
- Calculate combinations and permutations.
- Calculate probabilities with factorials.

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 learn about simple random samples. You will also be presented with the various counting rules used to calculate probabilities.

Simple Random Sample

A *simple random sample* of size *n* is one in which all samples of size *n* are equally likely to be selected. In other words, if *n* elements are selected from a population in such a way that every set of *n* elements in the population has an equal probability of being selected, then the *n* elements form a simple random sample.

Example: Suppose you randomly select 4 cards from a deck of playing cards, and all the cards selected are kings. Would you conclude that the deck is an ordinary deck, or would you conclude the deck is not an ordinary one and probably contains more than 4 kings?

The answer depends on how the cards were drawn. It is possible that the 4 kings were intentionally put on top of the deck, and hence, the drawing of the 4 kings was not unusual, and in fact, it was actually certain. However, if the deck was shuffled well, getting 4 kings is highly improbable.

Example: Suppose a lottery consists of 100 tickets, and one winning ticket is to be chosen. What would be a fair method of selecting a winning ticket?

First, we must require that each ticket has an equal chance of winning. That is, each ticket must have a probability of $\frac{1}{100}$ of being selected. One fair way of doing this is to mix up all the tickets in a container and blindly pick one ticket. This is an example of random sampling.
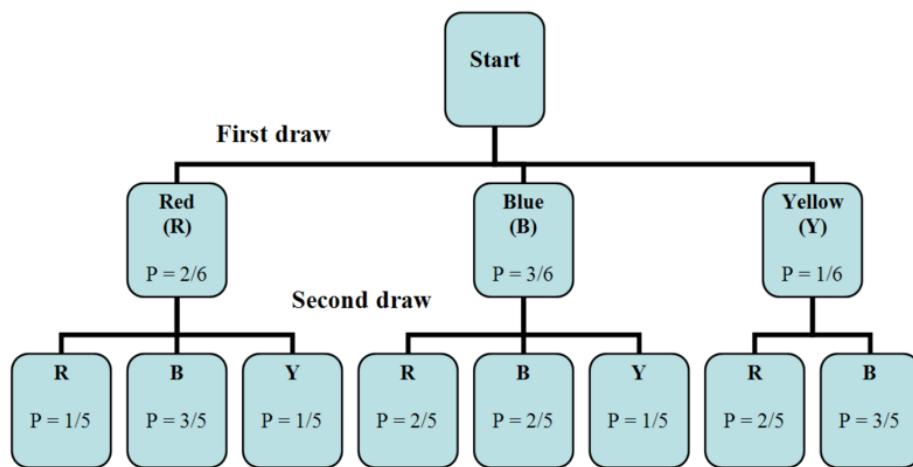
However, this method would not be too practical if we were dealing with a very large population, such as, say, a

million tickets, and we were asked to select 5 winning tickets. One method of picking a simple random sample is to give each element in the population a number. Then use a random number generator to pick 5 numbers. The people who were assigned one of the five numbers would then be the winners.

Some experiments have so many simple events that it is impractical to list them all. Tree diagrams are helpful in determining probabilities in these situations.

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, 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.

When taking the first ball and the second ball into account, 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.

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

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 $\left(\frac{2}{6}\right)\left(\frac{1}{5}\right)$. The method used to solve the example above can be generalized to any number of stages.

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 above 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.

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*.

Another way of stating this is as follows:

(II) Suppose you have *k* sets of elements, with n_1 elements in the first set, n_2 elements in the second set, and n_k elements in the kth set, and you want to take one sample from each of the *k* sets. The number of different samples that can be formed is the product $n_1 n_2 n_3 \ldots . n_k$.

Example: In how many different ways can you seat 8 people at a dinner table?

For the first seat, there are eight 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 seat, 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

$$n! = n(n-1)(n-2)(n-3).....(3)(2)(1)$$

Example: Suppose there are 30 candidates that are competing for three executive positions. How many different ways can you fill the three positions?

Since there are three executive positions and 30 candidates, let $n_1 =$ the number of candidates that are available to fill the first position, $n_2 =$ the number of candidates remaining to fill the second position, and $n_3 =$ the number of candidates remaining to fill the third position.

Hence, we have the following:

$$n_1 = 30$$
$$n_2 = 29$$
$$n_3 = 28$$

The number of different ways to fill the three executive positions with the given candidates is $(n_1)(n_2)(n_3) = (30)(29)(28) = 24,360$.

The arrangement of elements in a distinct order, as the example above shows, is called a *permutation*. Thus, from the example above, there are 24,360 possible permutations of three elements drawn from a set of 30 elements.

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 $P^n_r = \frac{n!}{(n-r)!}$.

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:

$$\begin{aligned} P^n_r &= \frac{n!}{(n-r)!} \\ &= \frac{8!}{(8-5)!} \\ &= \frac{8!}{3!} \\ &= \frac{40,320}{6} \\ &= 6,720 \end{aligned}$$

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.

A random sampling is a procedure in which each sample of a given size is equally likely to be selected.

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 $n! = n(n-1)(n-2)(n-3).....(3)(2)(1)$.

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
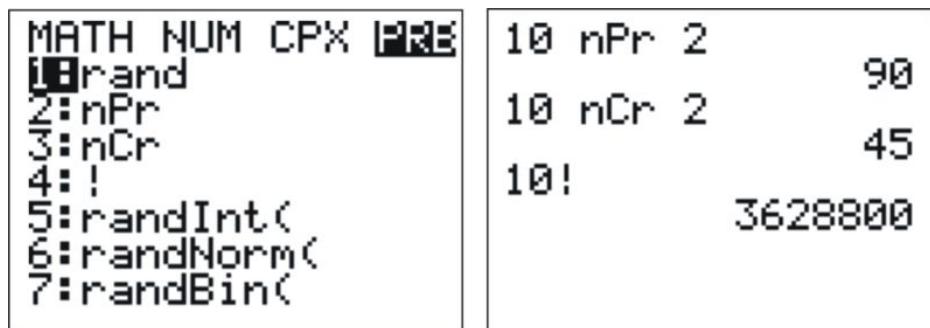
1. Determine the number of simple events when you toss a coin the following number of times. (Hint: As the numbers get higher, you will need to develop a systematic method of counting all the outcomes.)

   a. Twice
   b. Three times
   c. Five times
   d. n times (Look for a pattern in the results of a) through c).)

2. Flying into Los Angeles from Washington DC, you can choose one of three airlines and can also choose either first class or economy. How many travel options do you have?
3. How many different 5-card hands can be chosen from a 52-card deck?
4. 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?

Technology Note: Generating Random Numbers on the TI-83/84 Calculator

Press **[MATH]**, and then scroll to the right and choose **PRB**. Next, choose '1:rand' and press **[ENTER]** twice. The calculator returns a random number between 0 and 1. If you are taking a sample of 100, you need to use the first two digits of the random number that has been returned. If the calculator returns the same first two digits more than once, you can ignore them and press **[ENTER]** again.

Technology Note: Computing Factorials, Permutations and Combination on the TI-83/84 Calculator

Press **[MATH]**, and then scroll to the right and choose **PRB**. You will see the following choices, among others: '2:nPr', '3:nCr'. and '4:!'. The screenshots below show the menu and the proper uses of these commands.

Technology Note: Using EXCEL to Computer Factorials, Permutations and Combinations

In Excel, the commands shown above are entered as follows:

=PERMUT(10,2)

=COMBIN(10,2)

=FACT(10)

Keywords

Additive Rule of Probability

Classical probability

Combinations

Complement

Complement Rule

Compound event

Conditional probability

Counting Rule for Combinations

Counting Rule for Permutations

Event

Experiment

Factorial

Independent events

Intersection of events

Multiplicative Rule of Counting

Multiplicative Rule of Probability

Mutually exclusive

Natural frequencies approach

Permutation

Sample space

Simple events

Simple random sample

Tree diagram

Union of events

Venn diagram

CHAPTER **4**

Discrete Probability Distribution

Chapter Outline

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 $x = 0, 1, 2, 3, \ldots$. 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
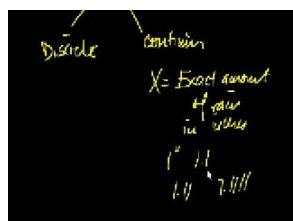
A random variable represents the numerical value of a simple event of an experiment.

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.

Multimedia Links

For an introduction to random variables and probability distribution functions **(3.0)**, see khanacademy, Introduction to Random Variables (12:04).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1067
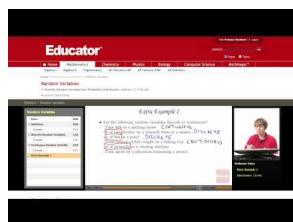
For examples of discrete and continuous random variables **(3.0)**, see EducatorVids, Statistics: Random Variables (Discrete or Continuous) (1:54).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1068

4.2 Probability Distribution for a Discrete Random Variable

Learning Objectives

- Know and understand the notion of discrete random variables.
- Learn how to use discrete random variables to solve 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:

$$P(x = 0) = P(TT) = \frac{1}{4}$$
$$P(x = 1) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$P(x = 2) = P(HH) = \frac{1}{4}$$

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.
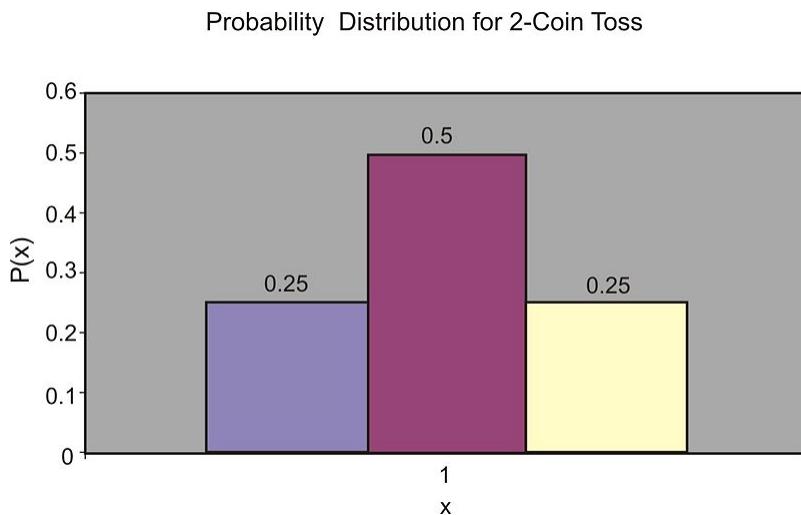
In tabular form:

TABLE 4.2:

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 the Chapter Introduction.)

Since each of the 8 outcomes is equally likely, the following table gives the probability of each value of the random variable.

TABLE 4.3:

Value of Random Variable (Number of Yes Votes)	Probability
3	$\frac{1}{8} = 0.125$
2	$\frac{3}{8} = 0.375$
1	$\frac{3}{8} = 0.375$
0	$\frac{1}{8} = 0.125$

Figure: Tabular representation of the probability distribution for the random variable in the first example in the Chapter Introduction.

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?

Mean and Standard Deviation of Discrete Random Variables

Learning Objectives

- Know the definition of the mean, or expected value, of a discrete random variable.
- Know the definition of the standard deviation of a discrete random variable.
- Know the definition of the variance of a discrete random variable.
- Find the expected value of a 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*.
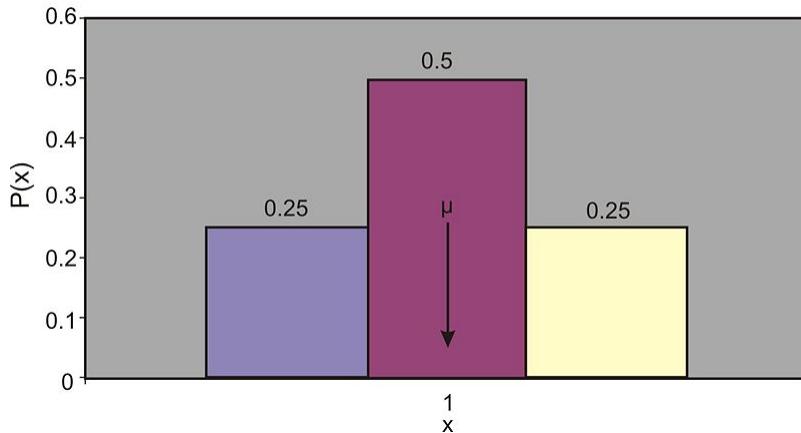
Example: Recall the probability distribution of the 2-coin experiment. Calculate the mean of this distribution.

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 4.4:

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:

$$\mu = E(x) = \sum xp(x)$$

$$\mu = (310)(99.9\%) + (310 - 15{,}000)(0.1\%)$$
$$= (310)(0.999) + (310 - 15{,}000)(0.001)$$
$$= 309.69 - 14.69 = \$295$$
$$\mu = \$295$$

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:

$$\mu = (0)(99.9\%) + (15{,}000)(0.1\%)$$
$$= 0 + 15$$
$$\mu = \$15$$

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:

$$\mu = E(x) = \sum xp(x)$$

$$\mu = (0)(0.002) + (1)(0.029) + (2)(0.132) + (3)(0.309) + (4)(0.360) + (5)(0.168)$$
$$\mu = 3.50$$
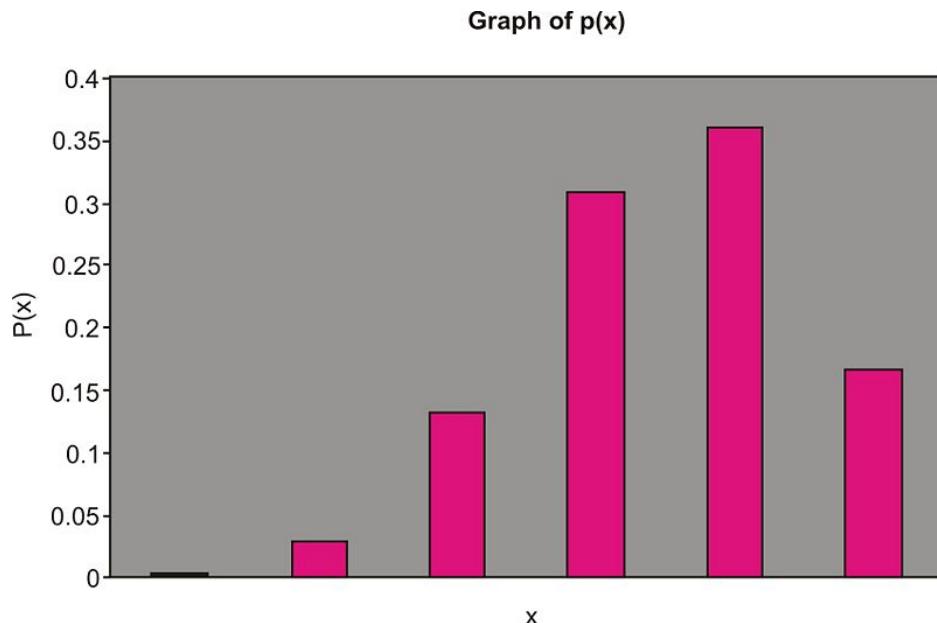
b. To find σ, we first calculate the variance of X:

$$\sigma^2 = \sum (x - \mu)^2 p(x)$$

$$= (0 - 3.5)^2(0.002) + (1 - 3.5)^2(0.029) + (2 - 3.5)^2(0.132)$$
$$+ (3 - 3.5)^2(0.309) + (4 - 3.5)^2(0.36) + (5 - 3.5)^2(0.168)$$
$$\sigma^2 = 1.05$$

Now we calculate the standard deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} = 1.02$$

c. The graph of $p(x)$ is shown below:

Graph of p(x)

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 value, or expected value, of the discrete random variable X is given by $\mu = E(x) = \sum x p(x)$.

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: $\sigma = \sqrt{\sigma^2}$.

Multimedia Links

For an example of finding the mean and standard deviation of discrete random variables **(5.0)(6.0)**, see Educator Vids, Statistics: Mean and Standard Deviation of a Discrete Random Variable (2:25).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1069

For a video presentation showing the computation of the variance and standard deviation of a set of data **(11.0)**, see American Public University, Calculating Variance and Standard Deviation (8:52).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1070
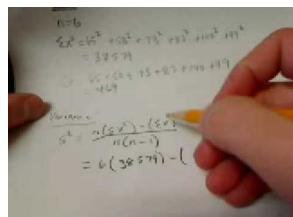
For an additional video presentation showing the calculation of the variance and standard deviation of a set of data **(11.0)**, see Calculating Variance and Standard Deviation (4:36).

> MEDIA
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1071

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

Figure: The probability distribution for question 1.

   a. Find the mean of the distribution.
   b. Find the variance.
   c. Find the standard deviation.

2. An officer at a prison was studying recidivism among the prison inmates. The officer questioned each inmate to find out how many times the inmate had been convicted prior to the inmate's current conviction. The officer came up with the following table that shows the relative frequencies of X, the number of times previously convicted: X, the number of times convicted:

x	0	1	2	3	4
$p(x)$	0.16	0.53	0.20	0.08	0.03

Figure: The probability distribution for question 2. If we regard the relative frequencies as approximate probabilities, what is the expected value of the number of previous convictions of an inmate?

4.4 Sums and Differences of Independent Random Variables

Learning Objectives

- Construct 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.

Sums and Differences of Independent Random Variables

Example: Create a table that shows all the possible outcomes when two dice are rolled simultaneously. (Hint: There are 36 possible outcomes.)

TABLE 4.5:

	2nd Die						
	1	2	3	4	5	6	
1	1, 1	1, 2	1, 3	1, 4	1, 5	1, 6	
2	2, 1	2, 2	2, 3	2, 4	2, 5	2, 6	
3	3, 1	3, 2	3, 3	3, 4	3, 5	3, 6	1st Die
4	4, 1	4, 2	4, 3	4, 4	4, 5	4, 6	
5	5, 1	5, 2	5, 3	5, 4	5, 5	5, 6	
6	6, 1	6, 2	6, 3	6, 4	6, 5	6, 6	

This table of possible outcomes when two dice are rolled simultaneously that is shown above can now be used to construct various probability distributions. The first table below displays the probabilities for all the possible sums of the two dice, and the second table shows the probabilities for each of the possible results for the larger of the two numbers produced by the dice.

TABLE 4.6:

Sum of Two Dice, x	Probability, $p(x)$
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$

<div align="center">TABLE 4.6: (continued)</div>

Sum of Two Dice, x	Probability, $p(x)$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$
Total	1

<div align="center">TABLE 4.7:</div>

Larger Number, x	Probability, $p(x)$
1	$\frac{1}{36}$
2	$\frac{3}{36}$
3	$\frac{5}{36}$
4	$\frac{7}{36}$
5	$\frac{9}{36}$
6	$\frac{11}{36}$
Total	1

When you roll the two dice, what is the probability that the sum is 4? By looking at the first table above, you can see that the probability is $\frac{3}{36}$.

What is the probability that the larger number is 4? By looking at the second table above, you can see that the probability is $\frac{7}{36}$.

Example: The Regional Hospital has recently opened a new pulmonary unit and has released the following data on the proportion of silicosis cases caused by working in the coal mines. Suppose two silicosis patients are randomly selected from a large population with the disease.

<div align="center">TABLE 4.8:</div>

Silicosis Cases	Proportion
Worked in the mine	0.80
Did not work in the mine	0.20

There are four possible outcomes for the two patients. With 'yes' representing "worked in the mines" and 'no' representing "did not work in the mines", the possibilities are as follows:

<div align="center">TABLE 4.9:</div>

	First Patient	Second Patient
1	No	No
2	Yes	No
3	No	Yes
4	Yes	Yes

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 multiplying the appropriate proportions as shown:

$$P(\text{no for } 1^{\text{st}}) \bullet P(\text{no for } 2^{\text{nd}}) = (0.2)(0.2) = 0.04$$
$$P(\text{yes for } 1^{\text{st}}) \bullet P(\text{no for } 2^{\text{nd}}) = (0.8)(0.2) = 0.16$$
$$P(\text{no for } 1^{\text{st}}) \bullet P(\text{yes for } 2^{\text{nd}}) = (0.2)(0.8) = 0.16$$
$$P(\text{yes for } 1^{\text{st}}) \bullet P(\text{yes for } 2^{\text{nd}}) = (0.8)(0.8) = 0.64$$

If X represents the number 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 disjoint, their probabilities can be added. The probability distribution for X is given in the table below:

TABLE 4.10:

x	Probability, $p(x)$
0	0.04
1	$0.16 + 0.16 = 0.32$
2	0.64

Expected Values and Standard Deviation

Example: Suppose an individual plays a gambling game where it is possible to lose $2.00, break even, win $6.00, or win $20.00 each time he plays. The probability distribution for each outcome is provided by the following table:

TABLE 4.11:

Winnings, x	Probability, $p(x)$
$-\$2$	0.30
$\$0$	0.40
$\$6$	0.20
$\$20$	0.10

The table can be used to calculate the expected value and the variance of this distribution:

$$\mu = \sum x p(x)$$
$$\mu = (-2 \cdot 0.30) + (0 \cdot 0.40) + (6 \cdot 0.20) + (20 \cdot 0.10)$$
$$\mu = 2.6$$

Thus, the player can expect to win $2.60 playing this game.

The variance of this distribution can be calculated as shown:

$$\sigma^2 = \sum (x - \mu)^2 p(x)$$
$$\sigma^2 = (-2 - 2.6)^2 (0.30) + (0 - 2.6)^2 (0.40) + (6 - 2.6)^2 (0.20) + (20 - 2.6)^2 (0.10)$$
$$\sigma^2 \approx 41.64$$
$$\sigma \approx \sqrt{41.64} \approx \$6.45$$

Example: The following probability distribution was constructed from the results of a survey at the local university. The random variable is the number of fast food meals purchased by a student during the preceding year (12 months). For this distribution, calculate the expected value and the standard deviation.

TABLE 4.12:

Number of Meals Purchased Within 12 Months, x	Probability, $p(x)$
0	0.04
$[1-6)$	0.30
$[6-11)$	0.29
$[11-21)$	0.17
$[21-51)$	0.15
$[51-60)$	0.05
Total	1.00

You must begin by estimating a mean for each interval, and this can be done by finding the center of each interval. For the first interval of $[1-6)$, 6 is not included in the interval, so a value of 3 would be the center. This same procedure can be used to estimate the mean of all the intervals. Therefore, the expected value can be calculated as follows:

$$\mu = \sum xp(x)$$
$$\mu = (0)(0.04) + (3)(0.30) + (8)(0.29) + (15.5)(0.17) + (35.5)(0.15) + (55)(0.05)$$
$$\mu = 13.93$$
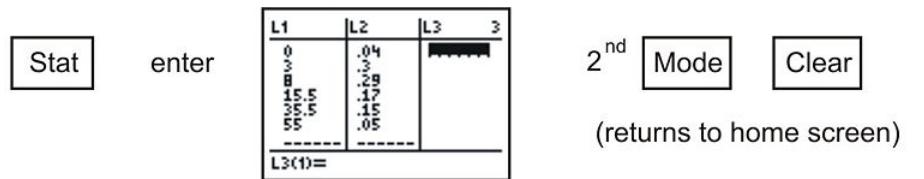
Likewise, the standard deviation can be calculated:

$$\sigma^2 = \sum (x-\mu)^2 p(x)$$
$$= (0-13.93)^2(0.04) + (3-13.93)^2(0.30)$$
$$+ (8-13.93)^2(0.29) + (15.5-13.93)^2(0.17)$$
$$+ (35.5-13.93)^2(0.15) + (55-13.93)^2(0.05)$$
$$\approx 208.3451$$

$$\sigma \approx 14.43$$

Thus, the expected number of fast food meals purchased by a student at the local university is 13.93, and the standard deviation is 14.43. Note that the mean should not be rounded, since it does not have to be one of the values in the distribution. You should also notice that the standard deviation is very close to the expected value. This means that the distribution will be skewed to the right and have a long tail toward the larger numbers.

Technology Note: Calculating mean and variance for probability distribution on TI-83/84 Calculator

Notice that the mean, which is denoted by $\bar{x}$ in this case, is 13.93, and the standard deviation, which is denoted by σ_x, is approximately 14.43.

Linear Transformations of

If you add the same value to all the numbers of a data set, the shape and standard deviation of the data set remain the same, but the value is added to the mean. This is referred to as *re-centering* the data set. Likewise, if you *rescale* the data, or multiply all the data values by the same nonzero number, the basic shape will not change, but the mean and the standard deviation will each be a multiple of this number. (Note that the standard deviation must actually be multiplied by the absolute value of the number.) If you multiply the numbers of a data set by a constant d and then add a constant c, the mean and the standard deviation of the transformed values are expressed as follows:

$$\mu_{c+dX} = c + d\mu_X$$
$$\sigma_{c+dX} = |d|\sigma_X$$

These are called *linear transformations*, and the implications of this can be better understood if you return to the casino example.
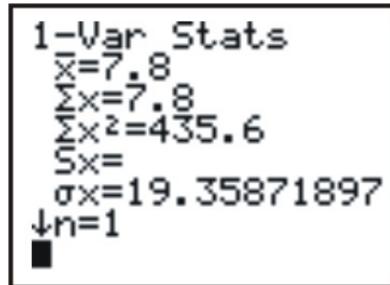
Example: The casino has decided to triple the prizes for the game being played. What are the expected winnings for a person who plays one game? What is the standard deviation? Recall that the expected value was $2.60, and the standard deviation was $6.45.

Solution:

The simplest way to calculate the expected value of the tripled prize is (3)($2.60), or $7.80, with a standard deviation of (3)($6.45), or $19.35. Here, $c = 0$ and $d = 3$. Another method of calculating the expected value and standard deviation would be to create a new table for the tripled prize:

TABLE 4.13:

Winnings, x	Probability, p
−$6	0.30
$0	0.40
$18	0.20
$60	0.10

The calculations can be done using the formulas or by using a graphing calculator. Notice that the results are the same either way.

This same problem can be changed again in order to introduce the *Addition Rule* and the *Subtraction Rule* for random variables. Suppose the casino wants to encourage customers to play more, so it begins demanding that customers play the game in sets of three. What are the expected value (total winnings) and standard deviation now?

Let X, Y and Z represent the total winnings on each game played. If this is the case, then μ_{X+Y+Z} is the expected value of the total winnings when three games are played. The expected value of the total winnings for playing one game was \$2.60, so for three games the expected value is:

$$\mu_{X+Y+Z} = \mu_X + \mu_Y + \mu_Z$$
$$\mu_{X+Y+Z} = \$2.60 + \$2.60 + \$2.60$$
$$\mu_{X+Y+Z} = \$7.80$$

Thus, the expected value is the same as that for the tripled prize.

Since the winnings on the three games played are independent, the standard deviation of X, Y and Z can be calculated as shown below:

$$\sigma^2_{X+Y+Z} = \sigma^2_X + \sigma^2_Y + \sigma^2_Z$$
$$\sigma^2_{X+Y+Z} = 6.45^2 + 6.45^2 + 6.45^2$$
$$\sigma^2_{X+Y+Z} \approx 124.8075$$
$$\sigma_{X+Y+Z} \approx \sqrt{124.8075}$$
$$\sigma_{X+Y+Z} \approx 11.17$$

This means that the person playing the three games can expect to win \$7.80 with a standard deviation of \$11.17. Note that when the prize was tripled, there was a greater standard deviation (\$19.36) than when the person played three games (\$11.17).

The Addition and Subtraction Rules for random variables are as follows:

If X and Y are random variables, then:

$$\mu_{X+Y} = \mu_X + \mu_Y$$
$$\mu_{X-Y} = \mu_X - \mu_Y$$

If X and Y are independent, then:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Variances are added for both the sum and difference of two independent random variables, because the variation in each variable contributes to the overall variation in both cases. (Subtracting is the same as adding the opposite.) Suppose you have two dice, one die, X, with the usual positive numbers 1 through 6, and another, Y, with the negative numbers -1 through -6. Next, suppose you perform two experiments. In the first, you roll the first die, X, and then the second die, Y, and you compute the difference of the two rolls. In the second experiment, you roll the first die and then the second die, and you calculate the sum of the two rolls.

Difference (X-Y)

Difference	Probability
12	1/36
11	2/36
10	3/36
9	4/36
8	5/36
7	6/36
6	5/36
5	4/36
4	3/36
3	2/36
2	1/36

Sum (X+Y)

Sum	Probability
5	1/36
4	2/36
3	3/36
2	4/36
1	5/36
0	6/36
-1	5/36
-2	4/36
-3	3/36
-4	2/36
-5	1/36

$$\mu_X = \sum x p(x)$$
$$\mu_X = 3.5$$
$$\sigma^2_X \approx \sum (x - \mu_X)^2 p(x)$$
$$\sigma^2_X \approx 2.917$$
$$\mu_{X-Y} = \mu_X - \mu_Y$$
$$\mu_{X-Y} = 3.5 - (-3.5) = 7$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} \approx 2.917 + 2.917 = 5.834$$

$$\mu_Y = \sum y p(y)$$
$$\mu_Y = -3.5$$
$$\sigma^2_Y \approx \sum (y - \mu_Y)^2 p(y)$$
$$\sigma^2_Y \approx 2.917$$
$$\mu_{X+Y} = \mu_X + \mu_Y$$
$$\mu_{X+Y} = 3.5 + (-3.5) = 0$$
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X+Y} \approx 2.917 + 2.917 = 5.834$$

Notice how the expected values and the variances for the two dice combine in these two experiments.

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 4.14:

Hours of Piano Lessons, x	Probability, $p(x)$
0	0.3
1	0.3
2	0.4

146

TABLE 4.15:

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:

$$E(x) = \mu_X = \sum x p(x) \qquad \sigma^2_X = \sum (x - \mu_X)^2 p(x)$$
$$\mu_X = (0)(0.3) + (1)(0.3) + (2)(0.4) \qquad \sigma^2_X = (0 - 1.1)^2(0.3) + (1 - 1.1)^2(0.3) + (2 - 1.1)^2(0.4)$$
$$\mu_X = 1.1 \qquad \sigma^2_X = 0.69$$
$$\sigma_X = 0.831$$

$$E(y) = \mu_Y = \sum y p(y) \qquad \sigma^2_Y = \sum (y - \mu_Y)^2 p(y)$$
$$\mu_Y = (1)(0.2) + (2)(0.3) + (3)(0.2) + (4)(0.3) \qquad \sigma^2_Y = (1 - 2.6)^2(0.2) + (2 - 2.6)^2(0.3) + (3 - 2.6)^2(0.2)$$
$$+ (4 - 2.6)^2(0.3)$$
$$\mu_Y = 2.6 \qquad \sigma^2_Y = 1.24$$
$$\sigma_Y = 1.11$$

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 with the *Linear Transformation Rule* as shown:

$\mu_{20X} = (20)(\mu_X) = (20)(1.1) = \$22$ by the Linear Transformation Rule.

Beth earns $25 for each hour of tutoring, so her mean weekly earnings from tutoring are as follows:

$\mu_{25Y} = (25)(\mu_Y) = (25)(2.6) = \$65$ by the Linear Transformation Rule.

Thus, Beth's expected weekly savings are:

$\mu_{25Y} - \mu_{20X} = \$65 - \$22 = \$43$ by the Subtraction Rule.

The standard deviation of the cost of her piano lessons is:

$\sigma_{20X} = (20)(0.831) = \$16.62$ by the Linear Transformation Rule.

The standard deviation of her earnings from tutoring is:

$\sigma_{25Y} = (25)(1.11) = \$27.75$ by the Linear Transformation Rule.

Finally, the variance and standard deviation of her weekly savings is:

$$\sigma^2_{25Y-20X} = \sigma^2_{25Y} + \sigma^2_{20X} = (27.75)^2 + (16.62)^2 = 1046.2896$$
$$\sigma_{25Y-20X} \approx \$32.35$$

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:

$$E(x) = \mu_X = \sum x p(x)$$
$$\sigma^2_X = \sum (x - \mu_X)^2 p(x)$$

For the random variables X and Y and constants c and d, the mean and the standard deviation of a linear transformation are given by the following:

$$\mu_{c+dX} = c + d\mu_X$$
$$\sigma_{c+dX} = |d|\sigma_X$$

If the random variables X and Y are added or subtracted, the mean is calculated as shown below:

$$\mu_{X+Y} = \mu_X + \mu_Y$$
$$\mu_{X-Y} = \mu_X - \mu_Y$$

If X and Y are independent, then the following formulas can be used to compute the variance:
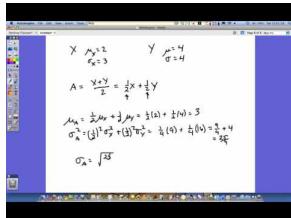
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Points to Consider

- Are these concepts applicable to real-life situations?
- Will knowing these concepts allow you estimate information about a population?

Multimedia Links

For examples of finding means and standard deviations of sums and differences of random variables **(5.0)**, see mrjaffesclass, Linear Combinations of Random Variables (6:41).

Review Questions

1. 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. What is the expected value of a 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?

3. A recent law has been passed banning the use of hand-held cell phones while driving, and a survey has revealed that 76% of drivers now refrain from using their cell phones while driving. Three drivers were randomly selected, and a probability distribution table was constructed to record the outcomes. Let N represent those drivers who never use their cell phones while driving and S represent those who do use their cell phones while driving. Calculate the expected value and the variance using your calculator.

4.5 The Binomial Probability Distribution

Learning Objectives

- Know the characteristics of a binomial random variable.
- Understand a binomial probability distribution.
- Know the definitions of 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.

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.

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 two other leading shampoos on the market. Let X be the number of the 100 customers who choose the new brand over the other two.

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 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 4.16:

Number Who Correctly Identify Percolated Coffee	Outcomes, C (correct), I (incorrect)	Number of Outcomes
0	IIII	1
1	ICII IIIC IICI CIII	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 $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{16}\right)$. The table shows six outcomes where two people guessed correctly, so the probability of getting two people who correctly identified the percolated coffee is $\frac{6}{16}$. Another way to determine the number of ways that exactly two people out of four people can identify the percolated coffee is simply to count how many ways two people can be selected from four people:

$$_4C_2 = \frac{4!}{2!\,2!} = \frac{24}{4} = 6$$

In addition, a graphing calculator can also be used to calculate binomial probabilities.

By pressing **[2ND][DISTR]**, you can enter 'binompdf (4,0.5,2)'. This command calculates the binomial probability for k (in this example, $k = 2$) successes out of n (in this example, $n = 4$) trials, when the probability of success on any one trial is p (in this example, $p = 0.5$).

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:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Let's return to the coffee experiment and look at the distribution of X (correct guesses):

TABLE 4.17:

k	$P(x = k)$
0	$\frac{1}{16}$
1	$\frac{4}{16}$
2	$\frac{6}{16}$
3	$\frac{4}{16}$
4	$\frac{1}{16}$

The expected value for the above distribution can be calculated as follows:

$$E(x) = (0)\left(\frac{1}{16}\right) + (1)\left(\frac{4}{16}\right) + (2)\left(\frac{6}{16}\right) + (3)\left(\frac{4}{16}\right) + (4)\left(\frac{1}{16}\right)$$

$$E(x) = 2$$

In other words, you would expect half of the four to guess correctly when given two equally-likely choices. $E(x)$ can be written as $(4)\left(\frac{1}{2}\right)$, which is equivalent to np.

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:

$E(x) = \mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$

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.
- Use the binomial formula to write the probability distribution of X.

Example: According to a study conducted by a telephone company, the probability is 25% that a randomly selected phone call will last longer than the mean value of 3.8 minutes. What is the probability that out of three randomly selected calls:

a. Exactly two last longer than 3.8 minutes?

b. None last longer than 3.8 minutes?

Using the first three steps listed above:

- A success is any call that is longer than 3.8 minutes.
- The probability of success is $p = 0.25$.
- 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: $p(x) = \binom{3}{x}(0.25)^x (1-0.25)^{3-x}$

a. For $x = 2$:

$$
\begin{aligned}
p(x) &= \binom{3}{2}(0.25)^2(1-0.25)^{3-2} \\
&= (3)(0.25)^2(1-0.25)^1 \\
&= 0.14
\end{aligned}
$$

Thus, the probability is 0.14 that exactly two out of three randomly selected calls will last longer than 3.8 minutes.

b. Here, $x = 0$. 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 calls will last longer than 3.8 minutes.

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:

a. Exactly three customers?

b. At most one customer?

c. At least one customer?

Also, determine the probability distribution for the number of sales.

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:

$$p(x) = \binom{5}{x}(0.2)^x(0.8)^{5-x}$$

a. 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.

b. The probability that the car dealer makes a sale to at most one customer can be calculated as follows:

$$p(x \le 1) = p(0) + p(1)$$
$$= \binom{5}{0}(0.2)^0(0.8)^{5-0} + \binom{5}{1}(0.2)^1(0.8)^{5-1}$$
$$= 0.328 + 0.410 = 0.738$$

c. The probability that the car dealer makes at least one sale is the sum of the probabilities of him making 1, 2, 3, 4, or 5 sales, as is shown below:

$$p(x \ge 1) = p(1) + p(2) + p(3) + p(4) + p(5)$$

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:

$$p(x \ge 1) = 1 - p(x < 1) = 1 - p(x = 0)$$
$$1 - p(0) = 1 - \binom{5}{0}(0.2)^0(0.8)^{5-0}$$
$$= 1 - 0.328 = 0.672$$

This tells us that the salesperson has a probability of 0.672 of making at least one sale in five attempts.

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 $x = 1, 2, 3, 4,$ and 5. We can use the binomial probability formula 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.

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 $x \le 10$.

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:

$$\mu = np = (20)(0.6) = 12$$
$$\sigma^2 = np(1-p) = (20)(0.6)(0.4) = 4.8$$
$$\sigma = \sqrt{4.8} = 2.2$$

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:

$$p(x \leq 10) = p(0) + p(1) + p(2) + \ldots + p(10)$$

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. Using a calculator (see the technology note below) with $n = 20, p = 0.6$, and $k \leq 10$, we get a probability of 0.245 that $x \leq 10$.

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:

$$p(x > 12) = p(13) + p(14) + \ldots + p(20) = \sum_{x=13}^{20} p(x)$$

Alternatively, using the Complement Rule, $p(x > 12) = 1 - p(x \leq 12)$.

Using a calculator (see the technology note below) with $n = 20, p = 0.6$, and $k = 12$, 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, it's possible to subtract the probability that less than or equal to 10 voters favor the candidate from the probability that less than or equal to 11 voters favor the candidate. These probabilities can be found using a calculator. Thus, the probability that exactly 11 voters favor the candidate can be calculated as follows:

$$p(x = 11) = p(x \leq 11) - p(x \leq 10) = 0.404 - 0.245 = 0.159$$

A graphing calculator will now be used to graph and compare different versions of a binomial distribution. Each binomial distribution will be entered into two lists and then displayed as a histogram. First, we will use the calculator to generate a sequence of integers, and next, we will use it to generate a corresponding list of binomial probabilities.

To generate a sequence of integers, press **[2ND][LIST]**, go to **OPS**, select '5:seq', enter '(X, X, 0, n, 1)', where n is the number of independent binomial trials, and press **[STO][2ND][L1]**.

To enter the binomial probabilities associated with this sequence of integers, press **[STAT]** and select '1:EDIT'.

Clear out **L2** and position the cursor on the **L2** list name.

Press **[2ND][DISTR]** to bring up the list of distributions.

Select 'A:binompdf(' and enter 'n, p)', where n is the number of independent binomial trials and p is the probability of success.
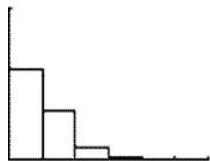
To graph the histogram, make sure your window is set correctly, press **[2ND][STAT PLOT]**, turn a plot on, select the histogram plot, choose **L1** for Xlist and **L2** for Freq, and press **[GRAPH]**. This will display the binomial histogram.

Horizontally, the following are examples of binomial distributions where n increases and p remains constant. Vertically, the examples display the results where n remains fixed and p increases.
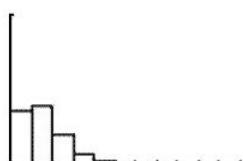
$n = 5$ and $p = 0.1$ $n = 10$ and $p = 0.1$ $n = 20$ and $p = 0.1$
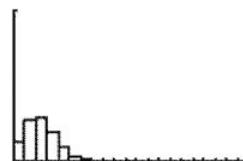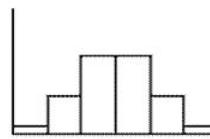
For a small value of p, the binomial distributions are skewed toward the higher values of X. As n increases, the skewness decreases and the distributions gradually move toward being more normal.

$n = 5$ and $p = 0.5$ $n = 10$ and $p = 0.5$ $n = 20$ and $p = 0.5$
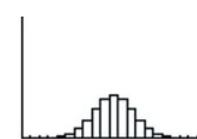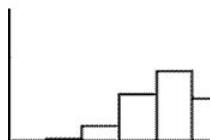
As p increases to 0.5, the skewness disappears and the distributions achieve perfect symmetry. The symmetrical, mound-shaped distribution remains the same for all values of n.

$n = 5$ and $p = 0.75$ $n = 10$ and $p = 0.75$ $n = 20$ and $p = 0.75$
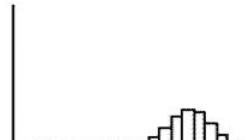
For a larger value of p, the binomial distributions are skewed toward the lower values of X. As n increases, the skewness decreases and the distributions gradually move toward being more normal.

Because $E(x) = np = \mu_X$, the expected value increases with both n and p. As n increases, so does the standard deviation, but for a fixed value of n, the standard deviation is largest around $p = 0.5$ and reduces as p approaches 0 or 1.

Technology Note: Calculating Binomial Probabilities on the TI-83/84 Calculator

Use the 'binompdf(' command to calculate the probability of exactly k successes. Press **[2ND][DIST]** and scroll down to 'A:binompdf('. Press **[ENTER]** to place 'binompdf(' on your home screen. Type values of n, p, and k, separated by commas, and press **[ENTER]**.

Use the 'binomcdf(' command to calculate the probability of at most x successes. The format is 'binomcdf(n, p, k)' to find the probability that $x \leq k$. (Note: It is not necessary to close the parentheses.)

Technology Note: Using Excel

In a cell, enter the function =binomdist(x, n, p,false). Press **[ENTER]**, and the probability of x successes will appear in the cell.

For the probability of at least x successes, replace 'false' with 'true'.

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 q. Thus, $q = 1 - p$.
- The trials are independent of each other.
- The binomial random variable X is the number of successes in n trials.
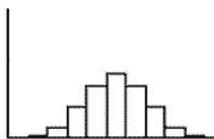
The binomial probability distribution is: $p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$.

For a binomial random variable, the mean is $\mu = np$.

The variance is $\sigma^2 = npq = np(1-p)$.
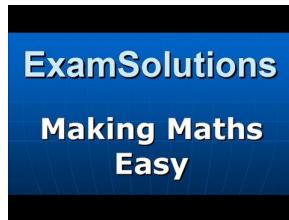
The standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$.

On the Web

http://tinyurl.com/268m56r Simulation of a binomial experiment. Explore what happens as you increase the number of trials.

Multimedia Links
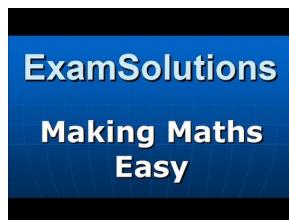
For an explanation of binomial distribution and notation used for it **(4.0)(7.0)**, see ExamSolutions, A-Level Statistics: Binomial Distribution (Introduction) (10:31).

ExamSolutions

Making Maths Easy

MEDIA

Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1073
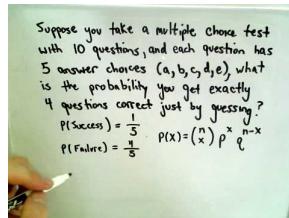
For an explanation on using tree diagrams and the formula for finding binomial probabilities **(4.0)(7.0)**, see Exam Solutions, A-Level Statistics: Binomial Distribution (Formula) (14:19).

ExamSolutions

Making Maths Easy

MEDIA

Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1074

For an explanation of using the binomial probability distribution to find probabilities **(4.0)**, see patrickJMT, The Binomial Distribution and Binomial Probability Function (6:45).

MEDIA

Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1075

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 $p = 0.5$.

   a. Display $p(x)$ in tabular form.

   a. Compute the mean and the variance of X.

3. Over the years, a medical researcher has found that one out of every ten diabetic patients receiving insulin develops antibodies against the hormone, thus, requiring a more costly form of medication.

   a. Find the probability that in the next five patients the researcher treats, none will develop antibodies against insulin.
   b. Find the probability that at least one 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.

   a. What is the expected number of households with annual incomes less than $56,400?
   b. What is the standard deviation of households with 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?

4.6 The Poisson Probability Distribution

Learning Objectives

- Know the definition of a Poisson distribution.
- Identify the characteristics of a Poisson distribution.
- Identify the type of statistical situation to which a Poisson distribution can be applied.
- Use a Poisson distribution to solve statistical problems.

Introduction

In this lesson, you will be introduced to Poisson distributions. Not only will you learn how to describe a Poisson distribution, but you will also learn how to apply the formula used with this type of distribution. Many real-world problems will be shown.

Poisson Distributions

A *Poisson probability distribution* is useful for describing the number of events that will occur during a specific interval of time or in a specific distance, area, or volume. Examples of such random variables are:

The number of traffic accidents at a particular intersection

The number of house fire claims per month that are received by an insurance company

The number of people who are infected with the AIDS virus in a certain neighborhood

The number of people who walk into a barber shop without an appointment

In a binomial distribution, if the number of trials, *n*, gets larger and larger as the probability of success, *p*, gets smaller and smaller, we obtain a Poisson distribution. The section below lists some of the basic characteristics of a Poisson distribution.

Characteristics of a Poisson Distribution

- The experiment consists of counting the number of events that will occur during a specific interval of time or in a specific distance, area, or volume.
- The probability that an event occurs in a given time, distance, area, or volume is the same.
- Each event is independent of all other events. For example, the number of people who arrive in the first hour is independent of the number who arrive in any other hour.

Poisson Random Variable

The probability distribution, mean, and variance of a Poisson random variable are given as follows:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, 3, \dots$$
$$\mu = \lambda$$
$$\sigma^2 = \lambda$$

where:

λ = the mean number of events in the time, distance, volume, or area

e = the base of the natural logarithm

Example: A lake, popular among boat fishermen, has an average catch of three fish every two hours during the month of October.

a. What is the probability distribution for X, the number of fish that you will catch in 7 hours?

b. What is the probability that you will catch 0 fish in seven hours of fishing? What is the probability of catching 3 fish? How about 10 fish?

c. What is the probability that you will catch 4 or more fish in 7 hours?

a. The mean number of fish is 3 fish in 2 hours, or 1.5 fish/hour. This means that over seven hours, the mean number of fish will be $\lambda = 1.5$ fish/hour $\bullet$ 7 hours $= 10.5$ fish. Thus, the equation becomes:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(10.5)^x e^{-10.5}}{x!}$$

b. To calculate the probabilities that you will catch 0, 3, or 10 fish, perform the following calculations:

$$p(0) = \frac{(10.5)^0 e^{-10.5}}{0!} \approx 0.000027 \approx 0\%$$
$$p(3) = \frac{(10.5)^3 e^{-10.5}}{3!} \approx 0.0053 \approx 0.5\%$$
$$p(10) = \frac{(10.5)^{10} e^{-10.5}}{10!} \approx 0.1236 \approx 12\%$$

This means that it is almost guaranteed that you will catch some fish in 7 hours.

c. The probability that you will catch 4 or more fish in 7 hours is equal to the sum of the probabilities that you will catch 4 fish, 5 fish, 6 fish, and so on, as is shown below:

$$p(x \geq 4) = p(4) + p(5) + p(6) + \ldots$$

The Complement Rule can be used to find this probability as follows:

$$\begin{aligned} p(x \geq 4) &= 1 - [p(0) + p(1) + p(2) + p(3)] \\ &\approx 1 - 0.000027 - 0.000289 - 0.00152 - 0.0053 \\ &\approx 0.9929 \end{aligned}$$

Therefore, there is about a 99% chance that you will catch 4 or more fish within a 7 hour period during the month of October.

Example: A zoologist is studying the number of times a rare kind of bird has been sighted. The random variable X is the number of times the bird is sighted every month. We assume that X has a Poisson distribution with a mean value of 2.5.

a. Find the mean and standard deviation of X.

b. Find the probability that exactly five birds are sighted in one month.

c. Find the probability that two or more birds are sighted in a 1-month period.

a. The mean and the variance are both equal to λ. Thus, the following is true:

$$\begin{aligned} \mu &= \lambda = 2.5 \\ \sigma^2 &= \lambda = 2.5 \end{aligned}$$

This means that the standard deviation is $\sigma = 1.58$.

b. Now we want to calculate the probability that exactly five birds are sighted in one month. For this, we use the Poisson distribution formula:

$$\begin{aligned} p(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ p(5) &= \frac{(2.5)^5 e^{-2.5}}{5!} \\ p(5) &= 0.067 \end{aligned}$$

c. The probability of two or more sightings is an infinite sum and is impossible to compute directly. However, we can use the Complement Rule as follows:

$$\begin{aligned} p(x \geq 2) &= 1 - p(x \leq 1) \\ &= 1 - [p(0) + p(1)] \\ &= 1 - \frac{(2.5)^0 e^{-2.5}}{0!} - \frac{(2.5)^1 e^{-2.5}}{1!} \\ &\approx 0.713 \end{aligned}$$

Therefore, according to the Poisson model, the probability that two or more sightings are made in a month is 0.713.

Technology Note: Calculating Poisson Probabilities on the TI-83/84 Calculator

Press **[2ND][DIST]** and scroll down to 'poissonpdf('. Press **[ENTER]** to place 'poissonpdf(' on your home screen. Type values of μ and x, separated by commas, and press **[ENTER]**.

Use 'poissoncdf(' for the probability of at most x successes.

Note: It is not necessary to close the parentheses.

Technology Note: Using Excel

In a cell, enter the function =Poisson(μ, x,false), where μ and x are numbers. Press **[ENTER]**, and the probability of x successes will appear in the cell.

For the probability of at least x successes, replace 'false' with 'true'.

Lesson Summary

Characteristics of a Poisson distribution:

- The experiment consists of counting the number of events that will occur during a specific interval of time or in a specific distance, area, or volume.
- The probability that an event occurs in a given time, distance, area, or volume is the same.
- Each event is independent of all other events.

Poisson Random Variable:

The probability distribution, mean, and variance of a Poisson random variable are given as follows:

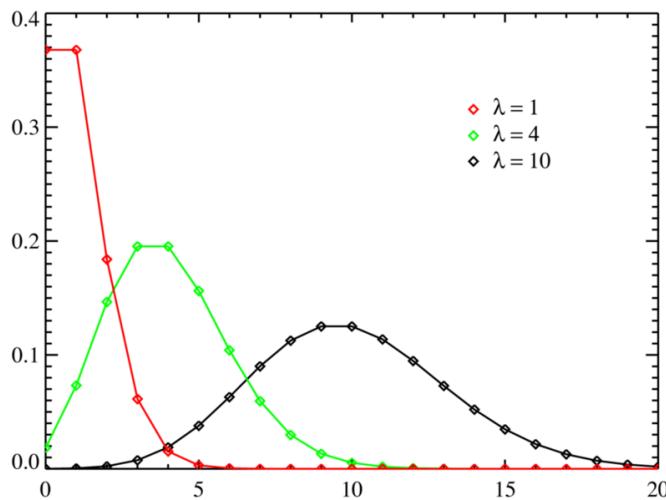$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, 3, \ldots$$
$$\mu = \lambda$$
$$\sigma^2 = \lambda$$
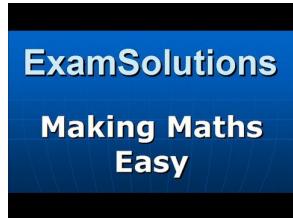
where:

λ = the mean number of events in the time, distance, volume or area

e = the base of the natural logarithm

Multimedia Links

For a discussion on the Poisson distribution and how to calculate probabilities **(4.0)(7.0)**, see ExamSolutions, Statistics: Poisson Distribution - Introduction (12:32).

For an example of finding probability in a Poisson situation **(7.0)**, see EducatorVids, Statistics: Poisson Probability Distribution (1:55).

Review Questions

A prison reports that the number of escape attempts per month has a Poisson distribution with a mean value of 1.5.

1. Calculate the probability that exactly three escapes will be attempted during the next month.
2. Calculate the probability that exactly one escape will be attempted during the next month.

4.7 Geometric Probability Distribution

Learning Objectives

- Know the definition of a geometric distribution.
- Identify the characteristics of a geometric distribution.
- Identify the type of statistical situation to which a geometric distribution can be applied.
- Use a geometric distribution to solve statistical problems.

Introduction

In this lesson, you will learn both the definition and the characteristics of a geometric probability distribution. In addition, you will be presented with a real-world problem that you can solve by applying what you have learned.

Geometric Probability Distributions

Like the Poisson and binomial distributions, a *geometric probability distribution* describes a discrete random variable. Recall that in the binomial experiments, we tossed a coin a fixed number of times and counted the number, X, of heads as successes.

A geometric distribution describes a situation in which we toss the coin until the first head (success) appears. We assume, as in the binomial experiments, that the tosses are independent of each other.

Characteristics of a Geometric Probability Distribution

- The experiment consists of a sequence of independent trials.
- Each trial results in one of two outcomes: success, S, or failure, F.
- The geometric random variable X is defined as the number of trials until the first S is observed.
- The probability $p(x)$ is the same for each trial.

Why would we wait until a success is observed? One example is in the world of business. A business owner may want to know the length of time a customer will wait for some type of service. Another example would be an employer who is interviewing potential candidates for a vacant position and wants to know how many interviews he/she has to conduct until the perfect candidate for the job is found. Finally, a police detective might want to know the probability of getting a lead in a crime case after 10 people are questioned.

Probability Distribution, Mean, and Variance of a Geometric Random Variable

The probability distribution, mean, and variance of a geometric random variable are given as follows:

$$p(x) = (1-p)^{x-1}p \quad x = 1,2,3,\ldots$$
$$\mu = \frac{1}{p}$$
$$\sigma^2 = \frac{1-p}{p^2}$$

where:

p = probability of an S outcome

x = the number of trials until the first S is observed

The figure below plots a few geometric probability distributions. Note how the probabilities start high and drop off, with lower p values producing a faster drop-off.

Example: A court is conducting a jury selection. Let X be the number of prospective jurors who will be examined until one is admitted as a juror for a trial. Suppose that X is a geometric random variable, and p, the probability of a juror being admitted, is 0.50.

a. Find the mean and the standard deviation of X.

b. Find the probability that more than two prospective jurors must be examined before one is admitted to the jury.

a. The mean and the standard deviation can be calculated as follows:

$$\mu = \frac{1}{p} = \frac{1}{0.5} = 2$$
$$\sigma^2 = \frac{1-p}{p^2} = \frac{1-0.5}{0.5^2} = 2$$
$$\sigma = \sqrt{2} = 1.41$$

To find the probability that more than two prospective jurors will be examined before one is selected, you could try to add the probabilities that the number of jurors to be examined before one is selected is 3, 4, 5, and so on, as follows:

$$p(x > 2) = p(3) + p(4) + p(5) + \ldots$$

However, this is an infinitely large sum, so it is best to use the Complement Rule as shown:

$$p(x > 2) = 1 - p(x \le 2)$$
$$= 1 - [p(1) + p(2)]$$

In order to actually calculate the probability, we need to find $p(1)$ and $p(2)$. This can be done by substituting the appropriate values into the formula:

$$p(1) = (1 - 0.5)^{1-1}(0.5) = (0.5)^0(0.5) = 0.5$$
$$p(2) = (1 - 0.5)^{2-1}(0.5) = (0.5)^1(0.5) = 0.25$$

Now we can go back to the Complement Rule and plug in the appropriate values for $p(1)$ and $p(2)$:

$$p(x > 2) = 1 - p(x \le 2)$$
$$= 1 - (0.5 + 0.25) = 0.25$$

This means that there is a 0.25 chance that more than two prospective jurors will be examined before one is admitted to the jury.

Technology Note: Calculating Geometric Probabilities on the TI-83/84 Calculator

Press **[2ND][DISTR]** and scroll down to 'geometpdf('. Press **[ENTER]** to place 'geometpdf(' on your home screen. Type in values of p and x separated by a comma, with p being the probability of success and x being the number of trials before you see your first success. Press **[ENTER]**. The calculator will return the probability of having the first success on trial number x.

Use 'geometcdf(' for the probability of at most x trials before your first success.

Note: It is not necessary to close the parentheses.

Lesson Summary

Characteristics of a Geometric Probability Distribution:

- The experiment consists of a sequence of independent trials.
- Each trial results in one of two outcomes: success, S, or failure, F.
- The geometric random variable X is defined as the number of trials until the first S is observed.
- The probability $p(x)$ is the same for each trial.

Geometric random variable:

The probability distribution, mean, and variance of a geometric random variable are given as follows:

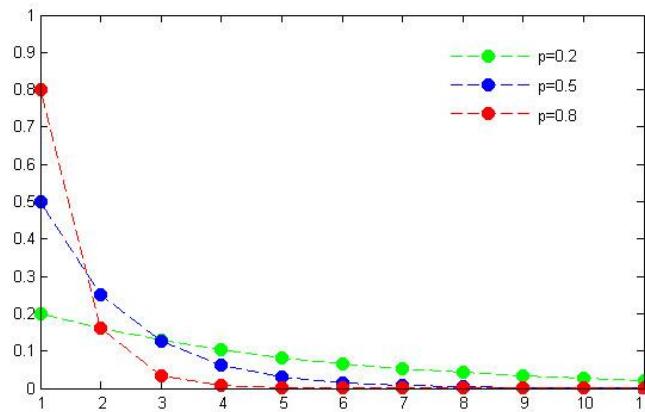$$p(x) = (1-p)^{x-1}p \quad x = 1, 2, 3, \dots$$
$$\mu = \frac{1}{p}$$
$$\sigma^2 = \frac{1-p}{p^2}$$

where:

p = probability of an S outcome

x = the number of trials until the first S is observed

Review Questions

1. The mean number of patients entering an emergency room at a hospital is 2.5. If the number of available beds today is 4 beds for new patients, what is the probability that the hospital will not have enough beds to accommodate its new patients?
2. An oil company has determined that the probability of finding oil at a particular drilling operation is 0.20. What is the probability that it would drill four dry wells before finding oil at the fifth one? (Hint: This is an example of a geometric random variable.)

Keywords

Addition Rule

Binomial experiment

Binomial probability distribution

Continuous

Continuous random variables

Discrete

Discrete random variables

Expected value

Geometric probability distribution

Linear Transformation Rule

Linear transformations

Poisson probability distribution

Probability distribution

Quantitative variables

Random variables

Re-centering

Rescale

Subtraction Rule

CHAPTER **5**

Normal Distribution

Chapter Outline

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.
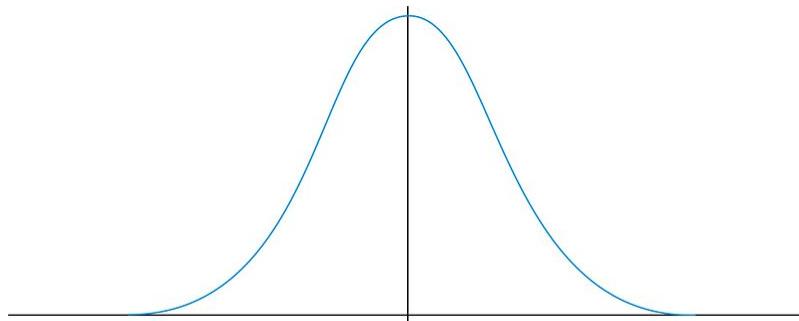
All of these examples show a typical pattern that seems to be a part of many real-life phenomena. In statistics, because this pattern is so pervasive, it seems to fit to call it normal, or more formally, the normal distribution. The normal distribution is an extremely important concept, because it occurs 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 the as a normal curve, or bell curve.
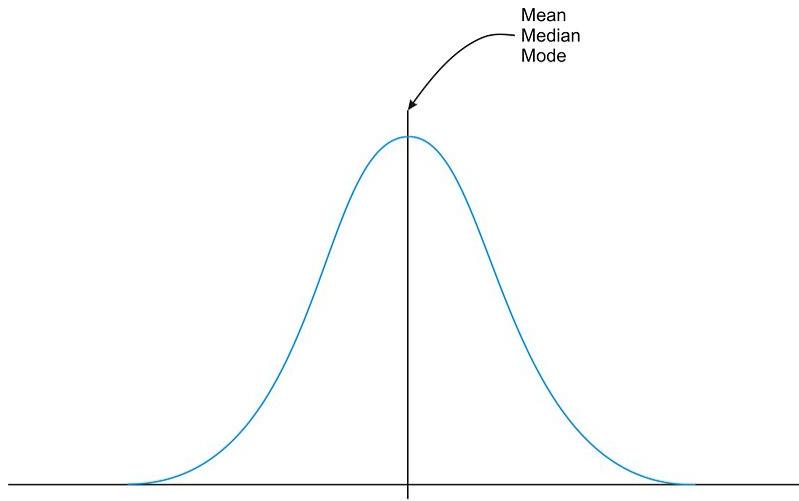
The Normal Distribution

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.
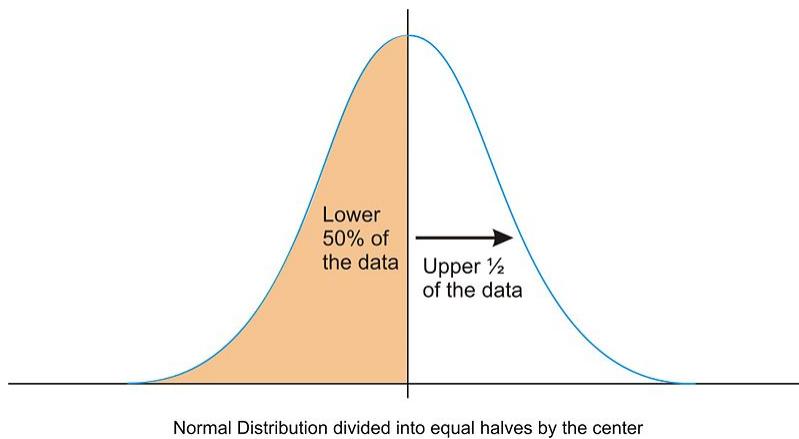
Center

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.

Mean
Median
Mode
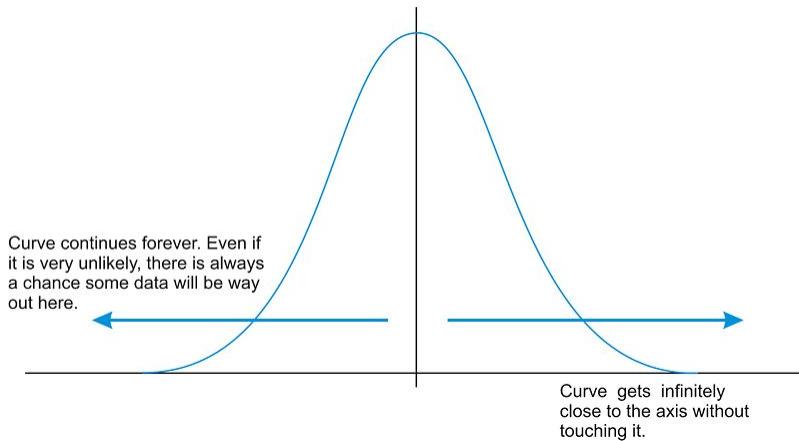
Normal Distribution Center

It is also important to realize that this center peak divides the data into two equal parts.

Normal Distribution divided into equal halves by the center
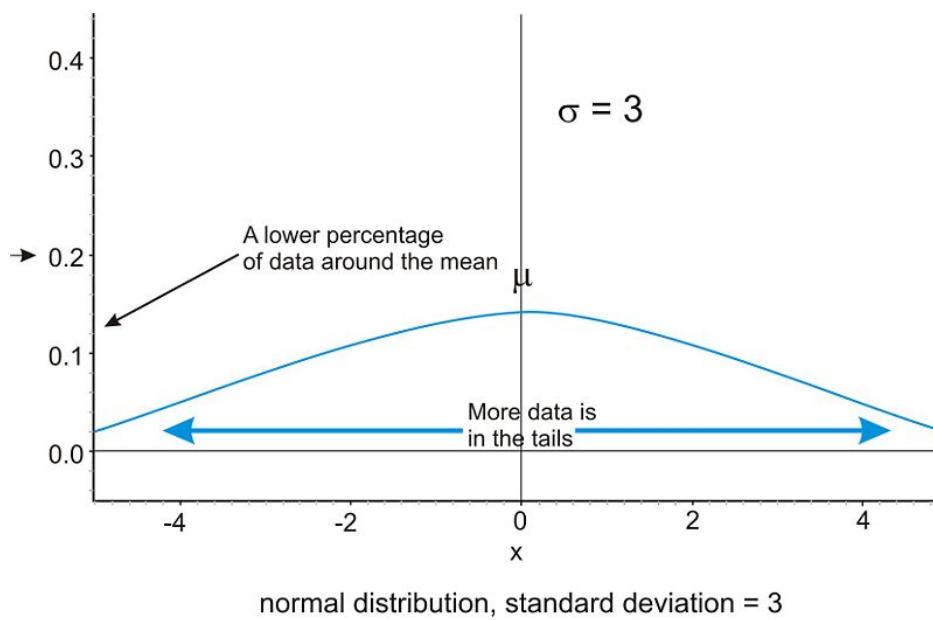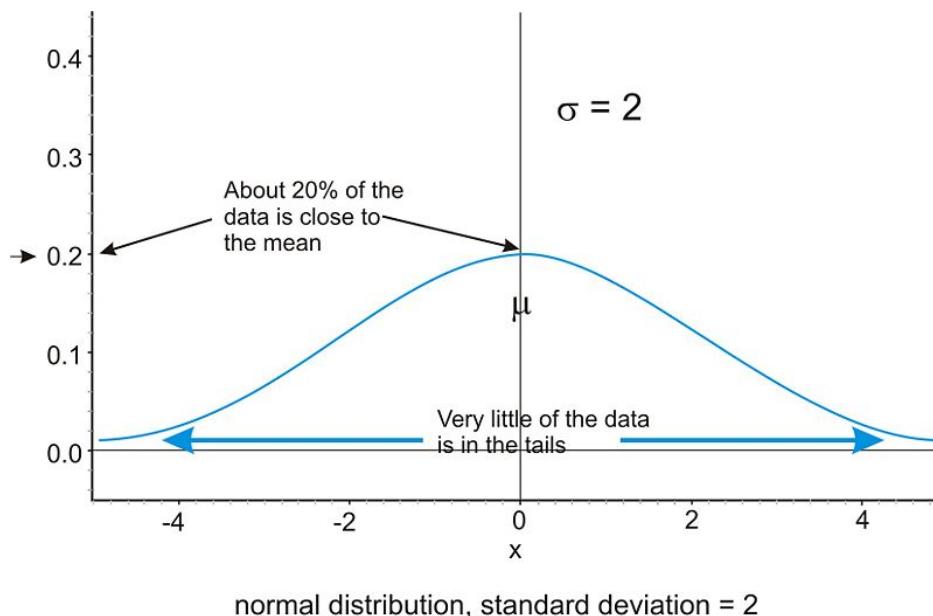
Spread

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):

normal distribution, standard deviation = 2

normal distribution, standard deviation = 3
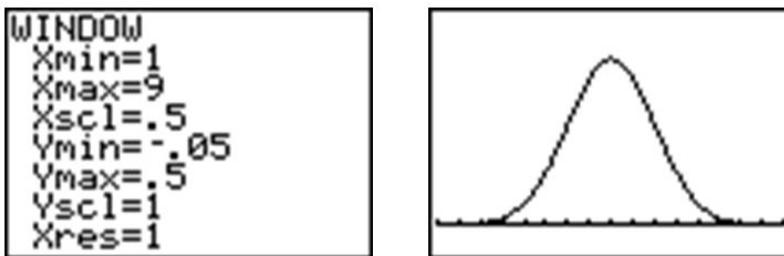
The first distribution pictured above has a smaller standard deviation, and so more of the data are heavily concentrated around the mean than in the second distribution. Also, in the first distribution, there are fewer data values at the extremes than in the second distribution. Because the second distribution 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
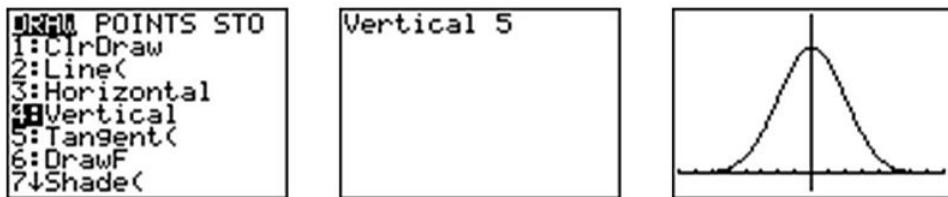
We can graph a normal curve for a probability distribution on the TI-83/84 calculator. To do so, first press **[Y=]**. To create a normal distribution, we will draw an idealized curve using something called a density function. The command is called 'normalpdf(', and it is found by pressing **[2nd][DISTR][1]**. Enter an X 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.

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]**.
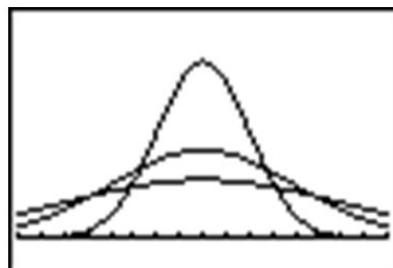
Remember that even though the graph appears to touch the *x*-axis, it is actually just very close to it.

In your **Y=** Menu, enter the following to graph 3 different normal distributions, each with a different standard deviation:
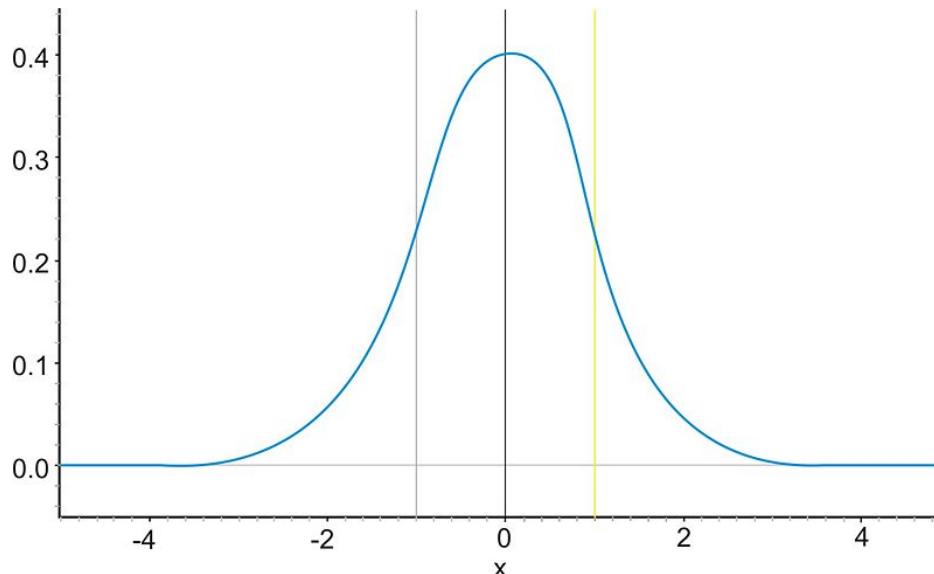
This makes it easy to see the change in spread when the standard deviation changes.
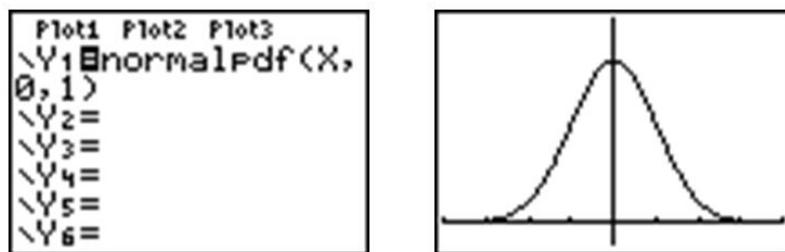
The Empirical Rule

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.
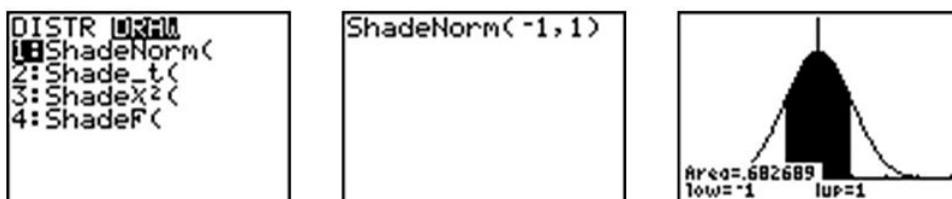
The Standard Normal Distribution
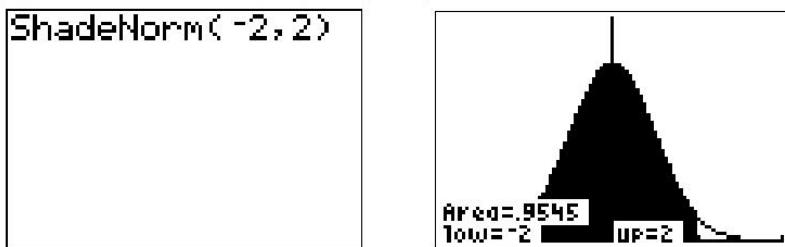
Notice that vertical lines are drawn at points that are exactly one standard deviation to the left and right of the mean. We have consistently described standard deviation as a measure of the typical distance away from the mean. How much of the data is actually within one standard deviation of the mean? 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.
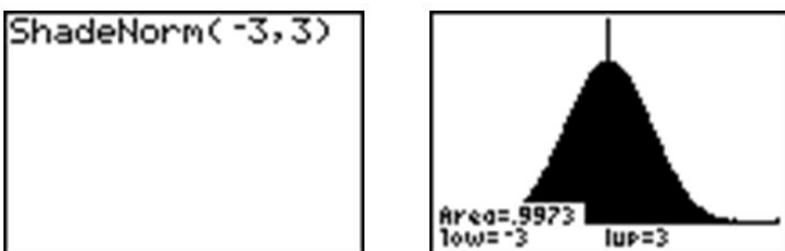
Now press **[2ND][DISTR]**, go to the **DRAW** menu, and choose 'ShadeNorm('. Insert '−1, 1' after the 'Shade-Norm(' command and press **[ENTER]**. It will shade the area within one standard deviation of the mean.

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.
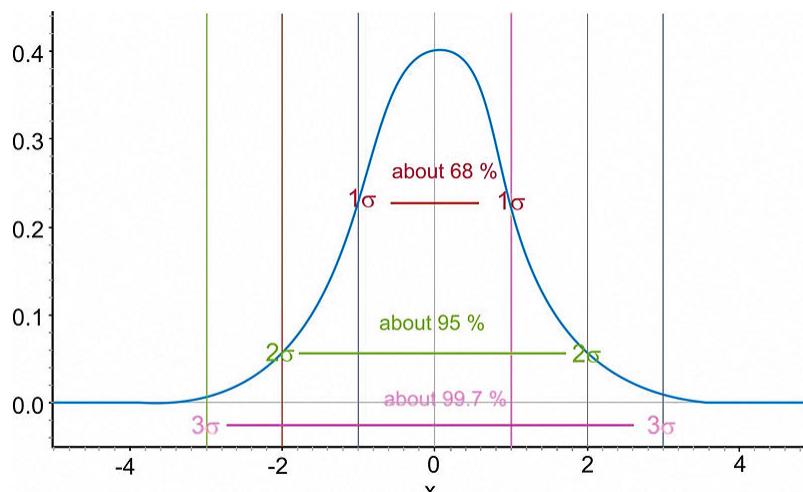
Notice that the calculator actually makes it look like the entire distribution is shaded because of the limitations of the screen resolution, but as we have already discovered, there is still some area under the curve further out than that. These three approximate percentages, 68%, 95%, and 99.7%, are 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.

The Empirical Rule

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:

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{93 - 82}{7}$$
$$z = \frac{11}{7} \approx 1.57$$

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?

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{81 - 65}{10}$$
$$z = \frac{16}{10}$$
$$z = 1.6$$
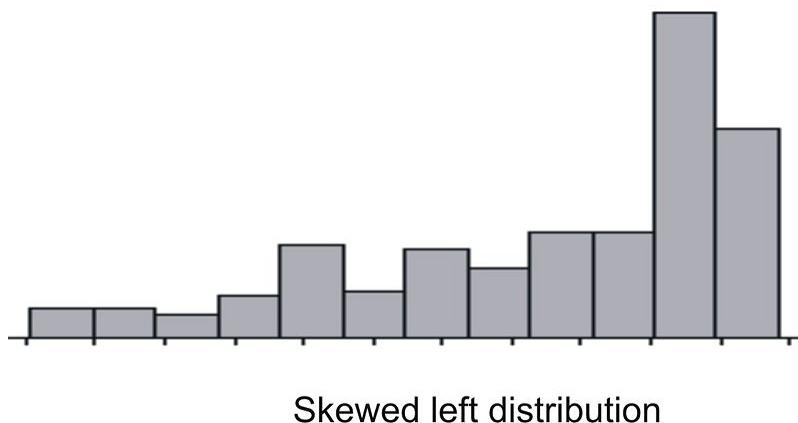
Example: On a college entrance exam, the mean was 70, and the standard deviation was 8. If Helen's *z*-score was -1.5, what was her exam score?
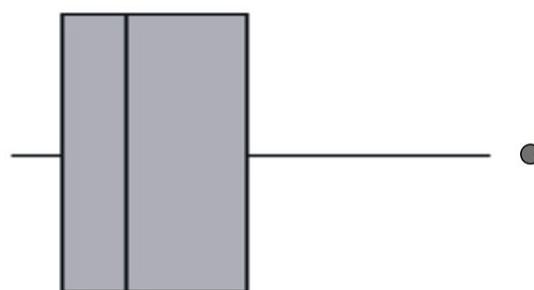
$$z = \frac{x - \mu}{\sigma}$$
$$\therefore z \cdot \sigma = x - \mu$$
$$x = \mu + z \cdot \sigma$$
$$x = 70 + (-1.5)(8)$$
$$x = 58$$

Assessing Normality
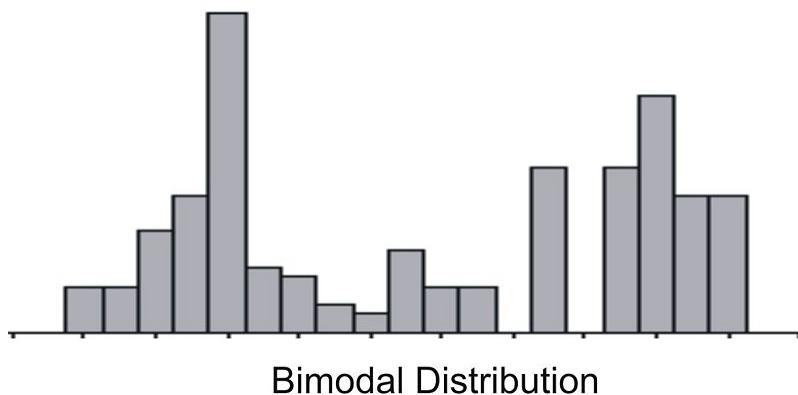
The best way to determine if a data set approximates a normal distribution is to look at a visual representation. Histograms and box plots can be useful indicators of normality, but they are not always definitive. It is often easier to tell if a data set is *not* normal from these plots.

Skewed left distribution

Skewed right distribution with outliers

Bimodal Distribution
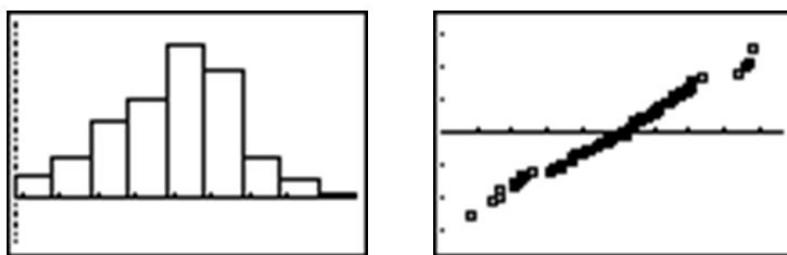
If a data set is skewed right, it means that the right tail is significantly longer than the left. Similarly, skewed left means the left tail has more weight than the right. A bimodal distribution, on the other hand, has two modes, or peaks. For instance, with a histogram of the heights of American 30-year-old adults, you will see a bimodal distribution−one mode for males and one mode for females.

There is a plot we can use to determine if a distribution is normal called a *normal probability plot* or *normal quantile plot*. To make this plot by hand, first order your data from smallest to largest. Then, determine the quantile of each data point. Finally, using a table of standard normal probabilities, determine the closest z-score for each quantile. Plot these z-scores against the actual data values. To make a normal probability plot using your calculator, enter your data into a list, then use the last type of graph in the **STAT PLOT** menu, as shown below:

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 of the quantiles of the data are plotted against the actual data values, the normal probability plot appears strongly linear, indicating that the data set closely approximates a normal distribution. The following example will allow you to see how a normal probability plot is made in more detail.
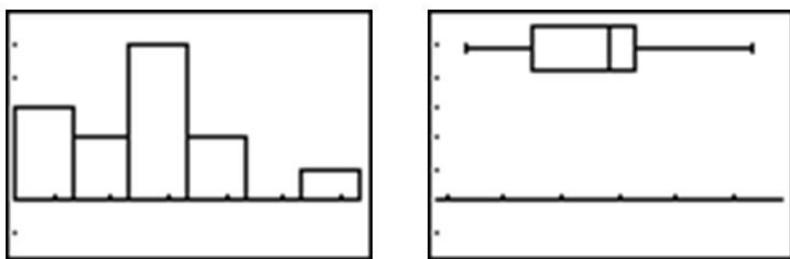
Example: The following data set tracked high school seniors' involvement in traffic accidents. The participants were asked the following question: "During the last 12 months, how many accidents have you had while you were driving (whether or not you were responsible)?"

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:
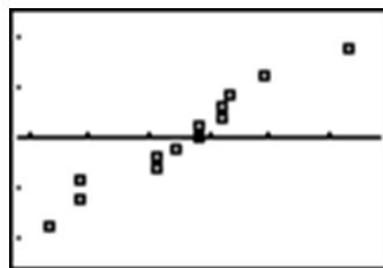
The histogram appears to show a roughly mound-shaped and symmetric distribution. The box plot does not appear to be significantly skewed, but the various sections of the plot also do not appear to be overly symmetric, either. In the following chart, the data has been reordered from smallest to largest, the quantiles have been determined, and the closest corresponding z-scores have been found using a table of standard normal probabilities.

TABLE 5.2:

Year	Percentage	Quantile	z-score
1996	74.1	$\frac{1}{13} = 0.078$	-1.42
1997	74.4	$\frac{2}{13} = 0.154$	-1.02
1998	74.4	$\frac{3}{13} = 0.231$	-0.74
1999	75.1	$\frac{4}{13} = 0.286$	-0.56
2000	75.1	$\frac{5}{13} = 0.385$	-0.29
1995	75.3	$\frac{6}{13} = 0.462$	-0.09
2001	75.5	$\frac{7}{13} = 0.538$	0.1
2002	75.5	$\frac{8}{13} = 0.615$	0.29
1991	75.7	$\frac{9}{13} = 0.692$	0.50
1994	75.7	$\frac{10}{13} = 0.769$	0.74
2003	75.8	$\frac{11}{13} = 0.846$	1.02
1993	76.1	$\frac{12}{13} = 0.923$	1.43
1992	76.9	$\frac{13}{13} = 1$	3.49

Figure: Table of quantiles and corresponding *z*-scores for senior no-accident data.

Here is a plot of the percentages versus the *z*-scores of their quantiles, or the normal probability plot:

Remember that you can simplify this process by simply entering the percentages into a L_1 in your calculator and selecting the normal probability plot option (the last type of plot) in **STAT PLOT**.

While not perfectly linear, this plot does have a strong linear pattern, and we would, therefore, conclude that the distribution is reasonably normal.

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 normally distributed 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.
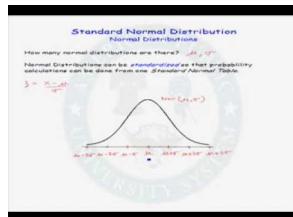
You should learn to recognize the normality of a distribution by examining the shape and symmetry of its visual display. A **normal probability plot**, or **normal quantile plot**, is a useful tool to help check the normality of a distribution. This graph is a plot of the *z*-scores of the data as quantiles against the actual data values. If a distribution is normal, this plot will be linear.

Points to Consider

- How can we use normal distributions to make meaningful conclusions about samples and experiments?
- How do we calculate probabilities and areas under the normal curve that are not covered by the Empirical Rule?
- What are the other types of distributions that can occur in different probability situations?

Multimedia Links

For an explanation of a standardized normal distribution **(4.0)(7.0)**, see APUS07, Standard Normal Distribution (4:22).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1078
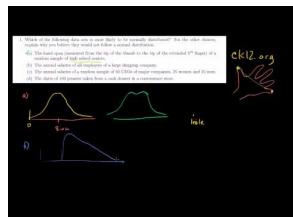
Review Questions

Sample explanations for some of the practice exercises below are available by viewing the following videos. Khan Academy: Normal Distribution Problems (10:52)
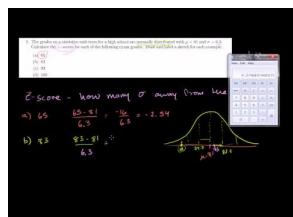
> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1079

Khan Academy: Normal Distribution Problems-z score (7:48)
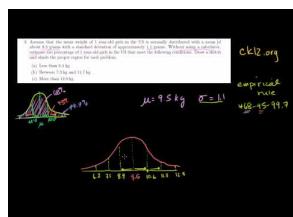
> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1080

Khan Academy: Normal Distribution Problems (Empirical Rule) (10:25)

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1081
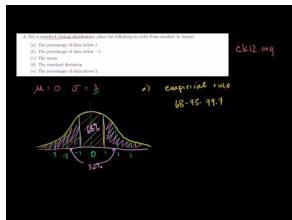
Khan Academy: Standard Normal Distribution and the Empirical Rule (8:15)

MEDIA

Click image to the left or use the URL below.

URL: http://www.ck12.org/flx/render/embeddedobject/1082

Khan Academy: More Empirical Rule and Z-score practice (5:57)

MEDIA

Click image to the left or use the URL below.

URL: http://www.ck12.org/flx/render/embeddedobject/1083

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.

[1]Data available on the College Board Website: http://professionals.collegeboard.com/data-reports-research/ap/archived/2007

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?

7. A statistics class bought some sprinkle (or jimmies) doughnuts for a treat and noticed that the number of sprinkles seemed to vary from doughnut to doughnut, so they counted the sprinkles on each doughnut. Here are the results: 241, 282, 258, 223, 133, 335, 322, 323, 354, 194, 332, 274, 233, 147, 213, 262, 227, and 366. (a) Create a histogram, dot plot, or box plot for this data. Comment on the shape, center and spread of the distribution. (b) Find the mean and standard deviation of the distribution of sprinkles. Complete the following chart by standardizing all the values:

$$\mu = \quad \sigma =$$

TABLE 5.3:

Number of Sprinkles	Quantile	z-score
241		
282		
258		
223		
133		
335		
322		
323		
354		
194		
332		
274		
233		
147		
213		
262		
227		
366		

Figure: A table to be filled in for the sprinkles question.

(c) Create a normal probability plot from your results.

(d) Based on this plot, comment on the normality of the distribution of sprinkle counts on these doughnuts.

References

[1]http://www.albany.edu/sourcebook/pdf/t352.pdf

5.2 The Density Curve of the Normal Distribution

Learning Objectives
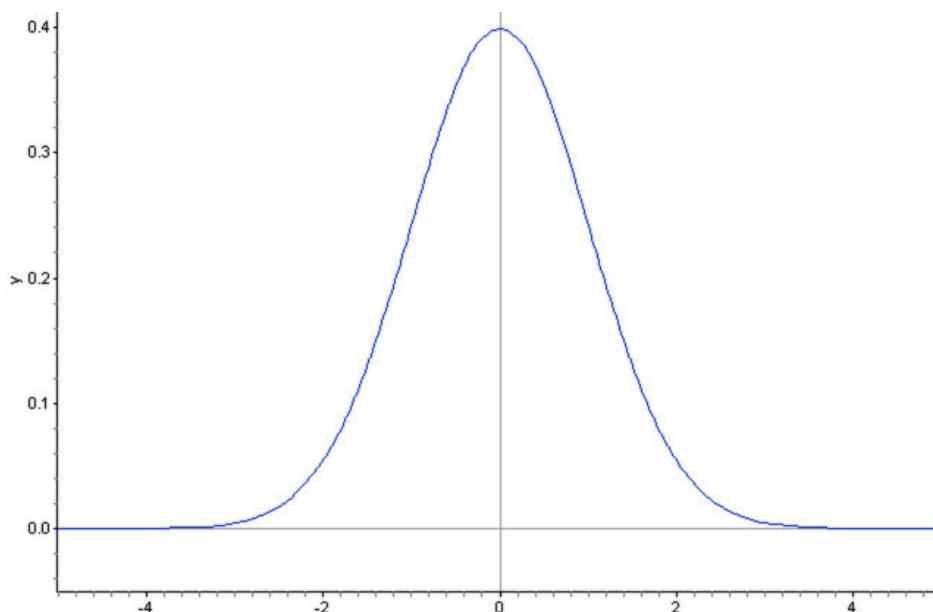
- Identify the properties of a normal density curve and the relationship between concavity and standard deviation.
- Convert between *z*-scores and areas under a normal probability curve.
- Calculate probabilities that correspond to left, right, and middle areas from a *z*-score table.
- Calculate probabilities 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
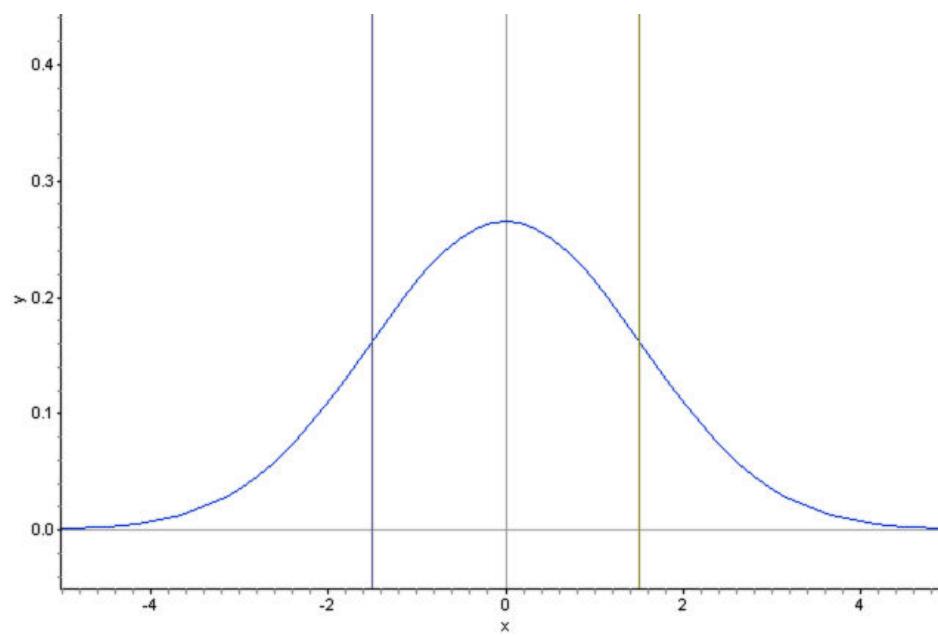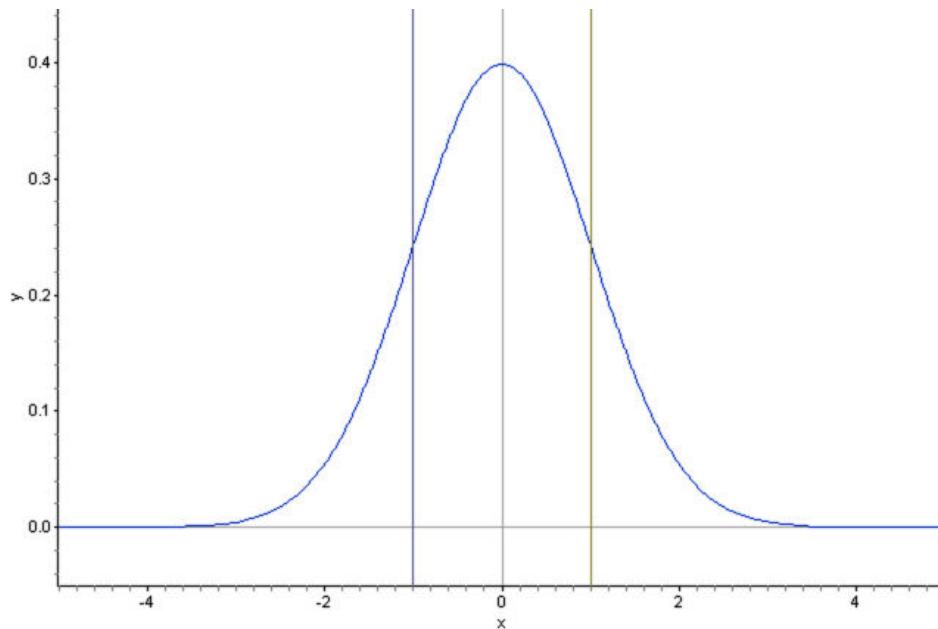
A *density curve* is an idealized representation of a distribution in which the area under the curve is defined to be 1. Density curves need not be normal, but the normal density curve will be the most useful to us.
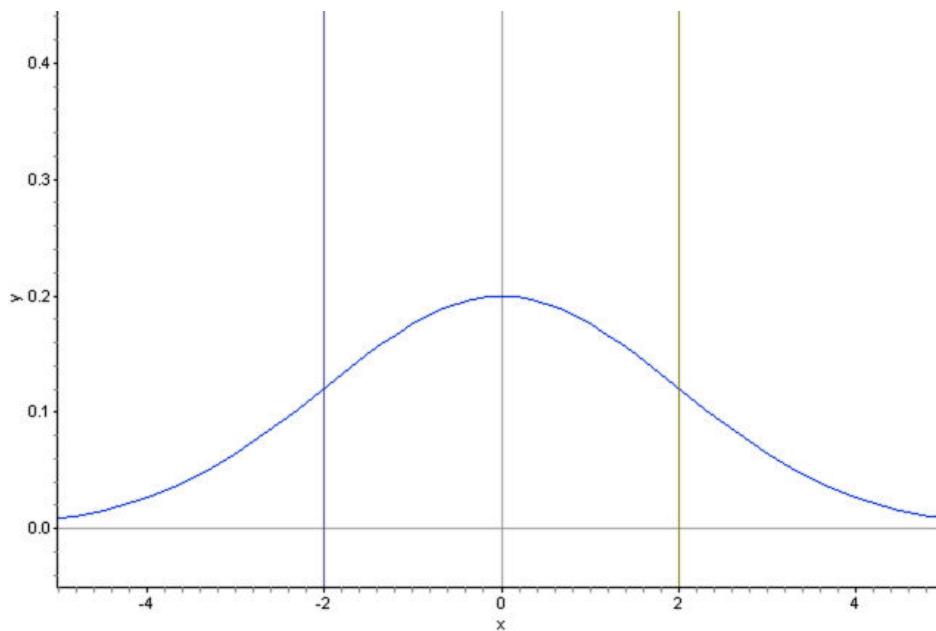
Inflection Points on a Normal Density Curve
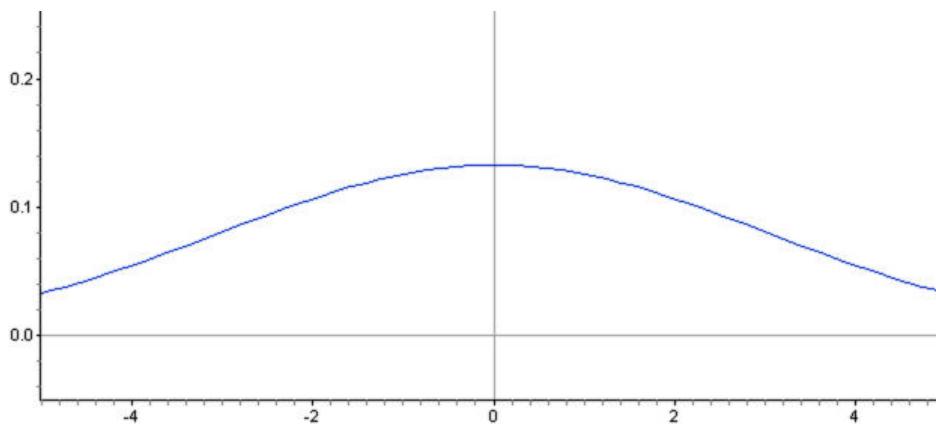
We already know from the Empirical Rule that approximately $\frac{2}{3}$ of the data in a normal distribution lies within 1 standard deviation of the mean. With a normal density curve, this means that about 68% of the total area under the

curve is within *z*-scores of ±1. Look at the following three density curves:
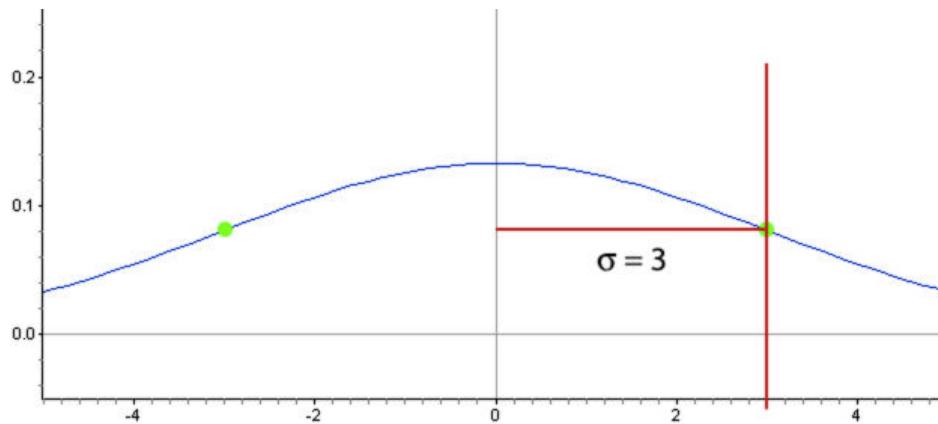
Notice that the curves are spread increasingly wider. Lines have been drawn to show the points that are one standard deviation on either side of the mean. Look at where this happens on each density curve. Here is a normal distribution with an even larger standard deviation.
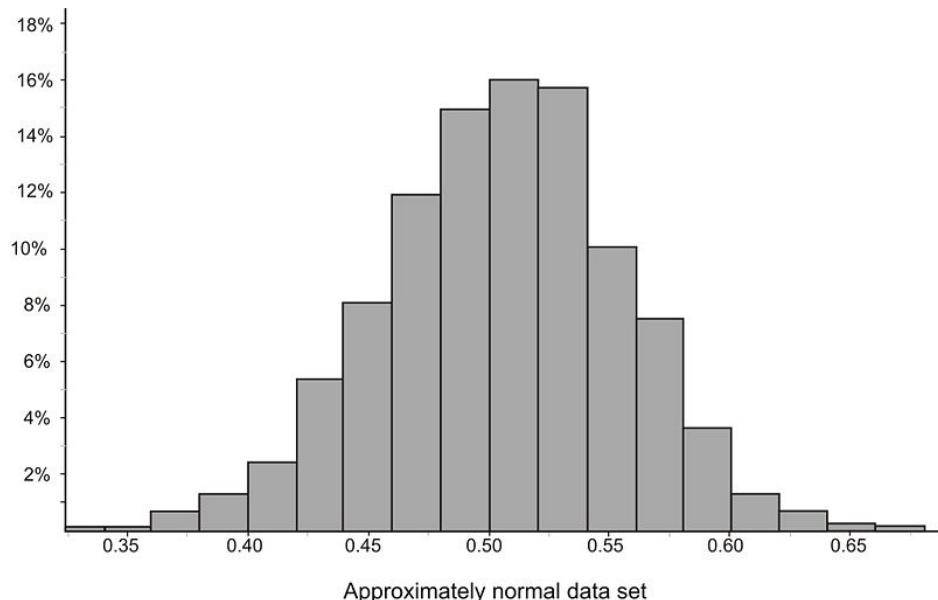
Is it possible to predict the standard deviation of this distribution by estimating the *x*-coordinate of a point on the density curve? Read on to find out!

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.
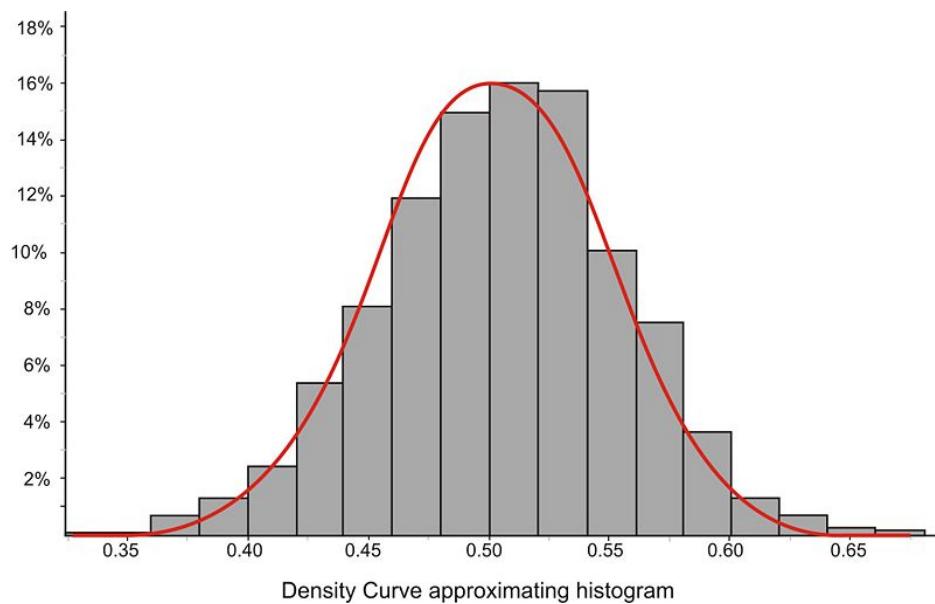
In this example, the standard deviation is 3 units. We can use this concept to estimate the standard deviation of a normally distributed data set.
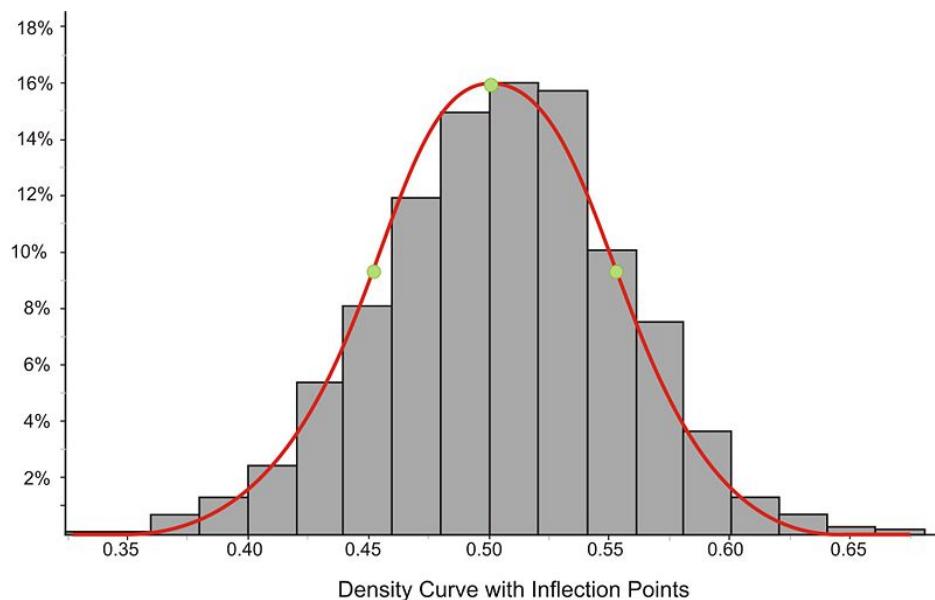
Example: Estimate the standard deviation of the distribution represented by the following histogram.

Approximately normal data set

This distribution is fairly normal, so we could draw a density curve to approximate it as follows:

Density Curve approximating histogram

Now estimate the inflection points as shown below:

Density Curve with Inflection Points
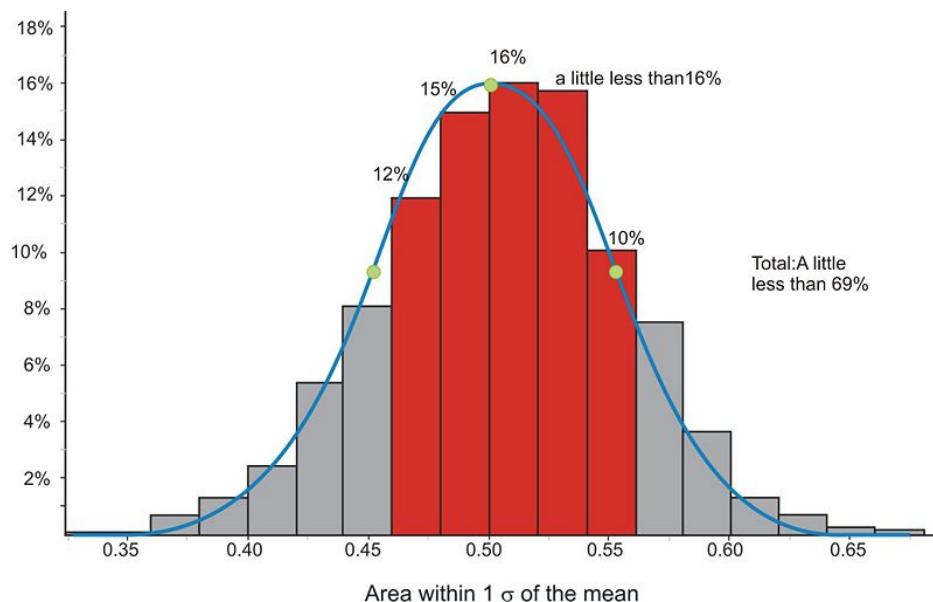
It appears that the mean is about 0.5 and that the *x*-coordinates of the inflection points are about 0.45 and 0.55, respectively. This would lead to an estimate of about 0.05 for the standard deviation.

The actual statistics for this distribution are as follows:

$$s \approx 0.04988$$
$$\bar{x} \approx 0.4997$$

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.

Area within 1 σ 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.

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?

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{81 - 65}{10}$$
$$z = \frac{16}{10}$$
$$z = 1.6$$

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?

$$z = \frac{x - \mu}{\sigma}$$
$$\therefore z \cdot \sigma = x - \mu$$
$$x = \mu + z \cdot \sigma$$
$$x = (70) + (-1.5)(8)$$
$$x = 58$$
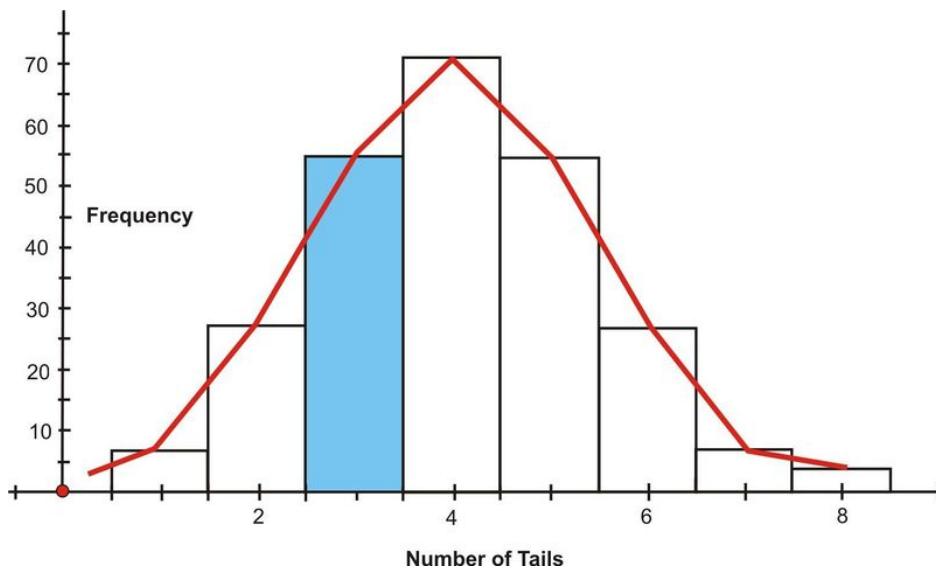
Now you will see how *z*-scores are used to determine the probability of an event.

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:

$$P(3 \text{ tails}) = \frac{{}_8C_3}{2^8}$$
$$P(3 \text{ tails}) = \frac{56}{256}$$
$$P(3 \text{ tails}) \cong 0.2188$$
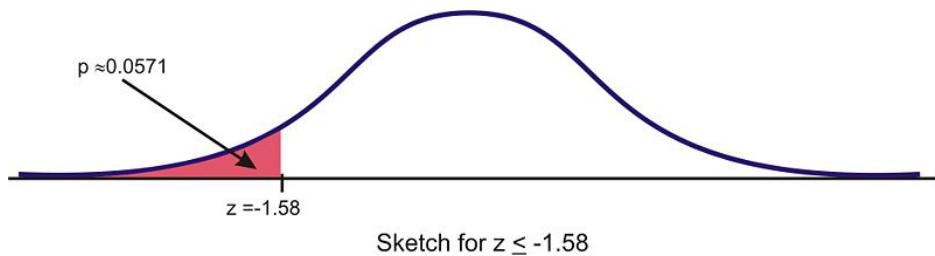
Geometrically, this probability represents the area of the blue shaded bar divided by the total area of the bars. The area of the blue shaded bar is approximately equal to the area under the normal curve from 2.5 to 3.5.

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/2ce9o gv .
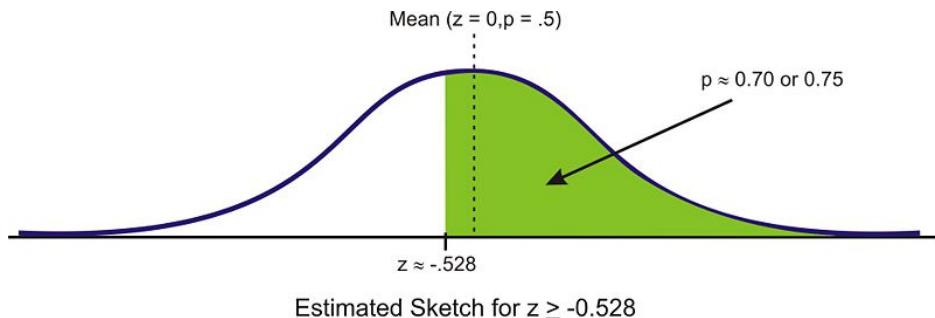
The values inside the given table represent the areas under the standard normal curve for values between 0 and the relative *z*-score. For example, to determine the area under the curve between *z*-scores of 0 and 2.36, look in the intersecting cell for the row labeled 2.3 and the column labeled 0.06. 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.)
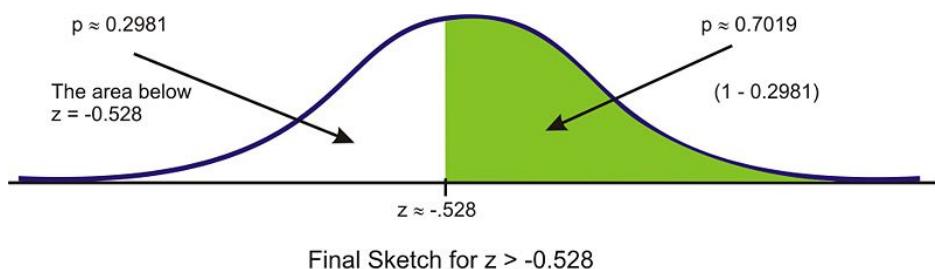
It is extremely important, especially when you first start with these calculations, that you get in the habit of relating it to the normal distribution by drawing a sketch of the situation. In this case, simply draw a sketch of a standard normal curve with the appropriate region shaded and labeled.

Sketch for z ≤ -1.58

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.

Estimated Sketch for z ≥ -0.528

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.

Final Sketch for z ≥ -0.528

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
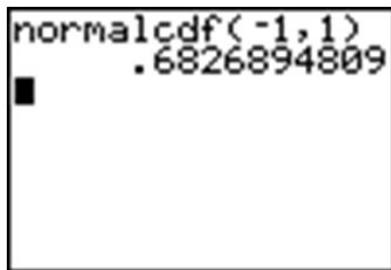
Your graphing calculator has already been programmed to calculate probabilities for a normal density curve using what is called a *cumulative density function*. The command you will use is found in the **DISTR** menu, which you can bring up by pressing **[2ND][DISTR]**.

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 $\sigma = 1$.
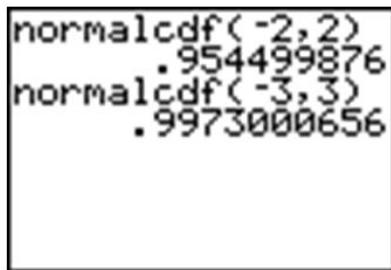
For example, entering 'normalcdf(−1, 1)' will specify the area within one standard deviation of the mean, which we already know to be approximately 0.68.

Try verifying the other values from the Empirical Rule.

Summary:

'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 $\sigma = 1$.

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.

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:
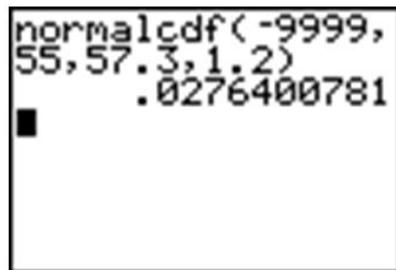
$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{55 - 57.3}{1.2}$$
$$z \approx -1.911666\ldots$$

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):

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.

Points to Consider

- How do we calculate areas/probabilities for distributions that are not normal?
- How do we calculate z-scores, means, standard deviations, or actual values given a probability or area?

On the Web

Tables

http://tinyurl.com/2ce9ogv This link leads you to a $z-$table and an explanation of how to use it.
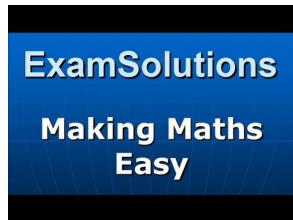
http://tinyurl.com/2aau5zy Investigate the mean and standard deviation of a normal distribution.

http://tinyurl.com/299hsjo The Normal Calculator.

http://www.math.unb.ca/~knight/utility/NormTble.htm Another online normal probability table.

Multimedia Links

For an example showing how to compute probabilities with normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: P(more than x) where x is less than the mean (8:40).

ExamSolutions

Making Maths
Easy

MEDIA

Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1084
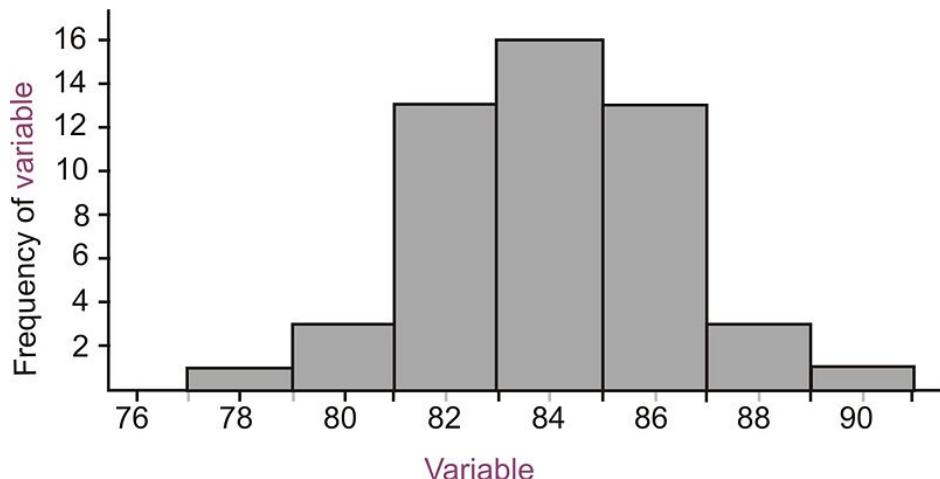
Review Questions

1. Estimate the standard deviation of the following distribution.

2. Calculate the following probabilities using only the z-table. Show all your work.

   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:

```
normalcdf(80,90)
                  0
```

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.
5. Teachers A and B have final exam scores that are approximately normally distributed, with the mean for Teacher A equal to 72 and the mean for Teacher B equal to 82. The standard deviation of Teacher A's scores is 10, and the standard deviation of Teacher B's scores is 5.

a. With which teacher is a score of 90 more impressive? Support your answer with appropriate probability calculations and with a sketch.
b. With which teacher is a score of 60 more discouraging? Again, support your answer with appropriate probability calculations and with a sketch.

5.3 Applications of the Normal Distribution

Learning Objective

- Apply the characteristics of a normal distribution to solving problems.

Introduction
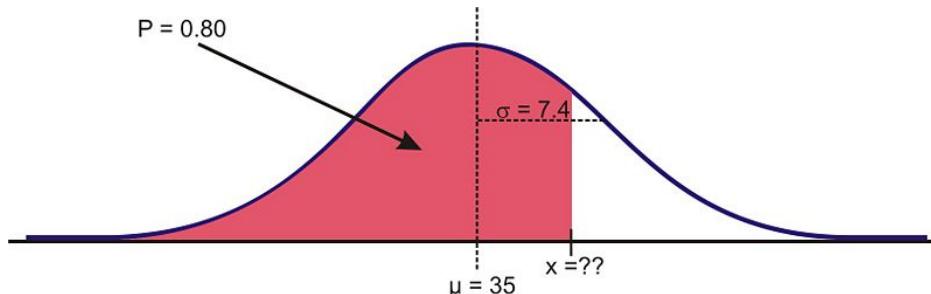
The normal distribution is the foundation for statistical inference and will be an essential part of many of those topics in later chapters. In the meantime, this section will cover some of the types of questions that can be answered using the properties of a normal distribution. The first examples deal 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.
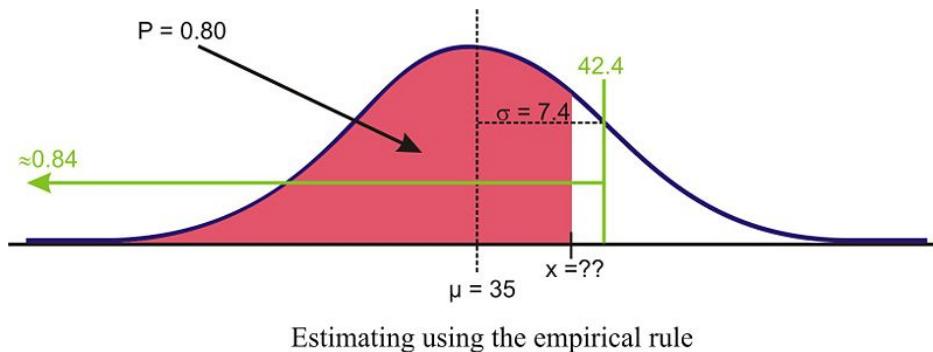
Sketch of distribution

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.

Estimating using the empirical rule
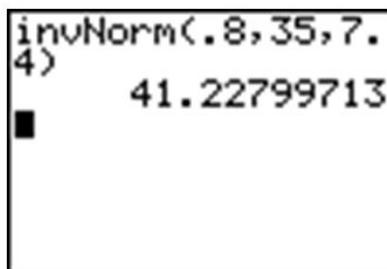
When we were given a value of the variable and were asked to find the percentage or probability, we used a *z*-table or the 'normalcdf(' command on a graphing calculator. But how do we find a value given the percentage? Again, the table has its limitations in this case, and graphing calculators and computer 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)'
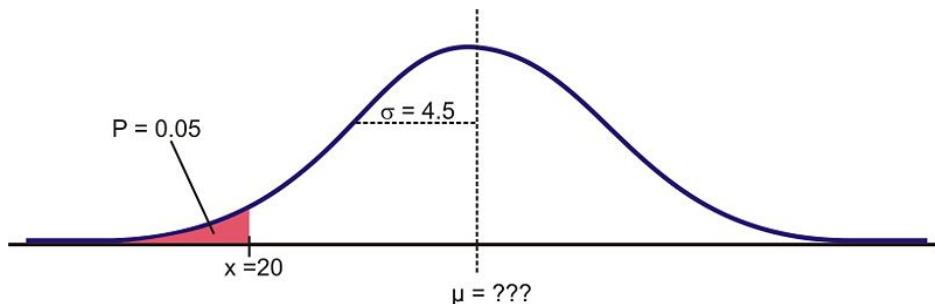
Make sure to enter the values in the correct order, such as in the example below:

Unknown Mean or Standard Deviation
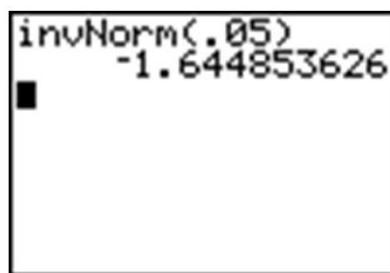
Example: For a normally distributed random variable, $\sigma = 4.5$, $x = 20$, and $p = 0.05$, Estimate μ.

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.
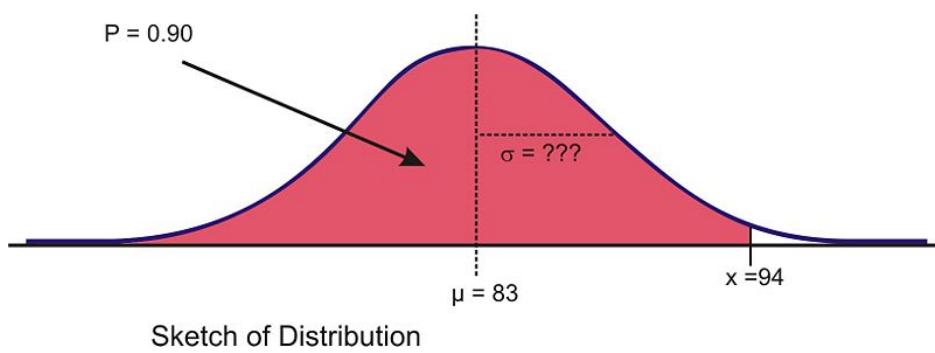
Now, plug in the known quantities into the z-score formula and solve for μ as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$-1.645 \approx \frac{20 - \mu}{4.5}$$
$$(-1.645)(4.5) \approx 20 - \mu$$
$$-7.402 - 20 \approx -\mu$$
$$-27.402 \approx -\mu$$
$$\mu \approx 27.402$$
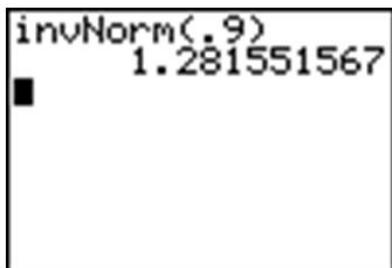
Example: For a normally-distributed random variable, $\mu = 83$, $x = 94$, and $p = 0.90$. Find σ.

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 $\sigma = 1$.

```
invNorm(.9)
            1.281551567
■
```
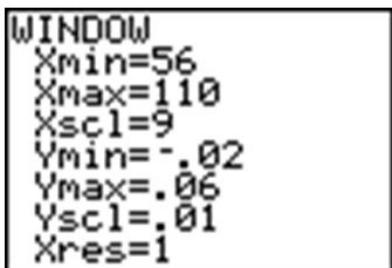
Now, use the *z*-score formula and solve for σ as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$1.282 \approx \frac{94 - 83}{\sigma}$$
$$\sigma \approx \frac{11}{1.282}$$
$$\sigma \approx 8.583$$

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.

```
WINDOW
 Xmin=56
 Xmax=110
 Xscl=9
 Ymin=-.02
 Ymax=.06
 Yscl=.01
 Xres=1
```

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.
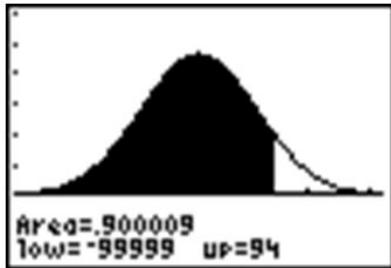
Choose the 'ShadeNorm(' command. With this command, you enter the values just as if you were doing a 'normalcdf(' calculation. The syntax for the 'ShadeNorm(' command is as follows:

'ShadeNorm(lower bound, upper bound, mean, standard deviation)'

Enter the values shown in the following screenshot:

```
ShadeNorm( -99999
,94,83,8.583)
```
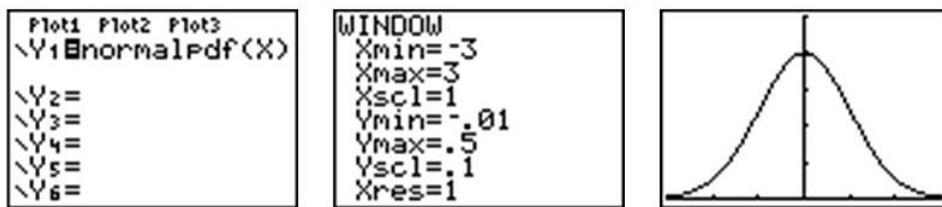
Next, press **[ENTER]** to see the result. It should appear as follows:

```
Area=.900009
low= -99999   up=94
```

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.

```
Plot1 Plot2 Plot3
\Y1▊normalpdf(X)
\Y2=
\Y3=
\Y4=
\Y5=
\Y6=
```
```
WINDOW
 Xmin=-3
 Xmax=3
 Xscl=1
 Ymin=-.01
 Ymax=.5
 Yscl=.1
 Xres=1
```

Normal Distributions with Real Data
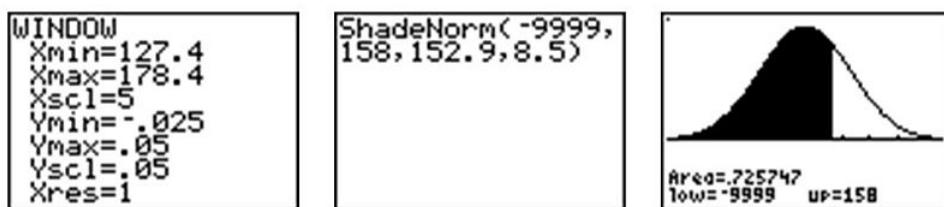
The foundation of performing experiments by collecting surveys and samples is most often based on the normal distribution, as you will learn in greater detail in later chapters. Here are two examples to get you started.

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 children[1]. 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?
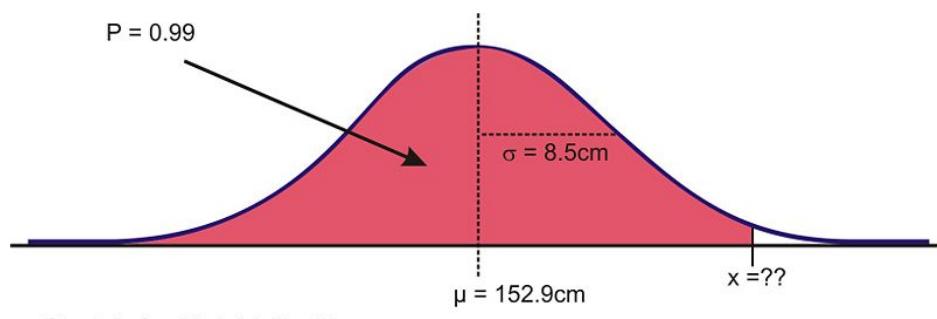
We first must assume that the height of 12-year-old boys in Britain is normally distributed, and this seems like a reasonable assumption to make. As always, draw a sketch and estimate a reasonable answer prior to calculating the percentage. In this case, let's use the calculator to sketch the distribution 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.
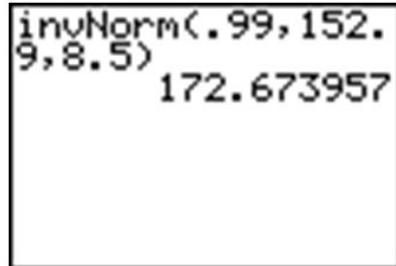
How tall would Cecil need to be in order to be in the top 1% of all 12-year-old boys in Britain?

Here is a sketch:

Sketch for Height Problem
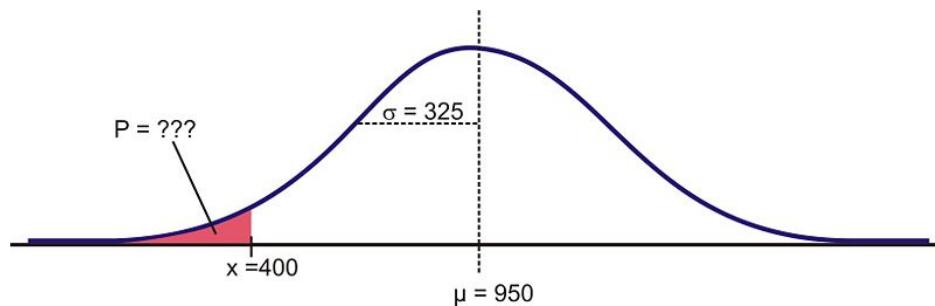
In this case, we are given the percentage, so we need to use the 'invNorm(' command as shown.

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:

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
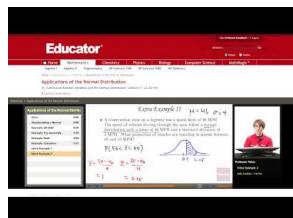
In order to find the percentage of data in-between two values (or the probability of a randomly chosen value being between those values) in a normal distribution, we can use the 'normalcdf(' command on the TI-83/84 calculator. When you know the percentage or probability, use the 'invNorm(' command to 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.

Point to Consider

- How do the probabilities of a standard normal curve apply to making decisions about unknown parameters for a population given a sample?

Multimedia Links

For an example of finding the probability between values in a normal distribution **(4.0)(7.0)**, see EducatorVids, Statistics: Applications of the Normal Distribution (1:45).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1085
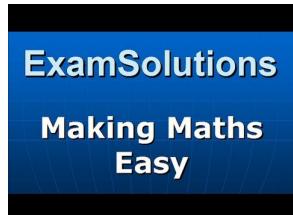
For an example showing how to find the mean and standard deviation of a normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: Finding the Mean and Standard Deviation (6:01).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1086
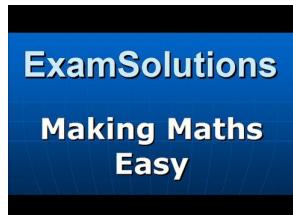
For the continuation of finding the mean and standard deviation of a normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: Finding the Mean and Standard Deviation (Part 2) (8:09).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1087

Review Questions

1. Which of the following intervals contains the middle 95% of the data in a standard normal distribution?

    a. $z < 2$
    b. $z \leq 1.645$
    c. $z \leq 1.96$
    d. $-1.645 \leq z \leq 1.645$
    e. $-1.96 \leq z \leq 1.96$

2. 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. The manufacturing process at a metal-parts factory produces some slight variation in the diameter of metal ball bearings. The quality control experts claim that the bearings produced have a mean diameter of 1.4 cm. If the diameter is more than 0.0035 cm too wide or too narrow, they will not work properly. In order to maintain its reliable reputation, the company wishes to insure that no more than one-tenth of 1% of the bearings that are defective. What would the standard deviation of the manufactured bearings need to be in order to meet this goal?
5. 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?
   d. If the manufacturer wants to adjust the production process so that no more than 1 candy bar in 1000 weighs less than the advertised weight, what would the mean of the actual weights need to be? (Assume the standard deviation remains the same.)
   e. If the manufacturer wants to adjust the production process so that the mean remains at 2.2 ounces and no more than 1 candy bar in 1000 weighs less than the advertised weight, how small does the standard deviation of the weights need to be?

References

http://www.ic.nhs.uk/default.asp?sID=1198755531686

http://www.nytimes.com/2008/04/04/us/04poll.html

On the Web

http://davidmlane.com/hyperstat/A25726.html Contains the formula for the normal probability density function.

http://www.willamette.edu/~mjaneba/help/normalcurve.html Contains background on the normal distribution, including a picture of Carl Friedrich Gauss, a German mathematician who first used the function.

http://en.wikipedia.org/wiki/Normal_distribution Is highly mathematical.

Keywords

Concave down

Concave up

Cumulative density function

Density curve

Empirical Rule

Inflection Points

Normal distribution

Normal probability plot

Normal quantile plot

Probability density function

Standard normal curve

Standard normal distribution

Standardize

z-score

CHAPTER **6**

Planning and Conducting an Experiment or Study

Chapter Outline

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.

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 appear[1]:

"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 hand[3]. 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/non members/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 poles, 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 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, on the other hand, would probably answer yes 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 blood[4]. 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

Randomization

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.
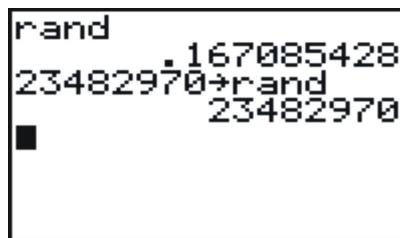
A Note about Randomness

Technology Note: Generating Random Numbers on the TI-83/84 Calculator
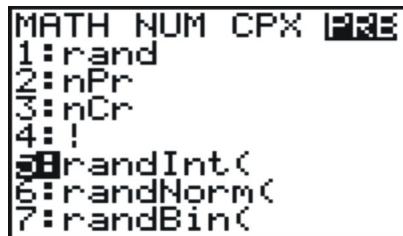
Your graphing calculator has a random number generator. Press **[MATH]** and move over to the **PRB** menu, which stands for probability. (Note: Instead of pressing the right arrow three times, you can just use the left arrow once!) Choose '1:rand' for the random number generator and press **[ENTER]** 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.
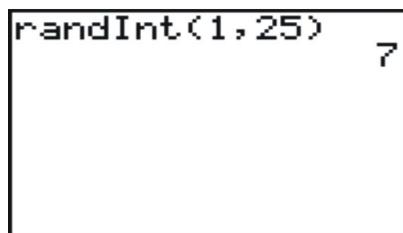
Now, back to our example. If we want to choose a student at random between 1 and 25, we need to generate a random integer between 1 and 25. To do this, press **[MATH][PRB]** and choose the 'randInt(' function.

```
MATH NUM CPX PRB
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(
```
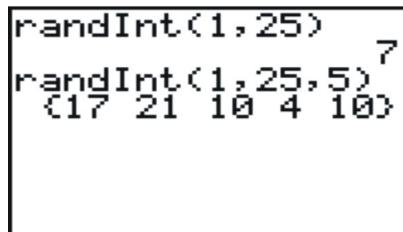
The syntax for this command is as follows:

'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
```
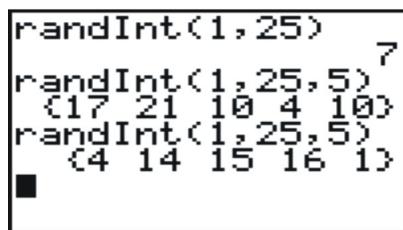
In this example, the student chosen would be student number 7. If we wanted to choose 5 students at random, we could enter the command shown below:

```
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 1}
■
```
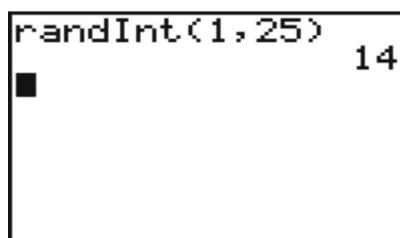
In this example, student number 4 has also already been chosen, so we would select student number 14 as our fifth student.

On the Web

http://tinyurl.com/395cue3 You choose the population size and the sample size and watch the random sample appear.

Systematic Sampling

There are other types of samples that are not simple random samples, 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.

On the Web

http://tinyurl.com/2wnhmok This statistical applet demonstrates five basic probability sampling techniques for a population of size 1000 that comprises two sub-populations separated by a river.

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.

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. What two types of sampling are evident in this example?

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.

10. There are 35 students taking statistics in your school, and you want to choose 10 of them for a survey about their impressions of the course. Use your calculator to select a SRS of 10 students. (Seed your random number generator with the number 10 before starting.) Assuming the students are assigned numbers from 1 to 35, which students are chosen for the sample?

References

http://www.nytimes.com/2008/04/04/us/04pollbox.html

http://www.gao.gov/cgi-bin/getrpt?GAO-04-37

http://edition.cnn.com/2011/TECH/innovation/02/04/census.digital.technology/index.html

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 Britain[1] 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 Health[2] 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
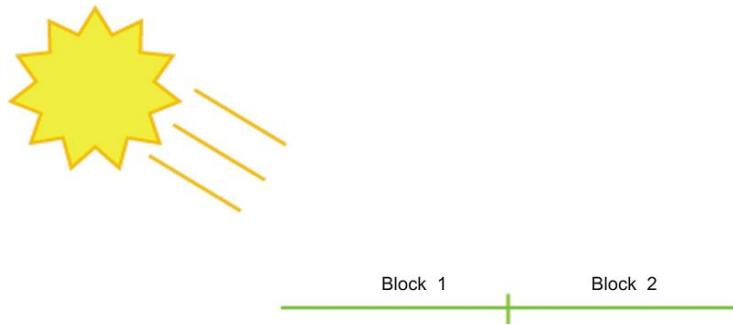
In medical studies, the treatment group usually receives some experimental medication or treatment that has the potential to offer a new cure or improvement for some medical condition. This would mean that the control group would not receive the treatment or medication. Many studies and experiments have 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.
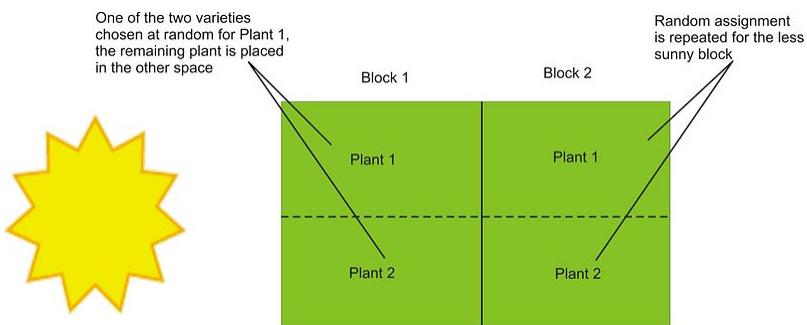
Blocking

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.
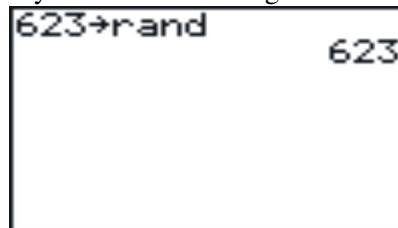
    a. How would you characterize the design of this study?

    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. Seed the random number generator on your calculator using 623.

```
623→rand
                      623
```

Use your calculator to randomly assign each student to a treatment. Explain how you made your assignments.

    c. Identify any potential lurking variables in this experiment.

    d. Explain how you could redesign this experiment as a repeated measures design?

4. A recent British study was attempting to show that a high-fat diet was effective in treating epilepsy in children. According to the *New York Times*, this involved, "...145 children ages 2 to 16 who had never tried the diet, who were having at least seven seizures a week and who had failed to respond to at least two anticonvulsant drugs."

    a. What is the population in this example?

    b. One group began the diet immediately; another group waited three months to start it. In the first group, 38% of the children experienced a 50% reduction in seizure rates, and in the second group, only 6

           percent saw a similar reduction prior to beginning the diet. What information would you need to be able to conclude that this was a valid experiment?

    c. Identify the treatment and control groups in this experiment.

    d. What conclusion could you make from the reported results of this experiment?

5. Researchers want to know how chemically fertilized and treated grass compares to grass grown using only organic fertilizer. Also, they believe that the height at which the grass is cut will affect the growth of the lawn. To test this, grass will be cut at three different heights: 1 inch, 2 inches, and 4 inches. A lawn area of existing healthy grass will be divided up into plots for the experiment. Assume that the soil, sun, and drainage for the test areas are uniform. Explain how you would implement a randomized block design to test the different effects of fertilizer and grass height. Draw a diagram that shows the plots and the assigned treatments.

Further reading:

http://www.nytimes.com/2008/05/06/health/research/06epil.html?ref=health

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

1. 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=15 6231&title=team-daily-polls What type of bias is the primary focus of this non-scientific, yet humorous, look at polling?

2. 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.

   a. Explain how you would use a randomized design to assign the treatments and carry out the experiment.
   b. 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?

3. 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

(a) Identify the treatments, and explain how the researcher could use blinding to improve the study.

(b) Explain how you would use a completely randomized design to assign the subjects to treatment groups.

(c) 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

Blind experiment

Blocking

Census

Cluster sampling

Confounding variables

Control group

Convenience sampling

Double blind experiment

Experiment

Incorrect response bias

Incorrect sampling frame

Judgement sampling

Lurking variable

Margin of error

Matched pairs design

Multi-stage sampling

Non-response bias

Observational study

Placebo

Placebo effect

Questionnaire bias

Quota sampling

Random sample

Randomization

Randomized block design

Randomly assigned

Repeated measures design

Replication

Response bias

Sample

Sampling error

Sampling frame

Seed value

Simple random sample

Size bias

Stratified sampling

Systematic sampling

Treatment

Undercoverage

Voluntary response bias

CHAPTER **7** **Sampling Distributions and Estimations**

Chapter Outline

7.1 Sampling Distribution
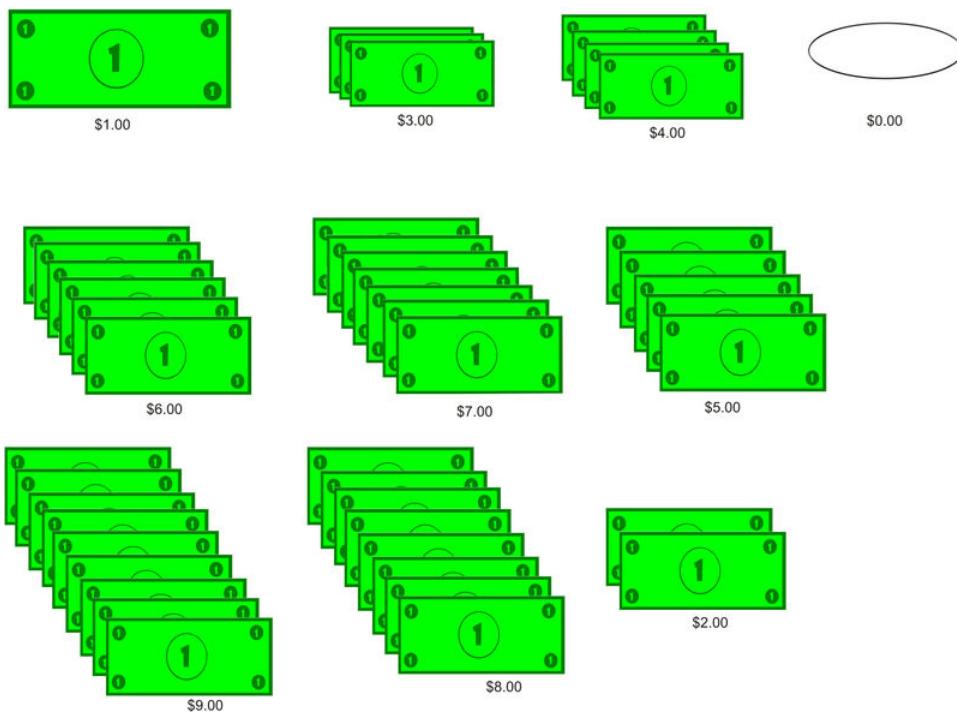
Learning Objectives

- Understand the inferential relationship between a sampling distribution and a population parameter.
- Graph a frequency distribution of sample means using a data set.
- Understand the relationship between sample size and the distribution of sample means.
- Understand sampling error.

Introduction

Have you ever wondered how the mean, or average, amount of money per person in a population is determined? It would be impossible to contact 100% of the population, so there must be a statistical way to estimate the mean number of dollars per person in the population.

Suppose, more simply, that we are interested in the mean number of dollars that are in each of the pockets of ten people on a busy street corner. The diagram below reveals the amount of money that each person in the group of ten has in his/her pocket. We will investigate this scenario later in the lesson.

Sampling Distributions

In previous chapters, you have examined methods that are good for the exploration and description of data. In this section, we will discuss how collecting data by random sampling helps us to draw more rigorous conclusions about the data.
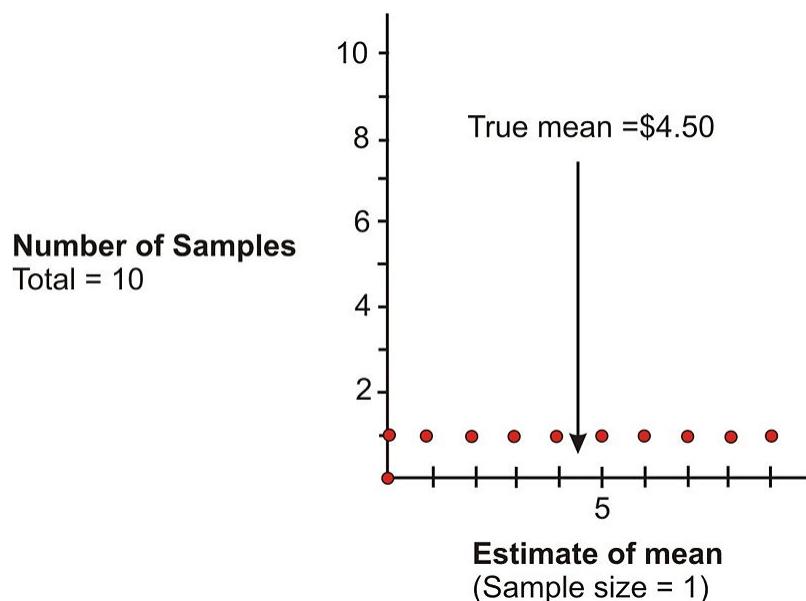
The purpose of sampling is to select a set of units, or elements, from a population that we can use to estimate the parameters of the population. Random sampling is one special type of probability sampling. Random sampling erases the danger of a researcher consciously or unconsciously introducing bias when selecting a sample. In addition, random sampling allows us to use tools from probability theory that provide the basis for estimating the characteristics of the population, as well as for estimating the accuracy of the samples.

Probability theory is the branch of mathematics that provides the tools researchers need to make statistical conclusions about sets of data based on samples. As previously stated, it also helps statisticians estimate the parameters of a population. A *parameter* is a summary description of a given variable in a population. A population mean is an example of a parameter. When researchers generalize from a sample, they're using sample observations to estimate population parameters. Probability theory enables them to both make these estimates and to judge how likely it is that the estimates accurately represent the actual parameters of the population.

Probability theory accomplishes this by way of the concept of *sampling distributions*. A single sample selected from a population will give an estimate of the population parameters. Other samples would give the same, or slightly different, estimates. Probability theory helps us understand how to make estimates of the actual population parameters based on such samples.
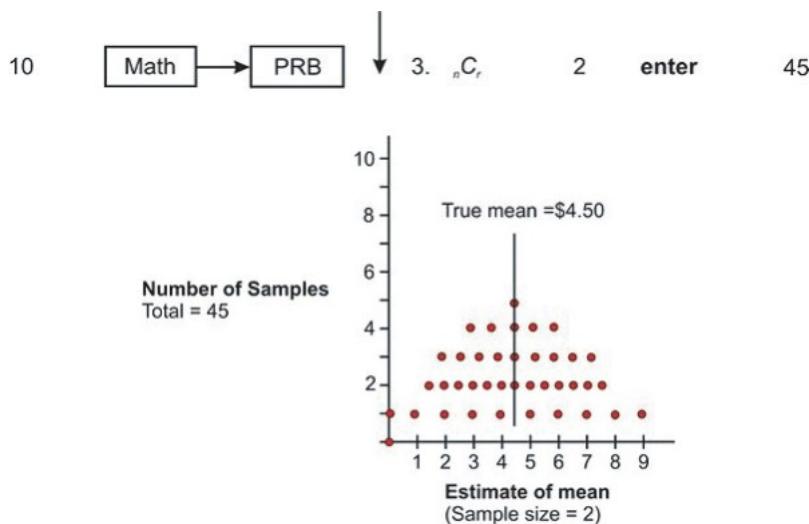
In the scenario that was presented in the introduction to this lesson, the assumption was made that in the case of a population of size ten, one person had no money, another had $1.00, another had $2.00, and so on. Until we reached the person who had $9.00.

The purpose of the task was to determine the average amount of money per person in this population. If you total the money of the ten people, you will find that the sum is $45.00, thus yielding a mean of $4.50. However, suppose you couldn't count the money of all ten people at once. In this case, to complete the task of determining the mean number of dollars per person of this population, it is necessary to select random samples from the population and to use the means of these samples to estimate the mean of the whole population. To start, suppose you were to randomly select a sample of only one person from the ten. The ten possible samples are represented in the diagram in the introduction, which shows the dollar bills possessed by each sample. Since samples of one are being taken, they also represent the means you would get as estimates of the population. The graph below shows the results:
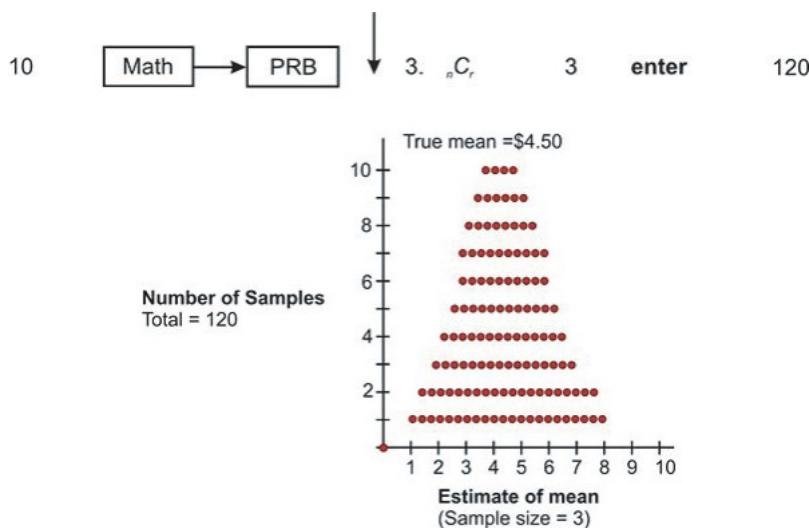
The distribution of the dots on the graph is an example of a sampling distribution. As can be seen, selecting a sample of one is not very good, since the group's mean can be estimated to be anywhere from $0.00 to $9.00, and the true mean of $4.50 could be missed by quite a bit.

What happens if we take samples of two? From a population of 10, in how many ways can two be selected if the order of the two does not matter? The answer, which is 45, can be found by using a graphing calculator as shown in the figure below. When selecting samples of size two from the population, the sampling distribution is as follows:

Increasing the sample size has improved your estimates. There are now 45 possible samples, such as ($0, $1), ($0, $2), ($7, $8), ($8, $9), and so on, and some of these samples produce the same means. For example, ($0, $6), ($1, $5), and ($2, $4) all produce means of $3. The three dots above the mean of 3 represent these three samples. In addition, the 45 means are not evenly distributed, as they were when the sample size was one. Instead, they are more clustered around the true mean of $4.50. ($0, $1) and ($8, $9) are the only two samples whose means deviate by as much as $4.00. Also, five of the samples yield the true estimate of $4.50, and another eight deviate by only plus or minus 50 cents.

If three people are randomly selected from the population of 10 for each sample, there are 120 possible samples, which can be calculated with a graphing calculator as shown below. The sampling distribution in this case is as follows:
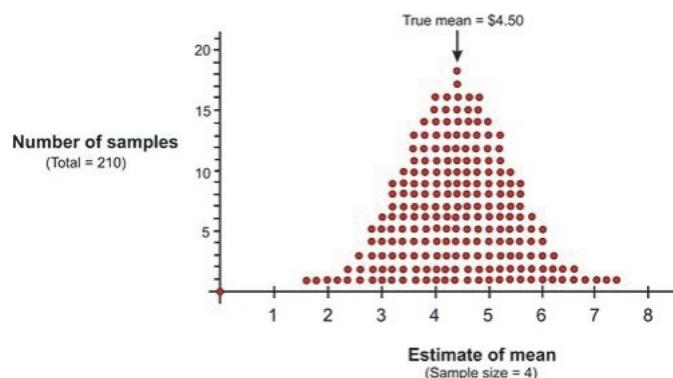
Here are screen shots from a graphing calculator for the results of randomly selecting 1, 2, and 3 people from the

population of 10. The 10, 45, and 120 represent the total number of possible samples that are generated by increasing the sample size by 1 each time.
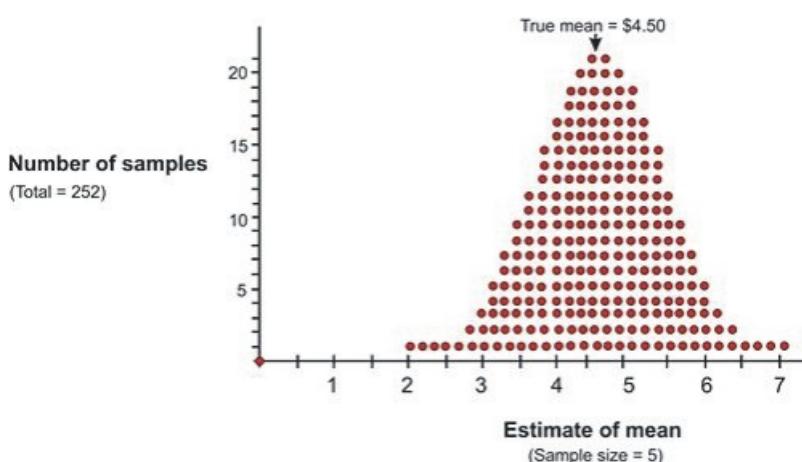
Next, the sampling distributions for sample sizes of 4, 5, and 6 are shown:
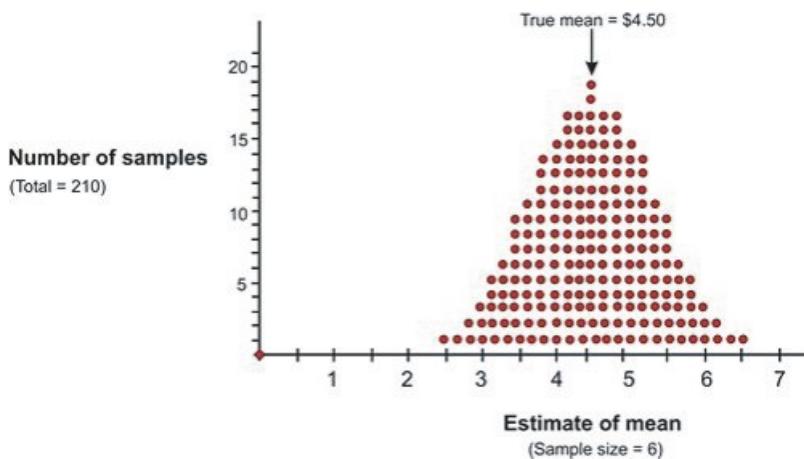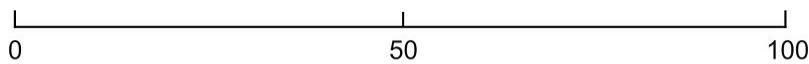
10 [Math] → [PRB] ↓ 3. $_nC_r$ 6 **enter** 210

From the graphs above, it is obvious that increasing the size of the samples chosen from the population of size 10 resulted in a distribution of the means that was more closely clustered around the true mean. If a sample of size 10 were selected, there would be only one possible sample, and it would yield the true mean of $4.50. Also, the sampling distribution of the *sample means* is approximately normal, as can be seen by the bell shape in each of the graphs.
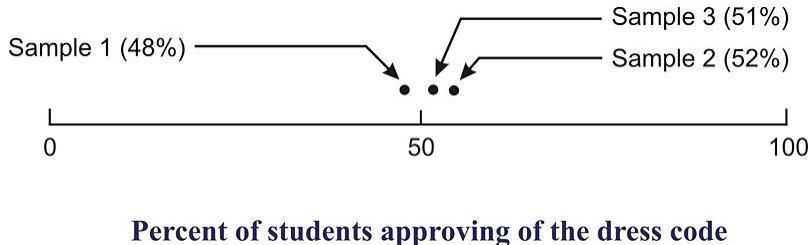
Now that you have been introduced to sampling distributions and how the sample size affects the distribution of the sample means, it is time to investigate a more realistic sampling situation. Assume you want to study the student population of a university to determine approval or disapproval of a student dress code proposed by the administration. The study's population will be the 18,000 students who attend the school, and the elements will be the individual students. A random sample of 100 students will be selected for the purpose of estimating the opinion of the entire student body, and attitudes toward the dress code will be the variable under consideration. For simplicity's sake, assume that the attitude variable has two variations: approve and disapprove. As you know from the last chapter, a scenario such as this in which a variable has two attributes is called binomial.

The following figure shows the range of possible sample study results. It presents all possible values of the parameter in question by representing a range of 0 percent to 100 percent of students approving of the dress code. The number 50 represents the midpoint, or 50 percent of the students approving of the dress code and 50 percent disapproving. Since the sample size is 100, at the midpoint, half of the students would be approving of the dress code, and the other half would be disapproving.
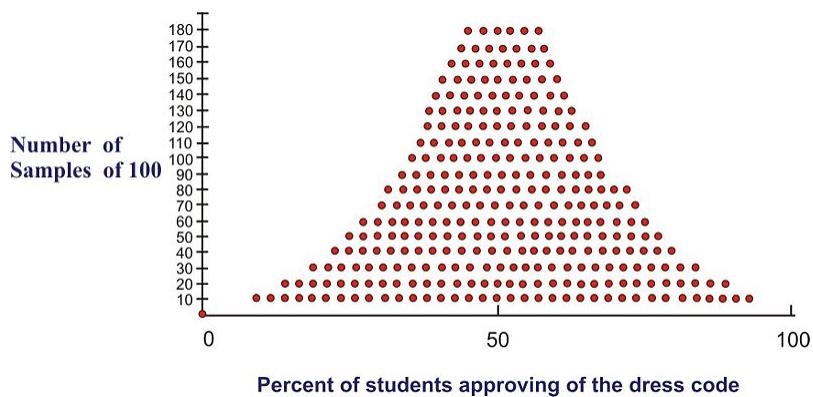
Percent of students approving of the dress code

To randomly select the sample of 100 students, every student is presented with a number from 1 to 18,000, and the sample is randomly chosen from a drum containing all of the numbers. Each member of the sample is then asked whether he or she approves or disapproves of the dress code. If this procedure gives 48 students who approve of the dress code and 52 who disapprove, the result would be recorded on the figure by placing a dot at 48%. This statistic is the *sample proportion*. Let's assume that the process was repeated, and it resulted in 52 students approving of the dress code. Let's also assume that a third sample of 100 resulted in 51 students approving of the dress code. The results are shown in the figure below.

Percent of students approving of the dress code

In this figure, the three different sample statistics representing the percentages of students who approved of the dress code are shown. The three random samples chosen from the population give estimates of the parameter that exists for the entire population. In particular, each of the random samples gives an estimate of the percentage of students in the total student body of 18,000 who approve of the dress code. Assume for simplicity's sake that the true proportion for the population is 50%. This would mean that the estimates are close to the true proportion. To more precisely estimate the true proportion, it would be necessary to continue choosing samples of 100 students and to record all of the results in a summary graph as shown:

Percent of students approving of the dress code

Notice that the statistics resulting from the samples are distributed around the population parameter. Although there is a wide range of estimates, most of them lie close to the 50% area of the graph. Therefore, the true value is likely to be in the vicinity of 50%. In addition, probability theory gives a formula for estimating how closely the sample statistics are clustered around the true value. In other words, it is possible to estimate the sampling error, or the degree of error expected for a given sample design. The formula $s = \sqrt{\dfrac{p(1-p)}{n}}$ contains three variables: the parameter, p, the sample size, n, and the *standard error, s*.

The symbols p and $1 - p$ in the formula represent the population parameters. For example, if 60 percent of the student body approves of the dress code and 40% disapproves, p and $1 - p$ would be 0.6 and 0.4,, respectively. The square root of the product of p and $1 - p$ is the population standard deviation. As previously stated, the symbol n represents the number of cases in each sample, and s is the standard error.

If the assumption is made that the true population parameters are 0.50 approving of the dress code and 0.50 disapproving of the dress code, when selecting samples of 100, the standard error obtained from the formula equals 0.05:

$$s = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05$$

This calculation indicates how tightly the sample estimates are distributed around the population parameter. In this case, the standard error is the standard deviation of the sampling distribution.

The Empirical Rule states that certain proportions of the sample estimates will fall within defined increments, each increment being one standard error from the population parameter. According to this rule, 34% of the sample

estimates will fall within one standard error above the population parameter, and another 34% will fall within one standard error below the population parameter. In the above example, you have calculated the standard error to be 0.05, so you know that 34% of the samples will yield estimates of student approval between 0.50 (the population parameter) and 0.55 (one standard error above the population parameter). Likewise, another 34% of the samples will give estimates between 0.5 and 0.45 (one standard error below the population parameter). Therefore, you know that 68% of the samples will give estimates between 0.45 and 0.55. In addition, probability theory says that 95% of the samples will fall within two standard errors of the true value, and 99.7% will fall within three standard errors. In this example, you can say that only three samples out of one thousand would give an estimate of student approval below 0.35 or above 0.65.

The size of the standard error is a function of the population parameter. By looking at the formula $s = \sqrt{\dfrac{p(1-p)}{n}}$, it is obvious that the standard error will increase as the quantity $p\,(1-p)$ increases. Referring back to our example, the maximum for this product occurred when there was an even split in the population. When $p = 0.5$, $p(1-p) = (0.5)(0.5) = 0.25$. If $p = 0.6$, then $p(1-p) = (0.6)(0.4) = 0.24$. Likewise, if $p = 0.8$, then $p(1-p) = (0.8)(0.2) = 0.16$. If p were either 0 or 1 (none or all of the student body approves of the dress code), then the standard error would be 0. This means that there would be no variation, and every sample would give the same estimate.

The standard error is also a function of the sample size. In other words, as the sample size increases, the standard error decreases, or the bigger the sample size, the more closely the samples will be clustered around the true value. Therefore, this is an inverse relationship. The last point about that formula that is obvious is emphasized by the square root operation. That is, the standard error will be reduced by one-half as the sample size is quadrupled.

On the Web

http://tinyurl.com/294stkw Explore the result of changing the population parameter, the sample size, and the number of samples taken for the proportion of Reese's Pieces that are brown or yellow.

Lesson Summary

In this lesson, we have learned about probability sampling, which is the key sampling method used in survey research. In the example presented above, the elements were chosen for study from a population by random sampling. The sample size had a direct effect on the distribution of estimates of the population parameter. The larger the sample size, the closer the sampling distribution was to a normal distribution.

Points to Consider

- 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?
- Are there any other components of sampling error estimates?

Review Questions

The following activity could be done in the classroom, with the students working in pairs or small groups. Before doing the activity, students could put their pennies into a jar and save them as a class, with the teacher also contributing. In a class of 30 students, groups of 5 students could work together, and the various tasks could be divided among those in each group.

1. If you had 100 pennies and were asked to record the age of each penny, predict the shape of the distribution. (The age of a penny is the current year minus the date on the coin.)
2. Construct a histogram of the ages of the pennies.
3. Calculate the mean of the ages of the pennies.

Have each student in each group randomly select a sample of 5 pennies from the 100 coins and calculate the mean of the five ages of the coins chosen. Have the students then record their means on a number line. Have the students repeat this process until all of the coins have been chosen.

4. How does the mean of the samples compare to the mean of the population (100 ages)? Repeat step 4 using a sample size of 10 pennies. (As before, allow the students to work in groups.)
5. What is happening to the shape of the sampling distribution of the sample means as the sample size increases?

Learning Objectives

- Understand the Central Limit Theorem and calculate a sampling distribution using the mean and standard deviation of a normally distributed random variable.
- Understand the relationship between the Central Limit Theorem and the normal approximation of a sampling distribution.

Introduction

In the previous lesson, you learned that sampling is an important tool for determining the characteristics of a population. Although the parameters of the population (mean, standard deviation, etc.) were unknown, random sampling was used to yield reliable estimates of these values. The estimates were plotted on graphs to provide a visual representation of the distribution of the sample means for various sample sizes. It is now time to define some properties of a sampling distribution of sample means and to examine what we can conclude about the entire population based on these properties.

Central Limit Theorem

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 7.1:

	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 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.
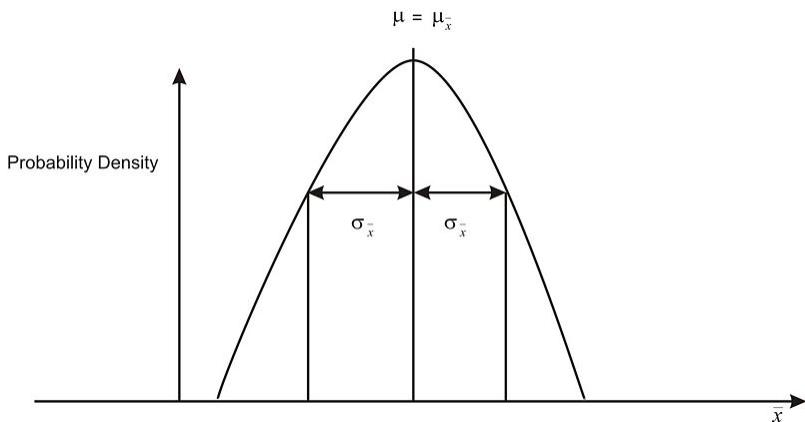
For the sampling distribution, $\mu_{\bar{x}} = \mu = 0.8$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{20}} = 0.268$.

We can use a graphing calculator as follows:

```
normalcdf(-1E99,
1.5,.8,.27)
          .9952371907
```

Therefore, the probability that the sample mean will be below 1.5 is 0.9952. 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.

The properties associated with the Central Limit Theorem are displayed in the diagram below:

The vertical axis now reads probability density, rather than frequency, since frequency can only be used when you are dealing with a finite number of sample means. Sampling distributions, on the other hand, are theoretical depictions of an infinite number of sample means, and probability density is the relative density of the selections from within this set.

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.

a) What is the expected shape of the resulting distribution?

b) Where is the sampling distribution of sample means centered?

c) 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.

a) The sampling distribution of the sample means will be approximately 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:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$\sigma_{\bar{x}} = \frac{4.3}{\sqrt{40}}$$
$$\sigma_{\bar{x}} = 0.68$$

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 value will the standard deviation of the sample means approach?

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 using '1 Vars Stat' 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: $\sigma_{\bar{x}} = \frac{4}{\sqrt{n}}$.

On the Web

http://tinyurl.com/2f969wj Explore how the sample size and the number of samples affect the mean and standard deviation of the distribution of sample means.

Lesson Summary
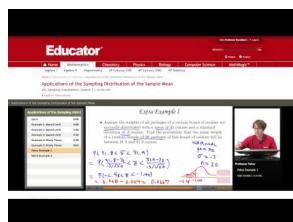
The Central Limit Theorem confirms the intuitive notion that as the sample size increases for a random variable, the distribution of the sample means will begin to approximate a normal distribution, with the mean equal to the mean of the underlying population and the standard deviation equal to the standard deviation of the population divided by the square root of the sample size, n.

Point to Consider

• How does sample size affect the variation in sample results?

Multimedia Links

For an example using the sampling distribution of x-bar **(15.0)(16.0)**, see EducatorVids, Statistics: Sampling Distribution of the Sample Mean (2:15).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1088

For another example of the sampling distribution of x-bar **(15.0)(16.0)**, see tcreelmuw, Distribution of Sample Mean (2:22).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1089
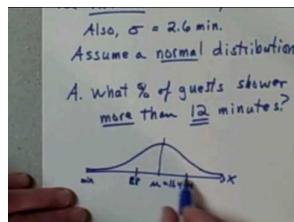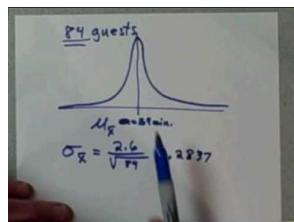
For an example of using the Central Limit Theorem **(9.0)**, see jsnider3675, Application of the Central Limit Theorem, Part 1 (5:44).

For the continuation of an example using the Central Limit Theorem (**9.0**), see jsnider3675, Application of the Central Limit Theorem, Part 2 (6:38).

Review Questions

1. A random sample of size 30 is selected from a known population with a mean of 13.2 and a standard deviation of 2.1. Samples of the same size are repeatedly collected, allowing a sampling distribution of sample means to be drawn.

   a. What is the expected shape of the resulting distribution?
   b. Where is the sampling distribution of sample means centered?
   c. What is the approximate standard deviation of the sample means?

2. What is the probability that a random sample of 40 families will have an average of 0.5 pets or fewer where the mean of the population is 0.8 and the standard deviation of the population is 1.2?

3. The scores of students on a college entrance exam were normally distributed with a mean of 19.4 and a standard deviation of 6.3.

   a. If a sample of 70 students who took the test (who have the same distribution as all scores) is collected, what are the mean and standard deviation of the sample mean for the 70 students?
   b. What is the probability that a random sample of 50 students will have an average score of 22 or higher?

4. 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 6000 new batteries, determine how many batteries will last:

   a. between 360 and 460 days.
   b. more than 320 days.
   c. less than 280 days.

7.3 Confidence Intervals

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
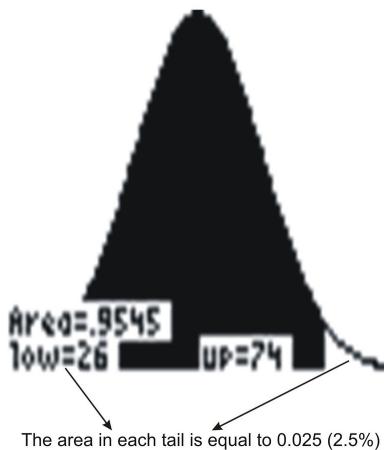
The objective of inferential statistics is to use sample data to increase knowledge about the entire population. In this lesson, we will examine how to use samples to make estimates about the populations from which they came. We will also see how to determine how wide these estimates should be and how confident we should be about them.

Confidence Intervals

Sampling distributions are the connecting link between the collection of data by unbiased random sampling and the process of drawing conclusions from the collected data. Results obtained from a survey can be reported as a *point estimate*. For example, a single sample mean is a point estimate, because this single number is used as a plausible value of the population mean. Keep in mind that some error is associated with this estimate-the true population mean may be larger or smaller than the sample mean. An alternative to reporting a point estimate is identifying a range of possible values the parameter might take, controlling the probability that the parameter is not lower than the lowest value in this range and not higher than the largest value. This range of possible values is known as a *confidence interval*. Associated with each confidence interval is a *confidence level*. This level indicates the level of assurance you have that the resulting confidence interval encloses the unknown population mean.

In a normal distribution, we know that 95% of the data will fall within two standard deviations of the mean. Another way of stating this is to say that we are confident that in 95% of samples taken, the sample statistics are within plus or minus two standard errors of the population parameter. As the confidence interval for a given statistic increases in length, the confidence level increases.

The selection of a confidence level for an interval determines the probability that the confidence interval produced will contain the true parameter value. Common choices for the confidence level are 90%, 95%, and 99%. These levels correspond to percentages of the area under the normal density curve. For example, a 95% confidence interval covers 95% of the normal curve, so the probability of observing a value outside of this area is less than 5%. Because the normal curve is symmetric, half of the 5% is in the left tail of the curve, and the other half is in the right tail of the curve. This means that 2.5% is in each tail.

Area=.9545
low=26   up=74

The area in each tail is equal to 0.025 (2.5%)

The graph shown above was made using a TI-83 graphing calculator and shows a normal distribution curve for a set of data for which $\mu = 50$ and $\sigma = 12$. A 95% confidence interval for the standard normal distribution, then, is the interval $(-1.96, 1.96)$, since 95% of the area under the curve falls within this interval. The ± 1.96 are the z-scores that enclose the given area under the curve. For a normal distribution, the *margin of error* is the amount that is added to and subtracted from the mean to construct the confidence interval. For a 95% confidence interval, the margin of error is 1.96σ. (Note that previously we said that 95% of the data in a normal distribution falls within ± 2 standard deviations of the mean. This was just an estimate, and for the remainder of this textbook, we'll assume that 95% of the data actually falls within ± 1.96 standard deviations of the mean.)

The following is the derivation of the confidence interval for the population mean, μ. In it, $z_{\frac{\alpha}{2}}$ refers to the positive z-score for a particular confidence interval. The Central Limit Theorem tells us that the distribution of $\bar{x}$ is normal, with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$. Consider the following:

$$-z_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}$$

All values are known except for μ. Solving for this parameter, we have:

$$-\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < -\mu < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} - \bar{x}$$

$$\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} > \mu > -z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} + \bar{x}$$

$$\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} > \mu > \bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$
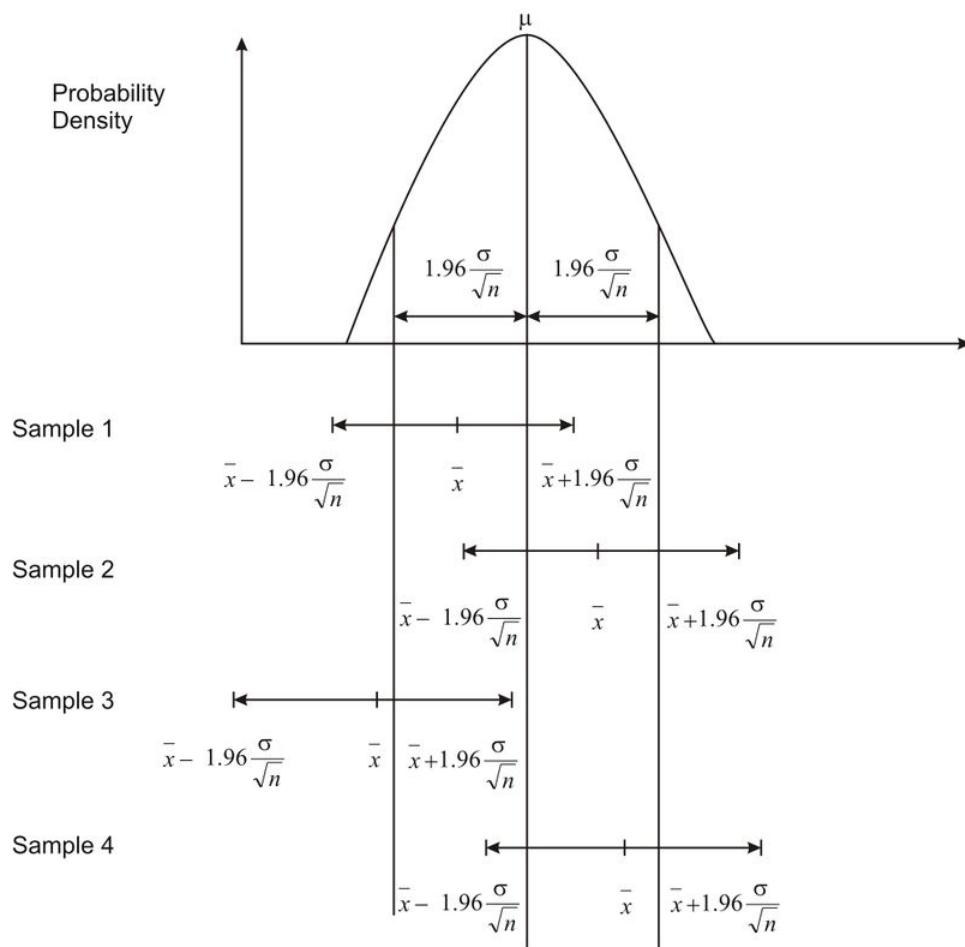
Another way to express this is: $\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$.

On the Web

http://tinyurl.com/27syj3x This simulates confidence intervals for the mean of the population.

Example: Jenny randomly selected 60 muffins of a particular brand and had those muffins analyzed for the number of grams of fat that they each contained. Rather than reporting the sample mean (point estimate), she reported the confidence interval. Jenny reported that the number of grams of fat in each muffin is between 10.3 grams and 12.1 grams with 95% confidence.

In this example, the population mean is unknown. This number is fixed, not variable, and the sample means are variable, because the samples are random. If this is the case, does the confidence interval enclose this unknown true mean? Random samples lead to the formation of confidence intervals, some of which contain the fixed population mean and some of which do not. The most common mistake made by persons interpreting a confidence interval is claiming that once the interval has been constructed, there is a 95% probability that the population mean is found within the confidence interval. Even though the population mean is unknown, once the confidence interval is constructed, either the mean is within the confidence interval, or it is not. Hence, any probability statement about this particular confidence interval is inappropriate. In the above example, the confidence interval is from 10.3 to 12.1, and Jenny is using a 95% confidence level. The appropriate statement should refer to the method used to produce the confidence interval. Jenny should have stated that the method that produced the interval from 10.3 to 12.1 has a 0.95 probability of enclosing the population mean. This means if she did this procedure 100 times, 95 of the intervals produced would contain the population mean. The probability is attributed to the method, not to any particular confidence interval. The following diagram demonstrates how the confidence interval provides a range of plausible values for the population mean and that this interval may or may not capture the true population mean. If you formed 100 intervals in this manner, 95 of them would contain the population mean.

Example: The following questions are to be answered with reference to the above diagram.

a) Were all four sample means within $1.96\frac{\sigma}{\sqrt{n}}$, or $1.96\sigma_{\bar{x}}$, of the population mean? Explain.

b) Did all four confidence intervals capture the population mean? Explain.

c) In general, what percentage of $\bar{x}'s$ should be within $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean?

d) In general, what percentage of the confidence intervals should contain the population mean?

a) The sample mean, $\bar{x}$, for Sample 3 was not within $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean. It did not fall within the vertical lines to the left and right of the population mean.

b) The confidence interval for Sample 3 did not enclose the population mean. This interval was just to the left of the population mean, which is denoted with the vertical line found in the middle of the sampling distribution of the sample means.

c) 95%

d) 95%

When the sample size is large $(n > 30)$, the confidence interval for the population mean is calculated as shown below:

$\bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$, where $z_{\frac{\alpha}{2}}$ is 1.96 for a 95% confidence interval, 1.645 for a 90% confidence interval, and 2.58 for a 99% confidence interval.

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:

$$107 \quad 112 \quad 109 \quad 115$$

a) For each sample, determine the 90% confidence interval.

b) Do all four confidence intervals enclose the population mean? Explain.

a)

$\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$107 \pm (1.645)(\dfrac{19}{\sqrt{60}})$

107 ± 4.04

from 102.96 to 111.04

$\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$112 \pm (1.645)(\dfrac{19}{\sqrt{60}})$

112 ± 4.04

from 107.96 to 116.04

$\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$109 \pm (1.645)(\dfrac{19}{\sqrt{60}})$

109 ± 4.04

from 104.96 to 113.04

$\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$

$115 \pm (1.645)(\dfrac{19}{\sqrt{60}})$

115 ± 4.04

from 110.96 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.

Technology Note: Simulation of Random Samples and Formation of Confidence Intervals on the TI-83/84 Calculator

Now it is time to use a graphing calculator to simulate the collection of three samples of sizes 30, 60, and 90, respectively. The three sample means will be calculated, as well as the three 95% confidence intervals. The samples will be collected from a population that displays a normal distribution, with a population standard deviation of 108 and a population mean of 2130. We will use the **randNorm(** function found in **[MATH]**, under the **PRB** menu. First, store the three samples in **L1**, **L2**, and **L3**, respectively, as shown below:

Store 'randNorm(μ, σ, n)' in **L1**. The sample size is $n = 30$.

Store 'randNorm(μ, σ, n)' in **L2**. The sample size is $n = 60$.

Store 'randNorm(μ, σ, n)' in **L3**. The sample size is $n = 90$.

The lists of numbers can be viewed by pressing **[STAT][ENTER]**. The next step is to calculate the mean of each of these samples.

To do this, first press **[2ND][LIST]** and go to the **MATH** menu. Next, select the 'mean(' command and press **[2ND][L1][ENTER]**. Repeat this process for **L2** and **L3**.

Note that your confidence intervals will be different than the ones calculated below, because the random numbers generated by your calculator will be different, and thus, your means will be different. For us, the means of **L1**, **L2**, and **L3** were 2139.1, 2119.2, and 2137.1, respectively, so the confidence intervals are as follows:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$2139.1 \pm (1.96)(\frac{108}{\sqrt{30}})$$

2139.1 ± 38.65

from 2100.45 to 2177.65

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$2119.2 \pm (1.96)(\frac{108}{\sqrt{60}})$$

2119.2 ± 27.33

from 2091.87 to 2146.53

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$2137.1 \pm (1.96)(\frac{108}{\sqrt{90}})$$

2137.1 ± 22.31

from 2114.79 to 2159.41

As was expected, the value of $\bar{x}$ varied from one sample to the next. The other fact that was evident was that as the sample size increased, the length of the confidence interval became smaller, or decreased. This is because with the increase in sample size, you have more information, and thus, your estimate is more accurate, which leads to a narrower confidence interval.

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. This formula is valid only for simple random samples. Since $z_{\frac{\alpha}{2}}\left(\frac{s_x}{\sqrt{n}}\right)$ is the margin of error, a confidence interval can be thought of simply as: $\bar{x} \pm$ the margin of error.

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.

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 Hypotheses about Population Proportions
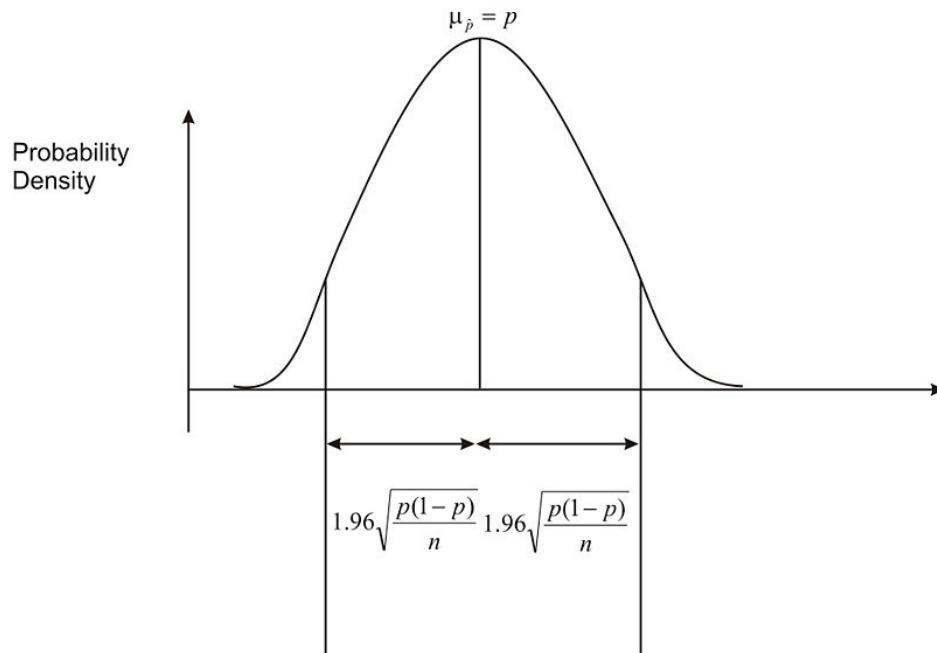
Often statisticians are interested in making inferences about a population proportion. For example, when we look at election results we often look at the proportion of people that vote and who this proportion of voters choose. Typically, we call these proportions percentages and we would say something like "Approximately 68 percent of the population voted in this election and 48 percent of these voters voted for Barack Obama."

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:

1. We must have a random sample.
2. 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.

$\sqrt{\dfrac{p(1-p)}{n}}$ 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, $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$. 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.

That is, we want the following:

$$-1.96 \text{ standard errors} < \hat{p} - p < 1.96 \text{ standard errors}$$

$$-\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < -p < -\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} > p > \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is a 95% confidence interval for the population proportion. If we generalize for any confidence level, the confidence interval is as follows:

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In other words, the confidence interval is $\hat{p} \pm z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$. Remember that $z_{\frac{\alpha}{2}}$ 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 remove all speed limits on interstate highways. 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.

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:

$$0.746 \pm (1.96)(0.0251)$$
$$0.697 < p < 0.795$$

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, 95 out of the 100 confidence intervals we constructed would capture the population proportion, p.

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.

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.

$$77\% - 2.9\% = 74.1\% \qquad 77\% + 2.9\% = 79.9\%$$

Thus, the confidence interval is from 0.741 to 0.799.

d) The 95% confidence interval from 0.741 to 0.799 for the population proportion is an interval calculated from a sample by a method that has a 0.95 probability of capturing the population proportion.

On the Web

http://tinyurl.com/27syj3x This simulates confidence intervals for the population proportion.

http://tinyurl.com/28z97lr Explore how changing the confidence level and/or the sample size affects the length of the confidence interval.

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%. You also learned that the probability is attributed to the method used to calculate the confidence interval.

In addition, you learned that you calculate the confidence interval for a population proportion by using the formula $\hat{p} \pm z_{\frac{\alpha}{2}} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$.

Points to Consider

- Does replacing σ with s change your chance of capturing the unknown population mean?
- Is there a way to increase the chance of capturing the unknown population mean?

Multimedia Links
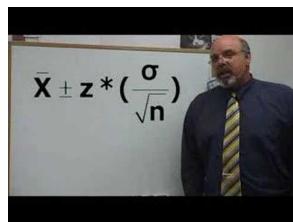
For an explanation of the concept of confidence intervals (**17.0**), see kbower50, What are Confidence Intervals? (3:24).
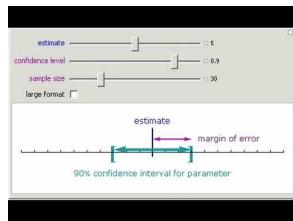
For a description of the formula used to find confidence intervals for the mean **(17.0)**, see mathguyzero, Statistics Confidence Interval Definition and Formula (1:26).
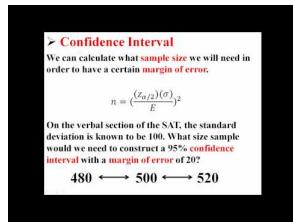
For an interactive demonstration of the relationship between margin of error, sample size, and confidence intervals **(17.0)**, see wolframmathematica, Confidence Intervals: Confidence Level, Sample Size, and Margin of Error (0:16).

For an explanation on finding the sample size for a particular margin of error **(17.0)**, see statslectures, Calculating Required Sample Size to Estimate Population Mean (2:18).

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

Confidence interval

Confidence level

Margin of error

Parameter

Point estimate

Sample means

Sample proportion

Sampling distributions

Standard error

CHAPTER **8**

Hypothesis Testing

Chapter Outline

8.1 Hypothesis Testing and the P-Value

Learning Objectives

- Develop null and alternative hypotheses to test for a given situation.
- Understand the critical regions of a graph 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.
- Understand Type I and Type II errors.
- Calculate the power of a test.

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.
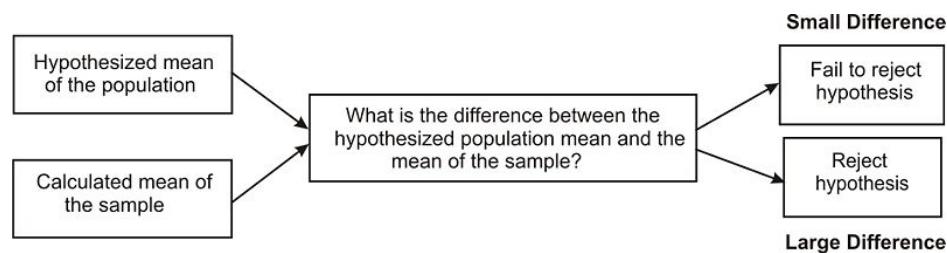
A hypothesis is simply a conjecture about a characteristic or set of facts. When performing statistical analyses, our hypotheses provide the general framework of what we are testing and how to perform the test.

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 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 populations in hypothesis testing, we are essentially determining the magnitude of the difference between the mean of the sample and they 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 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

$$H_0 : \mu = 3.2$$
$$H_a : \mu \neq 3.2$$

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 2- *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. Consider $H_0 : \mu \leq 3.2$ $H_a : \mu > 3.2$. In this situation we would reject H_0 for very large values of $\bar{x}$. This is called a *one tail test.* If, for this test, our data gives $\bar{x} = 15$, it would be highly unlikely that finding $\bar{x}$ this different from 3.2 would occur by chance and 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 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 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 that there is a difference between the null and alternative hypotheses 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 is different than 14 milligrams. This is two tailed.

Example: The 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 two tailed.
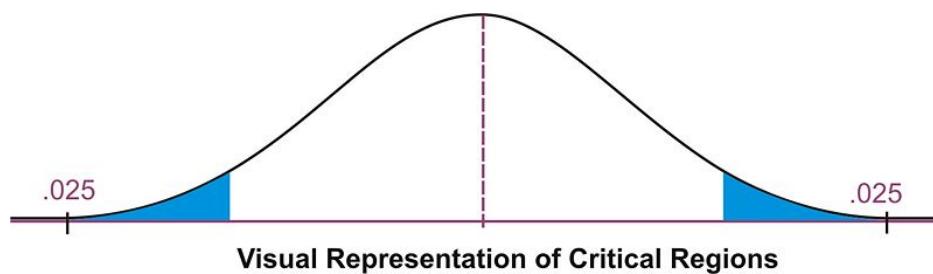
Deciding Whether to Reject the Null Hypothesis: One-Tailed and Two-Tailed Hypothesis Tests

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 that one trial. It is too likely that the discrepancy was merely due to chance.

We use statistical tests to determine if the sample data give good evidence against the claim (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, $\alpha = .01$ we are saying that we would get data at least as unusual as the data we have collected no more than 1% of the time when H_0 is true.

The most frequently used levels of significance are 0.05 and 0.01. If our data results in a statistic that falls within the region determined by the level of significance then we reject H_0. The region is therefore called the *critical region.* When choosing the level of significance, we need to consider the consequences of rejecting or failing to reject the null hypothesis. If there is the potential for health consequences (as in the case of active ingredients in prescription medications) or great cost (as in the case of manufacturing machine parts), we should use a more 'conservative' critical region with levels of significance such as .005 or .001.

When determining the critical regions for a two-tailed hypothesis test, the level of significance represents the extreme areas under the normal density curve. We call this a two-tailed hypothesis test because the critical region is located in both ends of the distribution. For example, if there was a significance level of 0.95 the critical region would be the most extreme 5 percent under the curve with 2.5 percent on each tail of the distribution.

Visual Representation of Critical Regions

Therefore, if the mean from the sample taken from the population falls within 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 the 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 hypothesis simply states that the mean is greater or less than the hypothesized value.
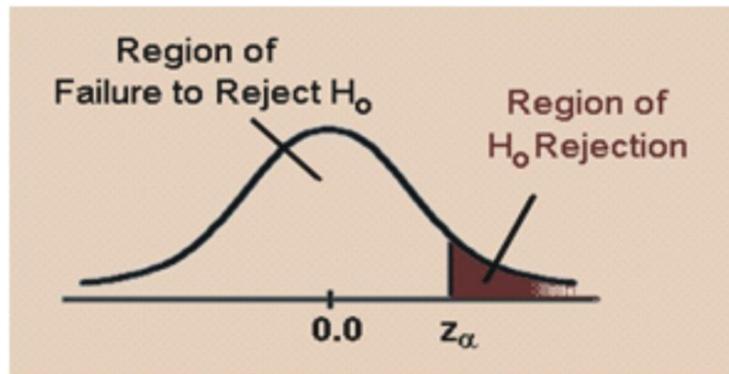
When performing a single-tail hypothesis test, our alternative hypothesis looks a bit different. When developing the alternative hypothesis in a single-tail hypothesis test we would use the symbols of greater than or less than. Using our example about SAT scores of graduating seniors, our null and alternative hypothesis could look something like:

$$H_0 : \mu = 1100$$
$$H_a : \mu > 1100$$

In this scenario, our null hypothesis states that the mean SAT scores would be equal to 1100 while the alternate hypothesis states that the SAT scores 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, the critical region is on the right side of the distribution. When the alternative hypothesis is that the sample is smaller, the critical region is on the left side of the distribution (see below).

To calculate the critical regions, we must first find the critical values or the cut-offs where the critical regions start. To find these values, we use the critical values found specified by the $z-$distribution. These values 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-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 alternative hypothesis is 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 the mean is less than a specified value the critical value would be -1.645.

Technology Note: Finding critical z values on the TI83/84 Calculator

You can also find this critical value using the TI83/84 calculator: 2nd **[DIST]** invNorm(.05,0,1) returns -1.64485. The syntax for this 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 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 distribution of $\overline{X}$ 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 formula:

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

$z =$ standardized score

$\bar{x}$ = sample mean

μ = the population mean under the null hypothesis

σ = population standard deviation. 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.

Following are the steps you must take when doing an hypothesis test:

1. Determine the null and alternative hypotheses.
2. Verify that necessary conditions are satisfied and summarize the data into a test statistic.
3. Determine the α level.
4. Determine the critical region(s).
5. Make a decision (Reject or fail to reject the null hypothesis)
6. 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. What are our hypotheses and the test statistic?

1. Let's first develop our null and alternative hypotheses:

$$H_0 : \mu = 1500$$
$$H_a : \mu \neq 1500$$

2. The test statistic is $z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(1450 - 1500)}{\frac{100}{\sqrt{125}}} \approx -5.59$

3. Choose $\alpha = .05$

4. This is a two sided test. If we choose $\alpha = .05$, the critical values will be -1.96 and 1.96. (Use invNorm (.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.

5. The value of the test statistic is -5.59. This is less than -1.96 and so our decision is to reject H_0.

6. Based on this sample we believe that the mean is not equal to 1500.

Example: A farmer is trying out a planting technique that he hopes will increase the yield on 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 the test statistic?

1. First, we develop our null and alternative hypotheses:

$$H_0 : \mu = 145$$
$$H_a : \mu > 145$$

This alternative hypothesis is >since he believes that there might be a gain in the number of pods.

2. Next, we calculate the test statistic for the sample of pea plants.

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(147 - 145)}{\frac{100}{\sqrt{144}}} \approx 0.24$$

3. If we choose $\alpha = .05$

4. The critical value will be 1.645. (Use invNorm (.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.

5. This is less than 1.645 and so our decision is to accept H_0.

6. Based on our sample we believe the mean is equal to 145.

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 we did if the null hypothesis is true?" This is called the $p-value.$

Example: Let's use the example about 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 a previous average of 145 pods, the null hypothesis). To determine the p-value we ask what is $P(z > .24)$? That is, what is the probability of obtaining a z value greater than .24 if the null hypothesis is true? Using the calculator **(normcdf (.24, 99999999, 0, 1)** we find this probability to be .405. This indicates that there is a 40.5% chance that under the null hypothesis the peas will produce 147 or more pods.

Type I and Type II Errors

When we decide to reject or not 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: accepting the null hypothesis when it is true and rejections 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 accepting 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**
Accept 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 may be 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.

$$\beta = P(\text{accepting } H_0 | H_0 \text{ is false}) = P(\text{making a type II error})$$

Once the value for the alternative hypothesis has been specified, it is possible to determine the probability of making a correct decision $(1 - \beta)$. This quantity, $1 - \beta$, is called the power of the 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.

On the Web

http://tinyurl.com/35zg7du This link leads you to a graphical explanation of the relationship between α and β

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.

Calculating the Power of a Test

The power of a test is defined as the probability of rejecting the null hypothesis when it is false (that is, making the correct decision). Obviously, we want to maximize this power if we are concerned about making Type II errors. To determine the power of the test, there must be a specified value for the alternative hypothesis.

Example: Suppose that a doctor is concerned about making a Type II error only if the active ingredient in the new medication is greater than 3 milligrams higher than what was specified in the null hypothesis (say, 250 milligrams with a sample of 200 and a standard deviation of 50). Now we have values for both the null and the alternative hypotheses.

$$H_0 : \mu = 250$$
$$H_a : \mu = 253$$

By specifying a value for the alternative hypothesis, we have selected one of the many values for H_a. In determining the power of the test, we must assume that H_a is true and determine whether we would correctly reject the null hypothesis

Calculating the exact value for the power of the test requires determining the area above the critical value set up to test the null hypothesis when it is re-centered around the alternative hypothesis. If we have an alpha level of .05 our critical value would be 1.645 for the one tailed test. Therefore,

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

$$1.645 = \frac{(\bar{x} - 250)}{\frac{50}{\sqrt{200}}}$$

Solving for $\bar{x}$ we find: $\bar{x} = 1.645 \left(\frac{50}{\sqrt{200}} \right) + 250 \approx 255.8$

Now, with a new mean set at the alternative hypothesis $H_a : \mu = 253$ we want to find the value of the critical score when centered around this score when we center this $\bar{x}$ around the population mean of the alternative hypothesis, $\mu = 253$. Therefore, we can figure that:

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(255.8 - 253)}{\frac{50}{\sqrt{200}}} \approx 0.79$$

Recall that we reject the null hypothesis if the critical value is to the right of .79. The question now is what is the probability of rejecting the null hypothesis when, in fact, the alternative hypothesis is true? We need to find the area to the right of 0.79. You can find this area using a z table or using the calculator with the Normcdf command (Invnorm (0.79, 9999999, 0, 1)). The probability is .2148. This means that since we assumed the alternative hypothesis to be true, there is only a 21.5% chance of rejecting the null hypothesis. Thus, the power of the test is .2148. In other words, this test of the null hypothesis is not very powerful and has only a 0.2148 probability of detecting the real difference between the two hypothesized means.

There are several things that affect the power of a test including:

- Whether the alternative hypothesis is a single-tailed or two-tailed test.
- The level of significance α
- The sample size.

On the Web

http://intuitor.com/statistics/CurveApplet.html Experiment with changing the sample size and the distance between the null and alternate hypotheses and discover what happens to the power.

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 alpha (α) level. If the value of the test statistic falls in 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.
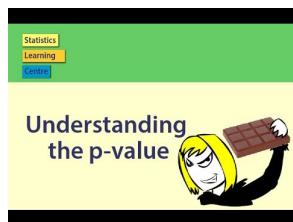
When we make a decision about a hypothesis, there are four different outcome and possibilities and two different types of errors. A Type I error is when we reject the null hypothesis when it is true and a Type II error is when we do not reject the null hypothesis, even when it is false. α, the level of significance of the test, is the probability of rejecting the null hypothesis when, in fact, the null hypothesis 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). We determine the power of a test by assigning a value to the alternative hypothesis and

using the $z-$score to calculate the probability of rejecting the null hypothesis when it is false. It is the probability of making a Type II error.

Multimedia Links

For an illustration of the use of the p-value in statistics **(4.0)** and how to interpret it **(18.0)**, see UCMSCI, Understanding the P-Value (4:04)

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/21645

Review Questions

1. If the difference between the hypothesized population mean and the mean of the sample is large, we ___ the null hypothesis. If the difference between the hypothesized population mean and the mean of the 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 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 the $z-$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 the test statistic?
6. During hypothesis testing, we use the $p-$value to predict the ___ of an event occurring if the null hypothesis is true.
7. A survey shows that California teenagers have an average of $500 in savings (standard error $=100$). What is the probability that a randomly selected teenager will have savings greater than $520?
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) ___

9. The ___ is defined as the probability of rejecting the null hypothesis when it is false (making the correct decision). We want to maximize___if we are concerned about making Type II errors.

10. The Governor's economic committee is investigating average salaries of recent college graduates in California. They decide to test the null hypothesis that the average salary is $24,500 (standard deviation is $4,800)) and is concerned with making a Type II error only if the average salary is less than $25,000. $H_a : \mu = \$25,100$ For an $\alpha = .05$ and a sample of 144 determine the power of a one-tailed test.

8.2 Testing a Proportion Hypothesis

Learning Objectives

- Test a hypothesis about a population proportion by applying the binomial distribution approximation.
- 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 (> 30).

Often statisticians are interest in making inferences about a population proportion. For example, when we look at election results we often look at the proportion of people that vote and who this proportion of 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 populations 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 by Applying the Binomial Distribution Approximation

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,

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 follow a series of steps:

- 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 proportion. We use the binomial distribution which illustrates situations in which two outcomes are possible (for example, voted for a

candidate, didn't vote for a candidate), remembering that when the sample size is relatively large, we can use the normal distribution to approximate the binomial distribution. The test statistic is

$$z = \frac{\text{sample estimate} - \text{value under the null hypothesis}}{\text{standard error under the null hypothesis}}$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

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 the test statistic?

Since we want to test the proportion of graduating seniors and we think that proportion is around 60 percent, our hypotheses are:

$$H_0 : p = .6$$
$$H_a : p \neq .6$$

The test statistic would be $z = \frac{\hat{p} - .6}{\sqrt{\frac{.6(1-.6)}{n}}}$. To complete this calculation we would have to have a value for the sample size (n).

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.
- Set the criterion for rejecting the null hypothesis.
- Calculate the test statistic.
- Decide whether to reject or fail to reject the null hypothesis.
- Interpret your 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 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.

$$H_0 : p = .7$$
$$H_a : p > .7$$

Next, we should set the criterion for rejecting the null hypothesis. Choose $\alpha = .05$ and since the null hypothesis is considering $p > .7$, this is 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 .05 to be 1.645.

The test statistic is $z = \frac{.74 - .7}{\sqrt{\frac{.7(1-.7)}{300}}} \approx 1.51$

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 conclude that the population proportion is greater than .70 with 95 percent certainty. 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 proportion of voters supporting the bill is over 70 percent, this could be due to chance and is not statistically significant.

Example: Admission staff from a local university is conducting a survey to determine the proportion of incoming freshman that 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 this an accurate guess?

First, we develop our null and alternative hypotheses.

$$H_0 : p = .3$$
$$H_a : p \neq .3$$

Next, we should set the criterion for rejecting the null hypothesis. The .05 alpha level is used and for a two tailed test the critical values of the test statistic are 1.96 and -1.96.

To calculate the test statistic:

$$z = \frac{.25 - .3}{\sqrt{\frac{.3(1-.3)}{400}}} \approx -2.18$$

Since our critical values are ± 1.96 and $-2.18 < -1.96$ we can reject the null hypothesis. This means that we can conclude that the population of freshman needing financial aid is significantly more or less than 30 percent. Since the test statistic is negative, we can conclude with 95% certainty that in the population of incoming freshman, less than 30 percent of the students will need financial aid.

Lesson Summary

In statistics, we also make inferences about proportions of a population. We use the same process as in testing hypotheses about 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

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:
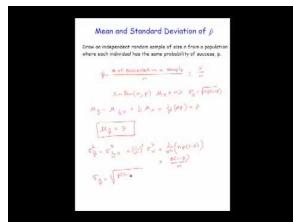
$\hat{p}$ is the sample proportion

p_0 is the hypothesized population proportion

We can construct something called the confidence interval that specifies the level of confidence that we have in our results. The formula for constructing a confidence interval for the population proportion is $\hat{p} \pm z_{\frac{\alpha}{2}} \left(\frac{\hat{p}(1-\hat{p})}{n} \right)$.

Multimedia Links

For an explanation on finding the mean and standard deviation of a sampling proportion, p, and normal approximation to binomials **(7.0)(9.0)(15.0)(16.0)**, see American Public University, Sampling Distribution of Sample Proportion (8:24)

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1100

For a calculation of the z-statistic and associated P-Value for a 1-proportion test **(18.0)**, see kbower50, Test of 1 Proportion: Worked Example (3:51)

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1101

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 .01?

8.3 Testing a Mean Hypothesis

Evaluating Hypotheses for Population Means using Large Samples

When testing a hypothesis for the mean of a normal distribution, we follow a series of four basic steps:

1. State the null and alternative hypotheses.
2. Choose an α level
3. Set the criterion (critical values) for rejecting the null hypothesis.
4. Compute the test statistic.
5. Make a decision (reject or fail to reject the null hypothesis)
6. 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 .05 significance level to evaluate the null and alternative hypotheses.

First, we develop our null and alternative hypotheses:

$$H_0 : \mu = 145$$
$$H_a : \mu > 145$$

Choose $\alpha = .05$. The critical value for this one tailed test is 1.64. Any test statistic greater than 1.64 will be in the rejection region.

Next, we calculate the test statistic for the sample of 7th graders.

$$z = \frac{147 - 145}{\frac{20}{\sqrt{200}}} \approx 1.414$$
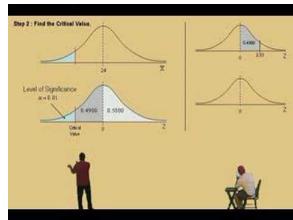
Since the calculated $z-$score of 1.414 is smaller than 1.64 and thus does not fall in the critical region. Our decision is to fail to reject the null hypothesis and conclude that the probability of obtaining a sample mean equal to 147 if the mean of the population is 145 is likely to have been due to chance.

When testing a hypothesis for the mean of a distribution, we follow a series of six basic steps:

1. State the null and alternative hypotheses.
2. Choose α
3. Set the criterion (critical values) for rejecting the null hypothesis.
4. Compute the test statistic.
5. Decide about the null hypothesis
6. Interpret our results.

Multimedia Links

For an step by step example of testing a mean hypothesis **(4.0)**, see MuchoMath, Z Test for the Mean (9:34).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1102

Review Questions

1. True or False: 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.
2. The dean from UCLA is concerned that the student's 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.

    a. What would the null and alternative hypotheses be for this scenario?
    b. What would the standard error be for this particular scenario?
    c. Describe in your own words how you would set the critical regions and what they would be at an alpha level of .05.
    d. Test the null hypothesis and explain your decision

3. For each of the following scenarios, state which one is more likely to lead to the rejection of the null hypothesis?

    a. A one-tailed or two-tailed test
    b. .05 or .01 level of significance
    c. A sample size of $n = 144$ or $n = 444$

8.4 Student's t-Distribution

Learning Objectives
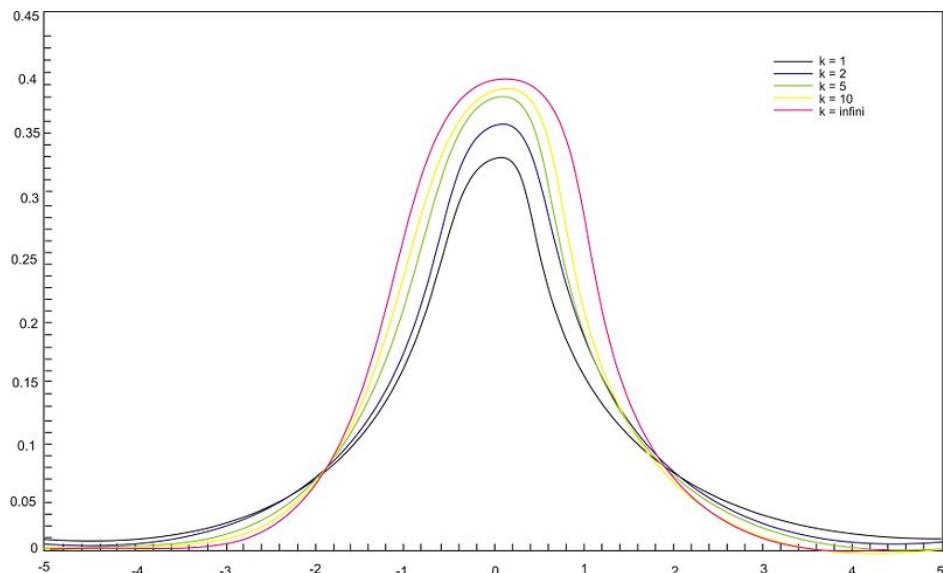
- Use Student's $t-$distribution to estimate population mean interval for smaller samples.
- Understand how the shape of Student's $t-$distribution corresponds to the sample size (which corresponds to a measure called the "degrees of freedom.")

Introduction

Hypothesis Testing with Small Populations and Sample Sizes

Back in the early 1900's a chemist at a brewery in Ireland discovered that when he was working with very small samples, the distributions of the mean 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."

$T-$distributions are a family of distributions that, like the normal distribution, are symmetrical and 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 the Student's $t-$distribution in hypothesis testing the same way that we use the normal distribution. Each row in the t distribution table (see link below) 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 used the $z-$scores.

http://tinyurl.com/ygcc5g9 Follow this link to the Student's $t-$table.

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 30 - we would use the normal distribution or the $z-$table instead. Note that usually in practice, if the standard deviation is known then the normal distribution is used regardless of the sample size.

In calculating the $t-$test statistic, we use the 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: The high school athletic director is asked if football players are doing as well academically as the other student athletes. We know from a previous study that the average GPA for the student athletes is 3.10. After an initiative to help improve the GPA of student athletes, the athletic director samples 20 football players and finds that the average GPA of the sample is 3.18 with a sample standard deviation of 0.54. Is there a significant improvement? Use a .05 significance level.

First, we establish our null and alternative hypotheses.

$$H_0 : \mu = 3.10$$
$$H_a : \mu \neq 3.10$$

Next, we use our alpha level of .05 and the $t-$distribution table to find our critical values. For a two-tailed test with 19 degrees of freedom and a .05 level of significance, our critical values are equal to ± 2.093.

In calculating the test statistic, we use the formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.18 - 3.10}{\frac{.54}{\sqrt{20}}} \approx 0.66$$

This means that the observed sample mean 3.18 of football players is .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 estimated 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 as 3.21 grams with a standard deviation of 0.067. What is the value of the test statistic for a test to determine how the mean differs from the estimated mean?

Solution:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{3.21 - 3.16}{\frac{0.067}{\sqrt{11}}}$$
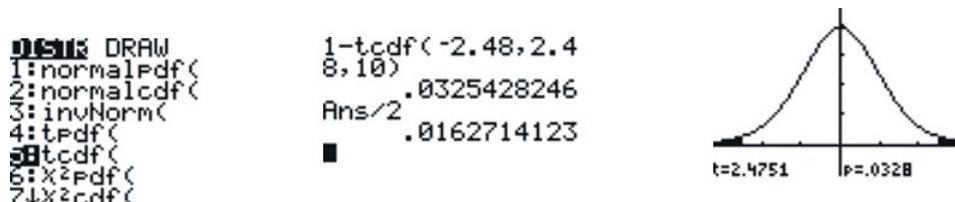
$$t \approx 2.48$$

If the value of t from the sample fits right into the middle of the distribution of t constructed by assuming the null hypothesis is true, the null hypothesis is true. On the other hand, if the value of t from the sample is way out in the tail of the $t-$distribution, then there is evidence to reject the null hypothesis. Now that the distribution of t is known when the null hypothesis is true, the location of this value on the distribution. The most common method used to determine this is to find a $p-$value (observed significance level). The $p-$value is a probability that is computed with the assumption that the null hypothesis is true.

The $p-$value for a two-sided test is the area under the $t-$distribution with $df = 11 - 1 = 10$ 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 to calculate probabilities associated with the t distribution

Press **2ND [DIST]** Use ↓ to select 5.tcdf (lower bound, upper bound, degrees of freedom) This will be the total area under both tails. To calculate the area under one tail divide by 2.

There is only a .016 chance of getting an absolute 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. The population mean differs from the estimated mean of 3.16.
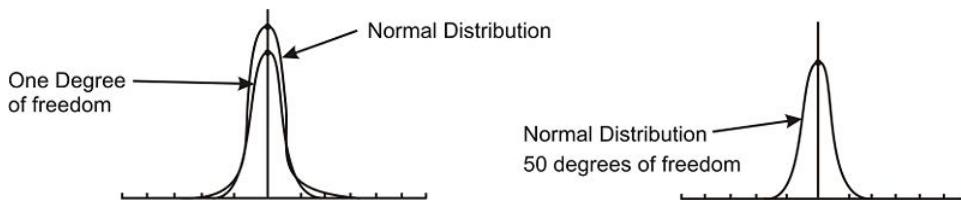
When the $p-$value is close to zero, there is strong evidence against the null hypothesis. When the $p-$value is large, the result from the sample is consistent with the estimated or hypothesized mean and there is no evidence against the null hypothesis.

A visual picture of the $P-$value can be obtained by using the graphing calculator.

Enter   T-Test
                   Inpt:Data **Stats**
                   μ0:3.16
                   x̄:3.21
                   Sx:.067
                   n:11
                   μ:**≠μ0** <μ0 >μ0
                   Calculate Draw   **Highlight Draw   Enter**

t=2.4751   p=.0328

The spread of any t distribution is greater than that of a standard normal distribution. This is due to the fact that 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.

Normal Distribution

One Degree
of freedom

Normal Distribution
50 degrees of freedom
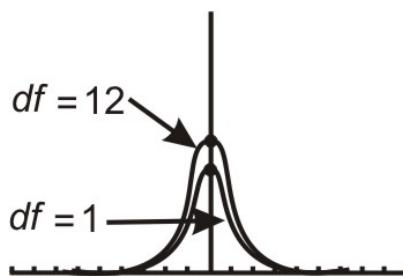
Notice in the first distribution graph the spread of the first (inner curve) is small but in the second one the both distributions are basically overlapping, so are roughly normal. This is due to the increase in the degrees of freedom.

Here are the $t-$distributions for $df = 1$ and for $df = 12$ as graphed on the graphing calculator

$Y = 2^{nd} \boxed{\text{Vars}}$ (Dist) $\downarrow$ 4. **tpdf(**

You are now on the $Y =$ screen.

$Y = \text{tpdf}(X, 1)$ **[Graph]**

$df = 12$

$df = 1$

Repeat the steps to plot more than one $t-$distribution on the same screen.

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. The Central Limit Theorem states the sampling distribution of a statistic will be close to normal with a large enough sample size. As a rough estimate, the Central Limit Theorem predicts a roughly normal distribution under 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.

The $t-$distribution also has some unique properties. These properties are:

- The mean of the distribution equals zero.
- The population standard deviation is unknown.
- The variance is equal to the degrees of freedom divided by the degrees of freedom minus 2. This means that the degrees of freedom must be greater than two to avoid the expression being undefined.
- The variance is always greater than one, although it approaches [1] as the degrees of freedom increase. This is due to the fact that as the degrees of freedom increase, the distribution is becoming more of a normal distribution.
- Although the Student $t-$distribution is bell-shaped, the smaller sample sizes produce a flatter curve. The distribution is not as mounded 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 life span of no more than 290 hours?

$t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ The degrees of freedom are $(n-1) = 15 - 1$. This means 14 degrees of freedom.

$t = \dfrac{290 - 300}{\frac{50}{\sqrt{15}}}$

$t = \dfrac{-10}{12.9099}$

$t = -.7745993$

Using the graphing calculator or a table of values, the cumulative probability is 0.226, which means that if the true life span of a battery were 300 hours, there is a 22.6% chance that the life span of the 15 tested batteries would be less than or equal to 290 days. This is not a high enough level of confidence to reject the null hypothesis and count the discrepancy as significant.

$$Y = 2^{nd} \boxed{Vars} \text{ (Dist)} \downarrow 4.\textbf{tcdf(}$$

You are now on the $Y =$ screen.

$Y = \text{tcdf}(-1E99, -.7745993, 14) = [0.226]$

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. Are these differences due to chance or is the distributor giving you less cheese than you deserve?

a) State the hypotheses.

b) Calculate the test statistic.

c) Find and interpret the p-value.

d) Would the null hypothesis be rejected at the 10% level? The 5% level? The 1% level?

Solution:

a) For H_0 the mean weight of cheese $\mu = 72$; and for $H_a : \mu \neq 72$.

b) Begin by determining the mean of the sample and the sample standard deviation. This can be done using the graphing calculator. $\bar{x} = 70.143$ and $s = 1.676$.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{70.143 - 72}{\frac{1.676}{\sqrt{7}}}$$

$$t \approx -2.9315$$

c) The test statistic computed in part b) was -2.9315. Using technology, the p value is .0262. If the mean weight of cheese is 72 ounces, the probability that the weight of 7 random measurements would give a value of t greater than 2.9315 or less than -2.9315 is about 0.0262.

d) Because the $p-$value of 0.0262 is less than both .10 and .05, the null hypothesis would be rejected at these levels. However, the $p-$value is greater than .01 so the null hypothesis would not be rejected if this level of confidence was required.

Lesson Summary
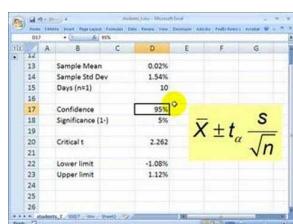
A test of significance is done when a claim is made about the value of a population parameter. The test can only be conducted if the random sample taken from the population came from a distribution that is normal or approximately normal. When you use s to estimate σ, you must use t instead of z to complete the significance test for a mean.

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?

Multimedia Links

For an explanation of the T distribution and an example using it **(7.0)(17.0)**, see bionicturtledotcom, Student's t distribution (8:32).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1103

Review Questions

1. 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.

2. You intend to use simulation to construct an approximate $t-$distribution with 8 degrees of freedom by taking random samples from a population with bowling scores that are normally distributed with mean, $\mu = 110$ and standard deviation, $\sigma = 20$.

   a. Explain how you will do one run of this simulation.
   b. Produce four values of t using this simulation.

3. 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 30 seniors from the last graduating class and finds that their mean GP is 2.85 with a sample standard deviation of 0.65. Suppose that the dean samples only 30 students. Would a $t-$distribution now be the appropriate sampling distribution for the mean? Why or why not?

4. Using the appropriate $t-$distribution, test the same null hypothesis with a sample of 30.

5. 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 to obtain statistical significance than with a sample size of 256? Explain your answer.

Testing a Hypothesis for Dependent and Independent Samples

Learning Objectives

- Identify situations that contain dependent or independent samples.
- Calculate the pooled standard deviation for two 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 proportion and means with both large and small samples. However, in the examples in those lessons only one sample was involved. In this lesson we will apply the principals 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 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 need to select a random sample from the population, measure that sample statistic and then make 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, in the case of medical research, the subjects are randomly assigned to a group), the two samples will vary only by chance and the difference will not be statistically significant. In short, when we have independent samples we assume that the scores of one sample do not affect the other.

Independent samples can occur in two scenarios.

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.

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 populations.

Dependent samples are a bit different. Two samples of data are dependent when each score in one sample is paired with a specific score in the other sample. In short, these types of samples are related to each other. Dependent samples can occur in two scenarios. In one, a group may be measured twice such as in a pretest-posttest 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.

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 delta symbol δ to symbolize the difference between the two samples. Therefore, in our null hypothesis we state that the difference of scores across the two measurements is equal to $0; \delta = 0$ or:

$$H_0 : \delta = \mu_1 - \mu_2$$

Calculating the Pooled Estimate of Population Variance

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, $s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}$.

Where n_1 and n_2 are the sizes of the two samples s^2 is the pooled sample variance, which is computed as $s^2 = \dfrac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$. Often, the top part of this formula is simplified by substituting the symbol SS for the sum of the squared deviations. Therefore, the formula often is expressed by $s^2 = \dfrac{SS_1 + SS_2}{n_1 + n_2 - 2}$

Example: Calculating s^2 Suppose we have two independent samples of student reading scores.

The data are as follows:

TABLE 8.3:

Sample 1	Sample 2
7	12
8	14
10	18
4	13
6	11
	10

From this sample, we can calculate a number of descriptive statistics that will help us solve for the pooled estimate of variance:

TABLE 8.4:

Descriptive Statistic	Sample 1	Sample 2
Number n	5	6
Sum of Observations $\sum x$	35	78
Mean of Observations $\bar{x}$	7	13
Sum of Squared Deviations $\sum_{i=1}^{n}(x_i - \bar{x})^2$	20	40

Using the formula for the pooled estimate of variance, we find that

$$s^2 = 6.67$$

We will use this information to calculate the test statistic needed to evaluate the hypotheses.

Testing Hypotheses with Independent Samples

When testing hypotheses with two independent samples, we follow similar steps as 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:

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

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:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s.e.(\bar{x}_1 - \bar{x}_2)}$$

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

$s.e.(\bar{x}_1 - \bar{x}_2)$ 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.

$$H_0 : \mu_1 = \mu_2$$
$$H_0 : \mu_1 \neq \mu_2$$

This is a two tailed test. For this example, we have two independent samples from the population and have a total of 17 students that we are examining. Since our sample size is so low, we use the $t-$distribution. In this example, we have 15 degrees of freedom (number in the samples minus 2) and with a .05 significance level and the t distribution, we find that our critical values are 2.131 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	68.875
Sum of Squared Deviations $\sum_{i=1}^{n}(x_i - \bar{x})^2$	862.22	1058.88

Therefore:

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = 128.07$$

and the standard error of the difference of the sample means is:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{128.07 \left(\frac{1}{9} + \frac{1}{8} \right)} \approx 5.50$$

Using this information, we can finally solve for the test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s.e.(\bar{x}_1 - \bar{x}_2)} = \frac{(49.44 - 68.875) - (0)}{5.50} \approx -3.53$$

Since -3.53 is less than the critical value of 2.13, we decide to reject the null hypothesis and conclude there is a significant difference in the achievement of the students assigned to different teachers.

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 similar method as described above. We might perform these hypotheses tests in the following scenarios:

- When examining the proportion 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 proportion 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, for small samples we generally use the t distribution and for samples over 30 we generally use the $z-$distribution).

When solving for the test statistic in large samples, we use the formula:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{se(p_1 - p_2)}$$

where:

$\hat{p}_1, \hat{p}_2$ are the observed sample proportions

p_1, p_2 are the population proportions under the null hypothesis

$se(p_1 - p_2)$ is the standard error of the difference between independent proportions

Similar to the standard error of the difference between independent samples, 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 estimate p.

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Now the standard error of the difference is $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

The test statistic is now $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (0)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Example: Suppose that we are interested in finding out which particular city is more 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 that 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 a large sample size we will use the $z-$distribution. At a .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.

$$\hat{p} = \frac{200(.61) + 150(.56)}{350} = .589$$

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{0.589(.411)\left(\frac{1}{200} + \frac{1}{150}\right)} \approx 0.053$$

Therefore, the test statistic is:

$$z = \frac{(0.61 - 0.56) - (0)}{0.053} \approx 0.94$$

Since 0.94 does not exceed the critical value 1.96, the null hypothesis is not rejected. Therefore, we can conclude that the difference in the probabilities could have occurred by chance and that there is no difference in the level of satisfaction between citizens of the two cities.

Testing Hypotheses with Dependent Samples

When testing a hypothesis about two dependent samples, we follow the same process as when testing one random sample or two independent samples:

- State the null and alternative hypotheses.
- Choose the level of significance
- 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 our results.

As mentioned in the section above, our hypothesis for two dependent samples states that there is no difference between the scores across the two samples $H_0 : \delta = \mu_1 - \mu_2 = 0$. We set the criterion for evaluating the hypothesis in the same way that we do with our other examples –by first establishing an alpha level and then finding the critical values by using the $t-$distribution table. Calculating the test statistic for dependent samples is a bit different since we are dealing with two sets of data. The test statistic that we first need calculate is $\bar{d}$, which is the difference in the means of the two samples. Therefore, $\bar{d} = \bar{x}_1 - \bar{x}_2$. We also need to know the standard error of the difference between the two samples. Since our population variance is unknown, we estimate it by first using the formula for the standard deviations 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:

$s_{\bar{d}}^2$ is the sample variance

d is the difference between corresponding pairs within the sample

$\bar{d}$ is the difference between the means of the two samples

n is the number in the sample

s_d is the standard deviation

With the standard deviation, we can calculate the standard error 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:

$$t = \frac{\bar{d} - \delta}{s_d}$$

Example: The 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 hypothesis is that there is no difference between the means of the two samples and our alternative hypothesis is that the two means of the samples are not equal. In other words, we are testing whether or not these two samples are related or:

$$H_0 : \delta = \mu_1 - \mu_2 = 0$$
$$H_a : \delta = \mu_1 - \mu_2 \neq 0$$

The results for the pre-and post-tests are below:

TABLE 8.8:

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 two samples, then the standard error of 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)

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}} = \frac{3.6 - 0}{1.40} \approx 2.57$$

With 8 degrees of freedom (number of observations - 1) and a significance level of .05, we find our critical values to be ± 2.306. Since our test statistic exceeds this critical value, we can reject the null hypothesis that the two samples are equal and conclude that the lesson had an effect on student achievement.

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 assume that the samples are related to each other.

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 formula:

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \text{ with } s^2 = \frac{ss_1 + ss_2}{n_1 + n_2 - 2}$$

We carry out the test on the means of two independent samples in a similar way as the testing of one random sample. However, we use the following formula to calculate the test statistic:

$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s.e.(\bar{x}_1 - \bar{x}_2)}$ with the standard error defined above.

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:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (0)}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \text{ with } \hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

We can also test the likelihood that two dependent samples are related. To calculate the test statistic for two dependent samples, we use the formula:

$$t = \frac{\bar{d} - \delta}{s_d} \text{ with } s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

Review Questions

1. In hypothesis testing, we have scenarios that have both dependent and independent samples. Give an example of an experiment with (1) dependent samples and (2) 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 and one group is given Drug A and the other group is given Drug B and the number of times that they push a lever is recorded. The following results for this test were calculated:

TABLE 8.9:

	Drug A	Drug B
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) Compute the pooled estimate for population variance.

(d) Calculate the estimated standard error for this scenario.

(e) What is the test statistic and at an alpha level of .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 8.10:

Husbands	Wives
16	15
20	18
10	13
15	10
8	12
19	16
14	11
15	12

(a) What would be the hypotheses for this scenario?

(b) Calculate the estimated standard deviation for this scenario.

(c) Compute the standard error of the difference for these samples.

(d) What is the test statistic and at an alpha level of .05 what conclusions would you make about the null hypothesis?

Keywords

Null hypothesis

Alternative hypothesis

One-tailed test

Two-tailed test

$p-$value

Power of a test

Level of significance

Critical region

Type I error

Type II error

α

β

Standard error

Dependent samples

t distribution

CHAPTER **9** **Regression and Correlation**

Chapter Outline

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
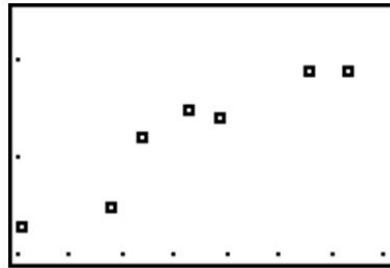
These types of studies are quite common, and we can use the concept of correlation to describe the relationship between the two variables.

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.
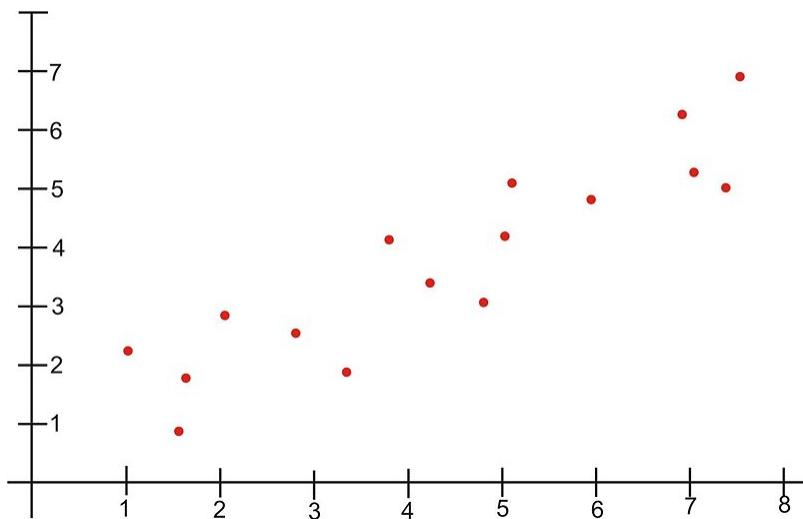
Scatterplots display these bivariate data sets and provide a visual representation of the relationship between variables. In a scatterplot, each point represents a paired measurement of two variables for a specific subject, and each subject is represented by one point on the scatterplot.
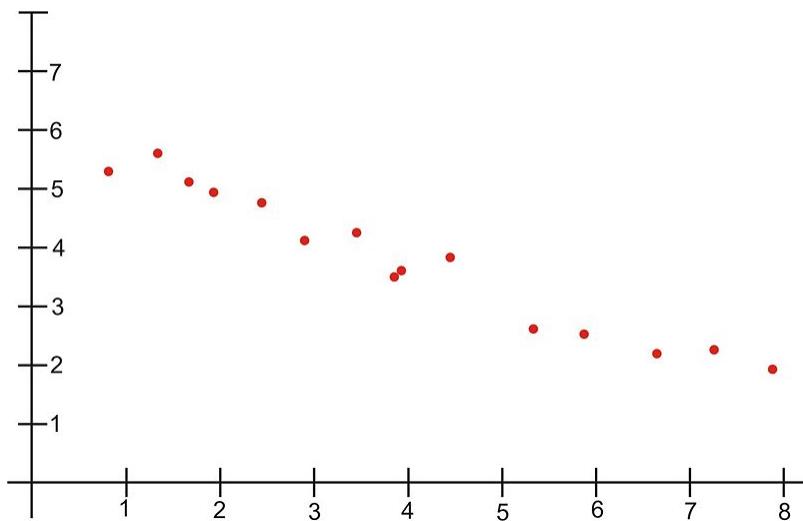
Correlation Patterns in Scatterplot Graphs

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.
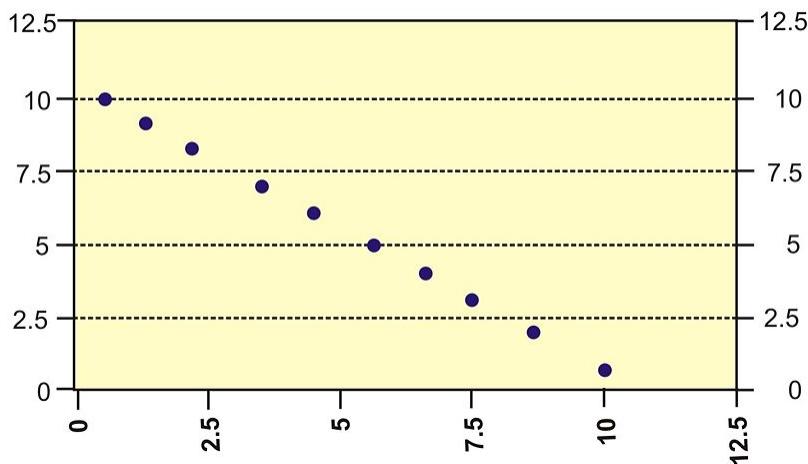
When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a *negative correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low, and vice versa.
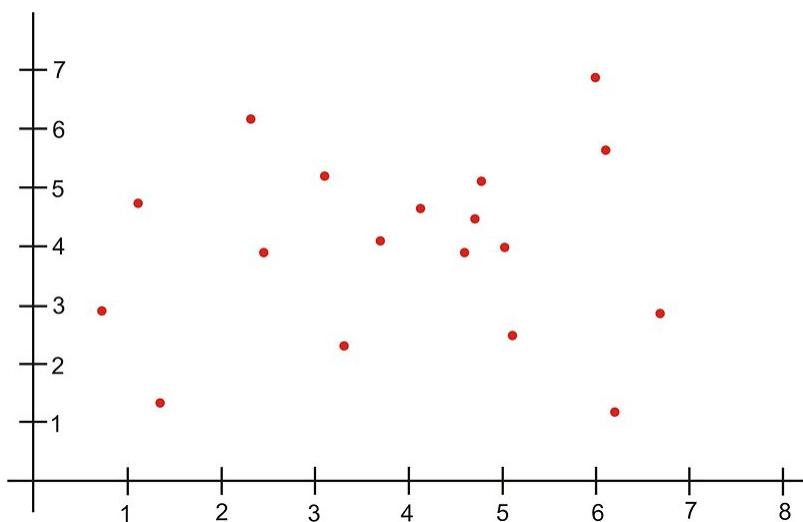
When all the points on a scatterplot lie on a straight line, you have what is called a *perfect correlation* between the two variables (see below).
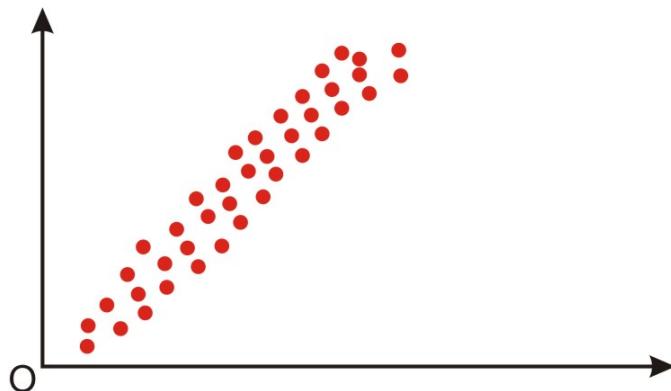
Perfect Negative Correlation
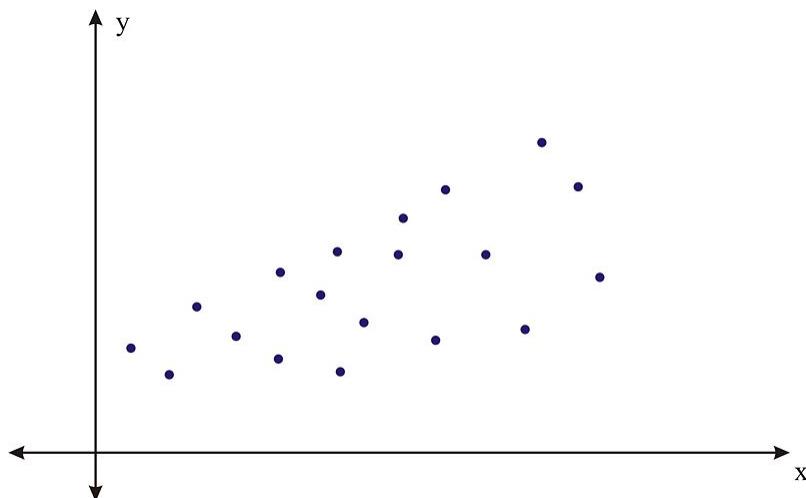
A scatterplot in which the points do not have a linear trend (either positive or negative) is called a *zero correlation* or a *near-zero correlation* (see below).

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:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}$$

Again, this formula is most often used when calculating correlation coefficients from original data. Note that n is used instead of $n-1$, because we are using actual data and not z-scores. 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

TABLE 9.3: (continued)

Student	SAT Score (X)	GPA (Y)	xy	x^2	y^2
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

Applying the formula to these data, we find the following:

$$r_{XY} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}} = \frac{(7)(13262) - (3970)(22.7)}{\sqrt{[(7)(2321400) - 3970^2][(7)(76.01) - 22.7^2]}}$$

$$= \frac{2715}{2864.22} \approx 0.95$$

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.

On the Web

http://tinyurl.com/ylcyh88 Match the graph to its correlation.

http://tinyurl.com/y8vcm5y Guess the correlation.

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html Regression by eye.

The Properties and Common Errors of Correlation
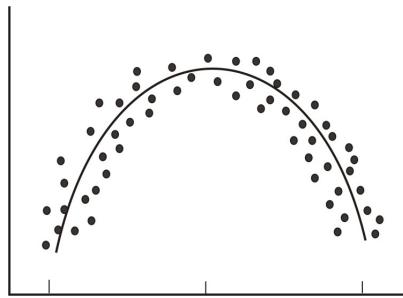
Correlation is a measure of the linear relationship between two variables-it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to 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 lie generally along a straight line.

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: $r_{XY} = \dfrac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]}\sqrt{\left[n\sum y^2 - (\sum y)^2\right]}}$.

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.

Multimedia Links

For an explanation of the correlation coefficient **(13.0)**, see kbower50, The Correlation Coefficient (3:59).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1104

Review Questions

1. Give 2 scenarios or research questions where you would use bivariate data sets.
2. In the space below, draw and label four scatterplot graphs. One should show:

    a. a positive correlation
    b. a negative correlation
    c. a perfect correlation
    d. a zero correlation

3. In the space below, draw and label two scatterplot graphs. One should show:

    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.

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

TABLE 9.5: (continued)

Student	Quiz 1	Quiz 2
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) 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?

9.2 Least-Squares Regression
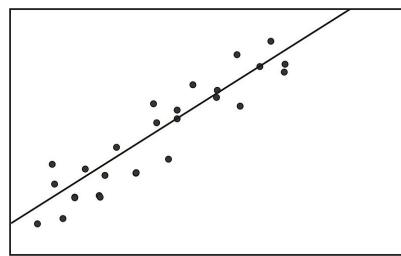
Learning Objectives

- 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).
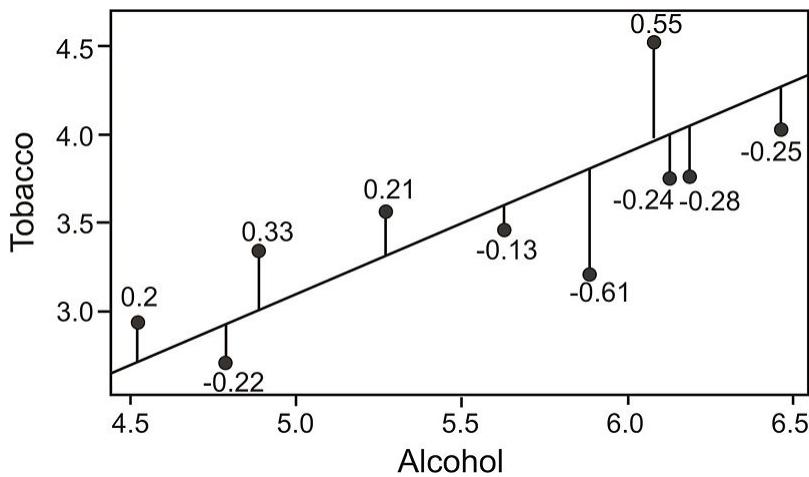
Calculating and Graphing the Regression Line

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 *predictor variable*) to predict the outcome of another (the *outcome variable*, or *criterion 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:

$$Y = bX + a$$

where:

Y is the score that we are trying to predict.

b is the slope of the line.

a is the y-intercept, or the value of Y when the value of X is 0.

To calculate the line itself, we need to find the values for b (the *regression coefficient*) and a (the *regression constant*). The regression coefficient explains the nature of the relationship between the two variables. Essentially, the regression coefficient tells us that a certain change in the predictor variable is associated with a certain change in the outcome, or criterion, variable. For example, if we had a regression coefficient of 10.76, 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 regression coefficient, we also need to calculate the regression constant. The regression constant is also the y-intercept and is the place where the line crosses the y-axis. For example, if we had an equation with a regression constant of 4.58, we would conclude that the regression line crosses the y-axis at 4.58. We use the following formula to calculate the regression constant:

$$a = \frac{\sum y - b\sum x}{n} = \bar{y} - b\bar{x}$$

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.

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
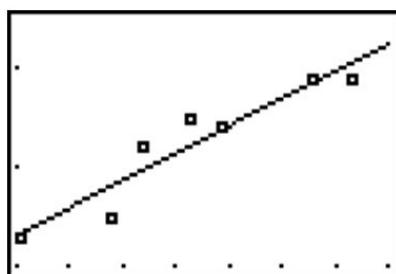
Using these data points, we first calculate the regression coefficient and the regression constant as follows:

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2} = \frac{(7)(13,262) - (3,970)(22.7)}{(7)(2,321,400) - 3,970^2} = \frac{2715}{488900} \approx 0.0056$$

$$a = \frac{\sum y - b\sum x}{n} \approx 0.094$$

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, 2.334) and (500, 2.89). From these pairs of coordinates, we can draw the regression line on the scatterplot.

Predicting Values Using Scatterplot Data

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 a predictor variable, X, into the regression equation and solving the equation for the outcome variable, 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, $Y = 0.0056X + 0.094$.

For example, say that we wanted to predict the GPA for two students, one who had an SAT score of 500 and the other who had an SAT score of 600. To predict the GPA scores for these two students, we would simply plug the two values of the predictor variable into the equation and solve for Y (see below).
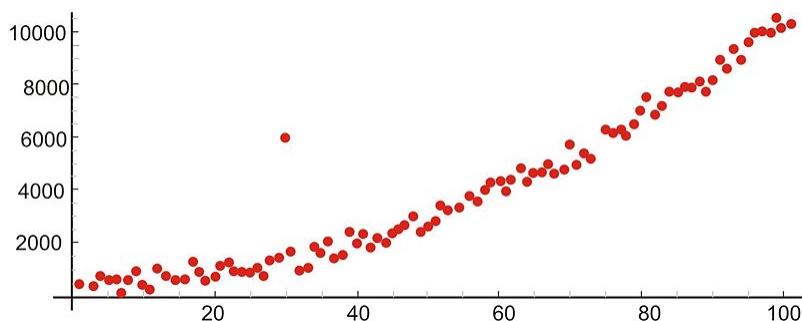
TABLE 9.7: GPA/SAT data, including predicted GPA values from the linear regression.

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). In the regression setting, outliers will be far away from the regression line in the y-direction. 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).
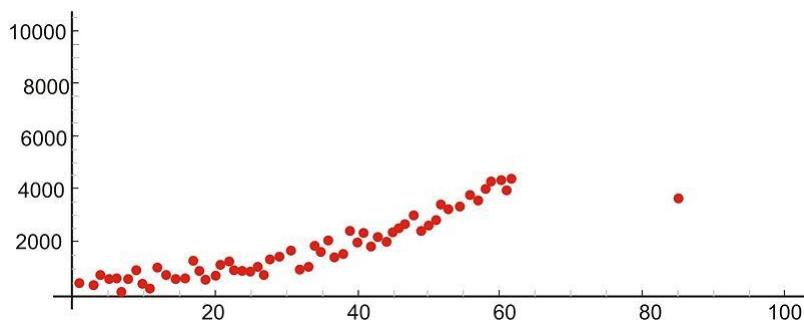
Let's use our example above to illustrate the effect of a single outlier. Say that we have a student who has a high GPA but who suffered from test anxiety the morning of the SAT verbal test and scored a 410. Using our original regression equation, we would expect the student to have a GPA of 2.2. But, 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. For univariate data, we

can use the IQR rule to determine whether or not a point is an outlier. 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.
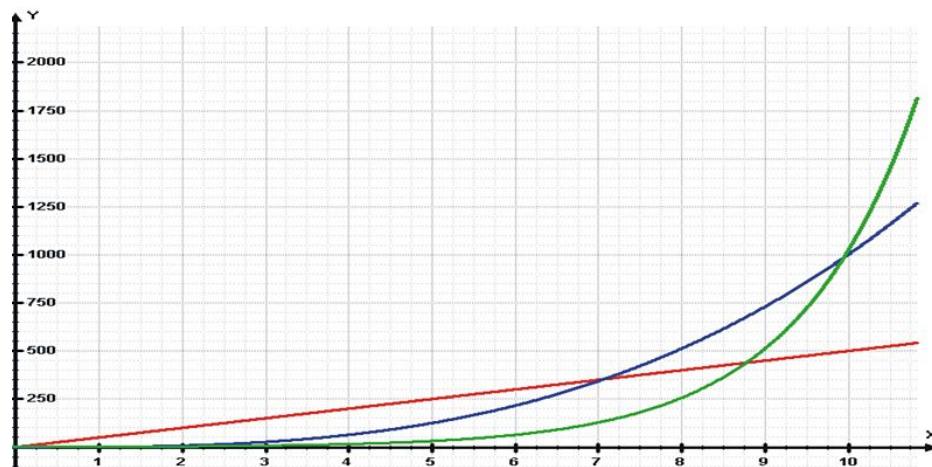
An *influential point* in regression is one whose removal would greatly impact the equation of the regression line. Usually, an influential point will be separated in the x direction from the other observations. It is possible for an outlier to be an influential point. However, there are some influential points that would not be considered outliers. These will not be far from the regression line in the y-direction (a value called a residual, discussed later) so you must look carefully for them. In the following scatterplot, the influential point has approximate coordinates of (85, 35,000).

It is important to determine whether influential points are 1) correct and 2) belong in the population. If they are not correct or do not belong, then they can be removed. If, however, an influential point is determined to indeed belong in the population and be correct, then one should consider whether other data points need to be found with similar x-values to support the data and regression line.

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.

Consider the following exponential relationship, and take the log of both sides as shown:

$$y = ab^x$$
$$\log y = \log(ab^x)$$
$$\log y = \log a + \log b^x$$
$$\log y = \log a + x \log b$$

In this example, *a* and *b* are real numbers (constants), so this is now a linear relationship between the variables *x* and $\log y$.
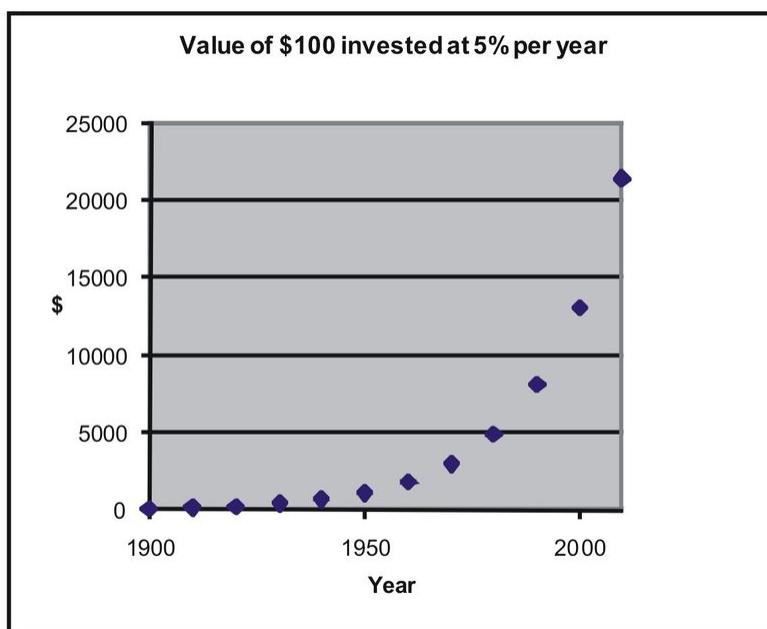
Thus, you can find a least squares line for these variables.

Let's take a look at an example to help clarify this concept. Say that we were interested in making a case for investing and examining how much return on investment one would get on $100 over time. Let's assume that we invested $100 in the year 1900 and that this money accrued 5% interest every year. The table below details how much we would have each decade:

TABLE 9.8: Table of account growth assuming $100 invested in 1900 at 5% annual growth.

Year	Investment with 5% Each Year
1900	100
1910	163
1920	265
1930	432
1940	704
1950	1147
1960	1868
1970	3043
1980	4956
1990	8073
2000	13150
2010	21420

If we graphed these data points, we would see that we have an exponential growth curve.
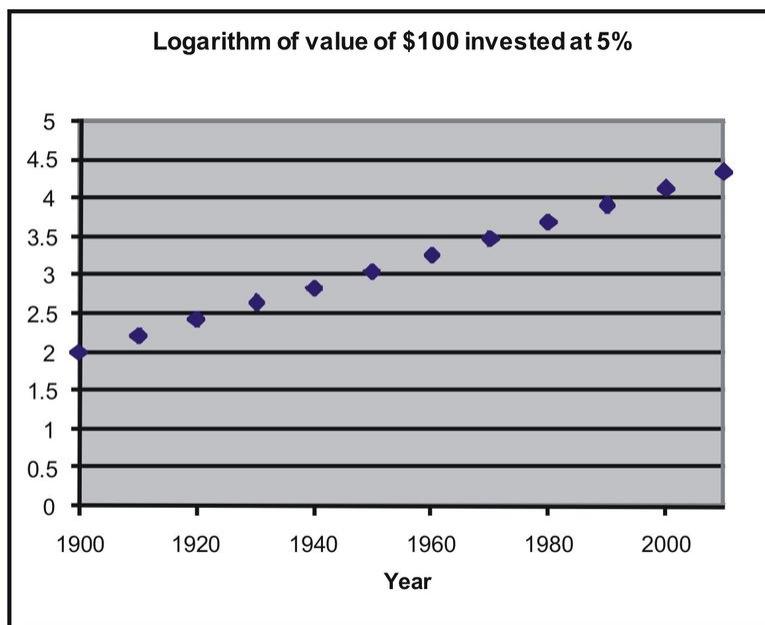
Say that we wanted to fit a linear regression line to these data. First, we would transform these data using logarithmic transformations as follows:

TABLE 9.9: Account growth data and values after a logarithmic transformation.

Year	Investment with 5% Each Year	Log of amount
1900	100	2
1910	163	2.211893
1920	265	2.423786
1930	432	2.635679
1940	704	2.847572
1950	1147	3.059465
1960	1868	3.271358
1970	3043	3.483251
1980	4956	3.695144
1990	8073	3.907037
2000	13150	4.118930
2010	21420	4.330823

If we plotted these transformed data points, we would see that we have a linear relationship as shown below:
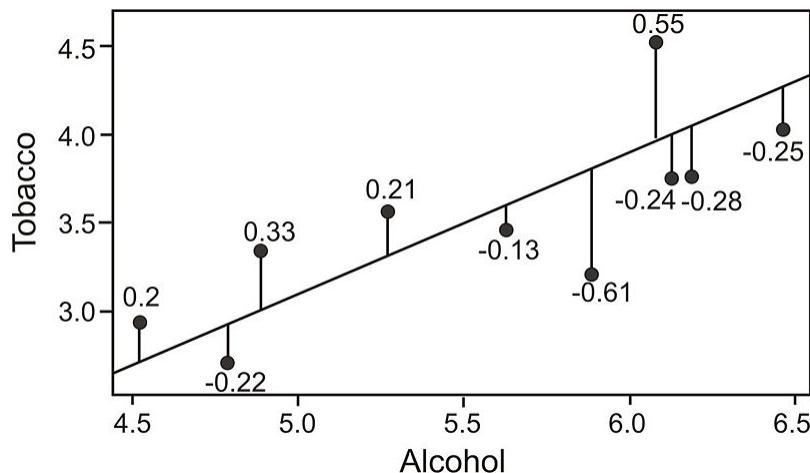
We can now perform a linear regression on (year, log of amount). If you enter the data into the TI-83/84 calculator, press **[STAT]**, go to the **CALC** menu, and use the 'LinReg(ax+b)' command, you find the following relationship:

$$Y = 0.021X - 38.2$$

with X representing year and Y representing log of amount.

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, as 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 $\sum(y-\hat{y})^2$.
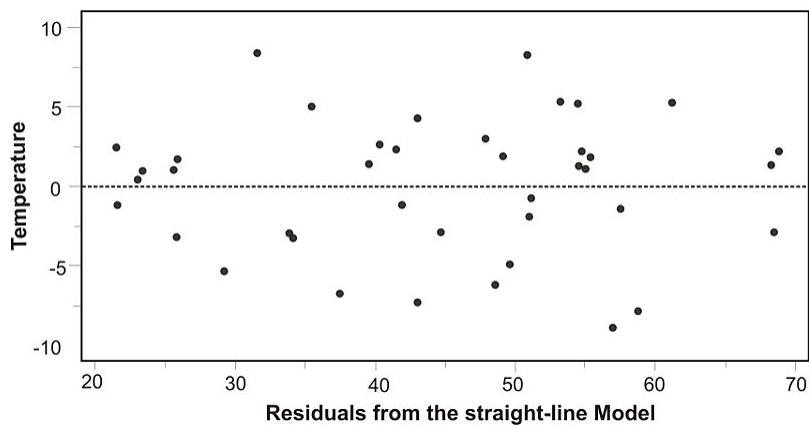
Example: Calculate the residuals for the predicted and the actual GPA's from our sample above.

TABLE 9.10: 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
$\sum(y-\hat{y})^2$					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.
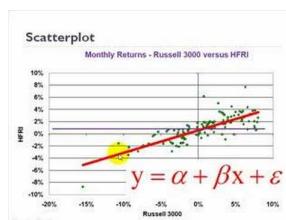
Using this regression line, we are able to use the slope, y-intercept, and the calculated regression coefficient to predict the scores of a variable. The predictions are represented by the variable $\hat{y}$.

When there is an exponential relationship between the variables, we can transform the data by taking the log of the dependent variable to achieve linearity between x and $\log y$. We can then fit a least squares regression line to the transformed data.

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 test for linearity.

Multimedia Links

For an introduction to what a least squares regression line represents **(12.0)**, see bionicturtledotcom, Introduction to Linear Regression (5:15).

> **MEDIA**
> Click image to the left or use the URL below.
> URL: http://www.ck12.org/flx/render/embeddedobject/1105

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 9.11: 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

Student	Exercise Per Week	Memory Test Score
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 regression coefficient 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) Do you think that a data transformation is necessary in order to build an accurate linear regression model? Why or why not?

(i) Calculate the residuals for each of the observations and plot these residuals on a scatterplot.

(j) Examine this scatterplot of the residuals. Is a transformation of the data necessary? Why or why not?

9.3 Inferences about Regression

Learning Objectives
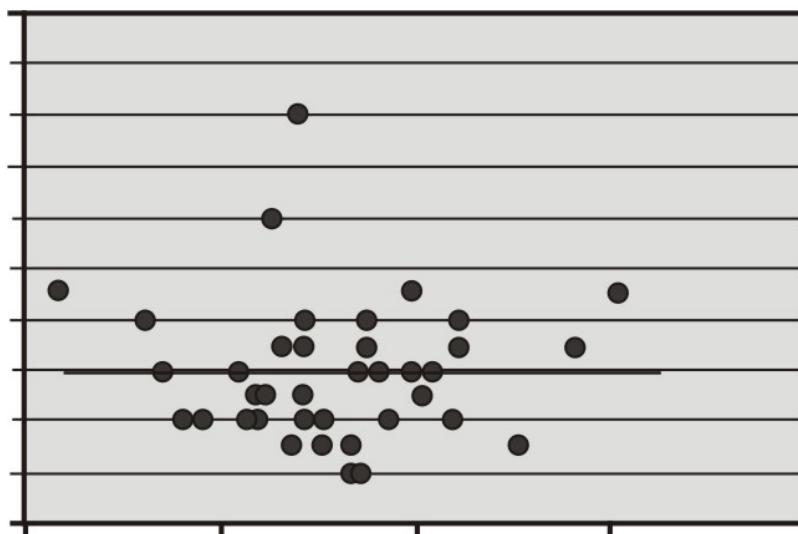
- Make inferences about regression models, including hypothesis testing for linear relationships.
- Make inferences about regression and predicted values, including the construction of confidence intervals.
- Check regression assumptions.

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 predicted values, or the mean of $\hat{Y}$. 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 regression coefficient, β, for the population does not equal zero. Furthermore, it is beneficial to test how strong (or far away) from zero the regression coefficient 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 regression coefficient, β, equals zero. Our alternative hypothesis is that our regression coefficient does not equal zero.

$$H_0: \ \beta = 0$$
$$H_a: \ \beta \neq 0$$

The test statistic for this hypothesis test is calculated as follows:

$$t = \frac{b - \beta}{s_b}$$

$$\text{where} \quad s_b = \frac{s}{\sqrt{\Sigma(x - \bar{x})^2}} = \frac{s}{\sqrt{SS_X}},$$

$$s = \sqrt{\frac{SSE}{n - 2}}, \text{ and}$$

$$SSE = \text{sum of residual error squared}$$

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 $Y = 0.635X + 1.22$, and the standard error of estimate is 0.56. The summary statistics are as follows:

Summary statistics for two foot ball fitness tests.

$n = 24$	$\Sigma xy = 591.50$
$\Sigma x = 118$	$\Sigma y = 104.3$
$\bar{x} = 4.92$	$\bar{y} = 4.35$
$\Sigma x^2 = 704$	$\Sigma y^2 = 510.01$
$SS_X = 123.83$	$SS_Y = 56.74$

Using $\alpha = 0.05$, test the null hypothesis that, in the population, the regression coefficient is zero, or $H_0: \ \beta = 0$.
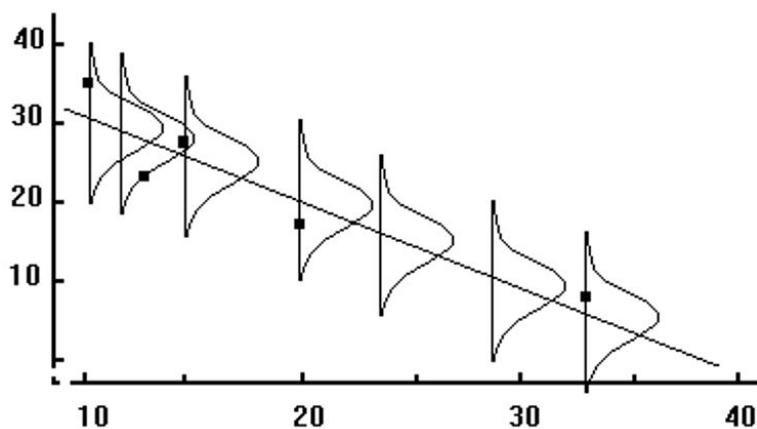
We use the t-distribution to calculate the test statistic and find that the critical values in the t-distribution at 22 degrees of freedom are 2.074 standard scores above and below the mean. Also, the test statistic can be calculated as follows:

$$s_b = \frac{0.56}{\sqrt{123.83}} = 0.05$$

$$t = \frac{0.635 - 0}{0.05} = 12.70$$

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 regression coefficient of 0.635 by chance less than 5% of the time.

Making Inferences about Predicted Scores

As we have mentioned, a regression line makes predictions about variables based on the relationship of the existing data. However, it is important to remember that the regression line simply infers, or estimates, what the value will be. These predictions are never accurate 100% of the time, unless there is a perfect correlation. What this means is that for every predicted value, we have a normal distribution (also known as the *conditional distribution*, since it is conditional on the X value) that describes the likelihood of obtaining other scores that are associated with the value of the predictor variable, X.

If we assume that these distributions are normal, we are able to make inferences about each of the predicted scores. We can ask questions like, "If the predictor variable, X, equals 4, what percentage of the distribution of Y scores will be lower than 3?"

The reason why we would ask questions like this depends on the scenario. Suppose, for example, that we want to know the percentage of students with a 5 on their short physical fitness test that have a predicted score higher than 5 on their long physical fitness test. If the coach is using this predicted score as a cutoff for playing in a varsity match, and this percentage is too low, he may want to consider changing the standards of the test.

To find the percentage of students with scores above or below a certain point, we use the concept of standard scores and the standard normal distribution.

Since we have a certain predicted value for every value of X, the Y values take on the shape of a normal distribution. This distribution has a mean (the regression line) and a standard error, which we found to be equal to 0.56. In short, the conditional distribution is used to determine the percentage of Y values above or below a certain value that are associated with a specific value of X.

Example: Using our example above, if a student scored a 5 on the short test, what is the probability that he or she would have a score of 5 or greater on the long physical fitness test?

From the regression equation $Y = 0.635X + 1.22$, we find that the predicted score when the value of X is 5 is 4.40. Consider the conditional distribution of Y scores when the value of X is 5. Under our assumption, this distribution is normally distributed around the predicted value 4.40 and has a standard error of 0.56.

Therefore, to find the percentage of Y scores of 5 or greater, we use the general formula for a z-score to calculate the following:

$$z = \frac{Y - \hat{Y}}{s} = \frac{5 - 4.40}{0.56} = 1.07$$

Using the z-distribution table, we find that the area to the right of a z-score of 1.07 is 0.1423. Therefore, we can conclude that the proportion of predicted scores of 5 or greater given a score of 5 on the short test is 0.1423, or 14.23%.

310

Prediction Intervals

Similar to hypothesis testing for samples and populations, we can also build a confidence interval around our regression results. This helps us ask questions like "If the predictor variable, X, is equal to a certain value, what are the likely values for Y?" A confidence interval gives us a range of scores that has a certain percent probability of including the score that we are after.

We know that the standard error of the predicted score is smaller when the predicted value is close to the actual value, and it increases as X deviates from the mean. This means that the weaker of a predictor that the regression line is, the larger the standard error of the predicted score will be. The formulas for the standard error of a predicted score and a confidence interval are as follows:

$$s_{\hat{Y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$
$$CI = \hat{Y} \pm t s_{\hat{y}}$$

where:

$\hat{Y}$ is the predicted score.

t is the critical value for $n - 2$ degrees of freedom.

$s_{\hat{Y}}$ is the standard error of the predicted score.

Example: Develop a 95% confidence interval for the predicted score of a student who scores a 4 on the short physical fitness exam.

We calculate the standard error of the predicted score using the formula as follows:

$$s_{\hat{Y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}} = 0.56 \sqrt{1 + \frac{1}{24} + \frac{(4 - 4.92)^2}{123.83}} = 0.57$$

Using the general formula for a confidence interval, we can calculate the answer as shown:

$$CI = \hat{Y} \pm t s_{\hat{Y}}$$
$$CI_{0.95} = 3.76 \pm (2.074)(0.57)$$
$$CI_{0.95} = 3.76 \pm 1.18$$
$$CI_{0.95} = (2.58, 4.94)$$

Therefore, we can say that we are 95% confident that given a student's short physical fitness test score, X, of 4, the interval from 2.58 to 4.94 will contain the student's score for the longer physical fitness test.

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 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 and test for significance.

For each predicted value, we have a normal distribution (also known as the conditional distribution, since it is conditional on the value of X) that describes the likelihood of obtaining other scores that are associated with the value of the predictor variable, X. We can use these distributions and the concept of standardized scores to make predictions about probability.

We can also build confidence intervals around the predicted values to give us a better idea about the ranges likely to contain a certain score.

We make several assumptions when dealing with a linear regression model including:

At each value of X, there is a distribution of Y.

A regression line is a good fit to the data. There is homoscedasticity, and the observations are independent.

Review Questions

1. A college counselor is putting on a presentation about the financial benefits of further education and takes a random sample of 120 parents. Each parent was asked a number of questions, including the number of years of education that he or she has (including college) and his or her yearly income (recorded in the thousands of dollars). The summary data for this survey are as follows:

$$n = 120 \quad r = 0.67 \quad \sum x = 1,782 \quad \sum y = 1,854 \quad s_X = 3.6 \quad s_Y = 4.2 \quad s_{XY} = 3.12 \quad SS_X = 1542$$

(a) What is the predictor variable? What is your reasoning behind this decision?

(b) Do you think that these two variables (income and level of formal education) are correlated? Is so, please describe the nature of their relationship.

(c) What would be the regression equation for predicting income, Y, from the level of education, X?

(d) Using this regression equation, predict the income for a person with 2 years of college (13.5 years of formal education).

(e) Test the null hypothesis that in the population, the regression coefficient for this scenario is zero.

- First develop the null and alternative hypotheses.
- Set the critical value to $\alpha = 0.05$.
- Compute the test statistic.
- Make a decision regarding the null hypothesis.

(f) For those parents with 15 years of formal education, what is the percentage who will have an annual income greater than \$18,500?

(g) For those parents with 12 years of formal education, what is the percentage who will have an annual income greater than \$18,500?

(h) Develop a 95% confidence interval for the predicted annual income when a parent indicates that he or she has a college degree (i.e., 16 years of formal education).

(i) If you were the college counselor, what would you say in the presentation to the parents and students about the relationship between further education and salary? Would you encourage students to further their education based on these analyses? Why or why not?

9.4 Multiple Regression

Learning Objectives

- Understand a multiple regression equation and the coefficients of determination for correlation of three or more variables.
- Calculate a multiple regression equation using technological tools.
- Calculate the standard error of a coefficient, test a coefficient for significance to evaluate a hypothesis, and calculate the confidence interval for a coefficient using technological tools.

Introduction

In the previous sections, we learned a bit about examining the relationship between two variables by calculating the correlation coefficient and the linear regression line. But, as we all know, often times we work with more than two variables. For example, what happens if we want to examine the impact that class size and number of faculty members have on a university's ranking. Since we are taking multiple variables into account, the linear regression model just won't work. In multiple linear regression, scores for one variable are predicted (in this example, a university's ranking) using multiple predictor variables (class size and number of faculty members).

Another common use of multiple regression models is in the estimation of the selling price of a home. There are a number of variables that go into determining how much a particular house will cost, including the square footage, the number of bedrooms, the number of bathrooms, the age of the house, the neighborhood, and so on. Analysts use multiple regression to estimate the selling price in relation to all of these different types of variables.
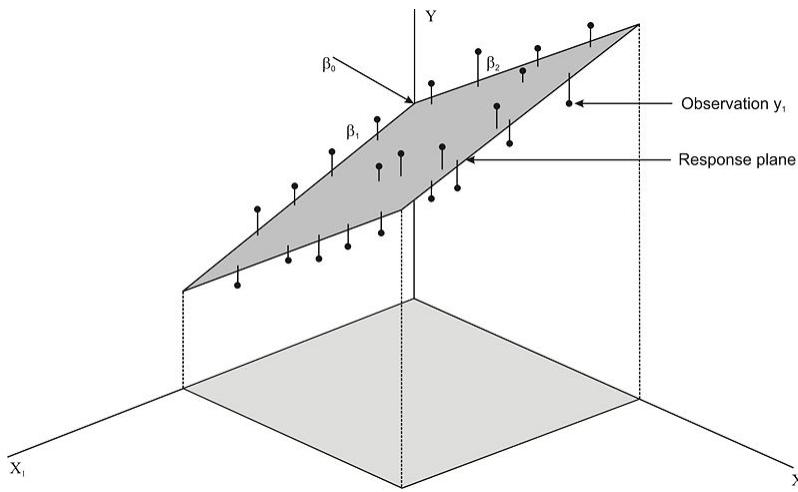
In this section, we will examine the components of a multiple regression equation, calculate an equation using technological tools, and use this equation to test for significance in order to evaluate a hypothesis.

Understanding a Multiple Regression Equation

If we were to try to draw a *multiple regression* model, it would be a bit more difficult than drawing a model for linear regression. Let's say that we have two predictor variables, X_1 and X_2, that are predicting the desired variable, Y. The regression equation would be as follows:

$$\hat{Y} = b_1 X_1 + b_2 X_2 + a$$

When there are two predictor variables, the scores must be plotted in three dimensions (see figure below). When there are more than two predictor variables, we would continue to plot these in multiple dimensions. Regardless of how many predictor variables there are, we still use the least squares method to try to minimize the distance between the actual and predicted values.

When predicting values using multiple regression, we first use the standard score form of the regression equation, which is shown below:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

where:

$\hat{Y}$ is the predicted variable, or criterion variable.

β_i is the ith regression coefficient.

X_i is the ith predictor variable.

To solve for the regression and constant coefficients, we need to determine multiple correlation coefficients, r, and coefficients of determination, also known as proportions of shared variance, r^2. In the linear regression model, we measured r^2 by adding the squares of the distances from the actual points to the points predicted by the regression line. So what does r^2 look like in the multiple regression model? Let's take a look at the figure above. Essentially, like in the linear regression model, the theory behind the computation of a multiple regression equation is to minimize the sum of the squared deviations from the observations to the regression plane.

In most situations, we use a computer to calculate the multiple regression equation and determine the coefficients in this equation. We can also do multiple regression on a TI-83/84 calculator. (This program can be downloaded.)

Technology Note: Multiple Regression Analysis on the TI-83/84 Calculator

http://www.wku.edu/~david.neal/manual/ti83.html

Download a program for multiple regression analysis on the TI-83/84 calculator by first clicking on the link above.

It is helpful to explain the calculations that go into a multiple regression equation so we can get a better understanding of how this formula works.

After we find the correlation values, r, between the variables, we can use the following formulas to determine the regression coefficients for the predictor variables, X_1 and X_2:

$$\beta_1 = \frac{r_{Y1} - (r_{Y2})(r_{12})}{1 - r_{12}^2}$$
$$\beta_2 = \frac{r_{Y2} - (r_{Y1})(r_{12})}{1 - r_{12}^2}$$

where:

β_1 is the correlation coefficient for X_1.

β_2 is the correlation coefficient for X_2.

r_{Y1} is the correlation between the criterion variable, Y, and the first predictor variable, X_1.

r_{Y2} is the correlation between the criterion variable, Y, and the second predictor variable, X_2.

r_{12} is the correlation between the two predictor variables, X_1 and X_2.

After solving for the beta coefficients, we can then compute the b coefficients by using the following formulas:

$$b_1 = \beta_1 \left(\frac{s_Y}{s_1} \right)$$
$$b_2 = \beta_2 \left(\frac{s_Y}{s_2} \right)$$

where:

s_Y is the standard deviation of the criterion variable, Y.

s_1 is the standard deviation of the particular predictor variable (1 for the first predictor variable, 2 for the second, and so on).

After solving for the regression coefficients, we can finally solve for the regression constant by using the formula shown below, where k is the number of predictor variables:

$$a = \bar{y} - \sum_{i=1}^{k} b_i \bar{x}_i$$

Again, since these formulas and calculations are extremely tedious to complete by hand, we usually use a computer or a TI-83/84 calculator to solve for the coefficients in a multiple regression equation.

Calculating a Multiple Regression Equation using Technological Tools

As mentioned, there are a variety of technological tools available to calculate the coefficients in a multiple regression equation. When using a computer, there are several programs that help us calculate the multiple regression equation, including Microsoft Excel, the Statistical Analysis Software (SAS), and the Statistical Package for the Social Sciences (SPSS). Each of these programs allows the user to calculate the multiple regression equation and provides summary statistics for each of the models.

For the purposes of this lesson, we will synthesize summary tables produced by Microsoft Excel to solve problems with multiple regression equations. While the summary tables produced by the different technological tools differ slightly in format, they all provide us with the information needed to build a multiple regression equation, conduct hypothesis tests, and construct confidence intervals. Let's take a look at an example of a summary statistics table so we get a better idea of how we can use technological tools to build multiple regression equations.

Example: Suppose we want to predict the amount of water consumed by football players during summer practices. The football coach notices that the water consumption tends to be influenced by the time that the players are on the field and by the temperature. He measures the average water consumption, temperature, and practice time for seven practices and records the following data:

TABLE 9.12:

Temperature (degrees F)	Practice Time (hrs)	H_2O Consumption (in ounces)
75	1.85	16
83	1.25	20
85	1.5	25
85	1.75	27
92	1.15	32
97	1.75	48
99	1.6	48

Figure: Water consumption by football players compared to practice time and temperature.

Technology Note: Using Excel for Multiple Regression

- Copy and paste the table into an empty Excel worksheet.
- Click the Data choice on the toolbar, then select 'Data Analysis,' and then choose 'Regression' from the list that appears (Note, if Data Analysis does not appear as a choice on your Data page need to follow the add-in instructions below).
- Place the cursor in the 'Input Y range' field and select the third column.
- Place the cursor in the 'Input X range' field and select the first and second columns.
- Place the cursor in the 'Output Range' field and click somewhere in a blank cell below and to the left of the table.
- Click 'Labels' so that the names of the predictor variables will be displayed in the table.
- Click 'OK', and the results shown below will be displayed.

Note: In Excel 2007, to add **Data Analysis** to your **Data** page, perform the following functions. Click the **Microsoft Office Button** in the upper left, then click on **Excel Options**. Click on **Add-ins**, then highlight the **Analysis ToolPak**, click **Go**, make sure the **Analysis ToolPak box** is checked off, and then click **OK**. The **Data Analysis** choice should now appear on your Excel **Data** page. Follow the remaining instructions above.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.996822
R Square	0.993654
Adjusted R Square	0.990481
Standard Error	1.244877
Observations	7

TABLE 9.13:

	Df	SS	MS	F	Significance F
Regression	2	970.6583	485.3291	313.1723	4.03E-05
Residual	4	6.198878	1.549719		
Total	6	976.8571			

TABLE 9.14:

	Coefficients	Standard Error	t Stat	P- value	Lower 95%	Upper 95%
Intercept	-121.655	6.540348	-18.6007	4.92e-05	-139.814	-103.496
Temperature	1.512364	0.060771	24.88626	1.55E-05	1.343636	1.681092
Practice Time	12.53168	1.93302	6.482954	0.002918	7.164746	17.89862

In this example, we have a number of summary statistics that give us information about the regression equation. As you can see from the results above, we have the regression coefficient and standard error for each variable, as well as the value of r^2. We can take all of the regression coefficients and put them together to make our equation.

Using the results above, our regression equation would be $\hat{Y} = -121.66 + 1.51(\text{Temperature}) + 12.53(\text{Practice Time})$.

Each of the regression coefficients tells us something about the relationship between the predictor variable and the predicted outcome. The temperature coefficient of 1.51 tells us that for every 1.0-degree increase in temperature, we predict there to be an increase of 1.5 ounces of water consumed, if we hold the practice time constant. Similarly, we find that with every one-hour increase in practice time, we predict players will consume an additional 12.53 ounces of water, if we hold the temperature constant. That equates to about 2.1 extra ounces of water for every 10 minutes increase in practice time.

With a value of 0.99 for r^2, we can conclude that approximately 99% of the variance in the outcome variable, Y, can be explained by the variance in the combined predictor variables. With a value of 0.99 for r^2, we can conclude that almost all of the variance in water consumption is attributed to the variance in temperature and practice time.

Testing for Significance to Evaluate a Hypothesis, the Standard Error of a Coefficient, and Constructing Confidence Intervals

When we perform multiple regression analysis, we are essentially trying to determine if our predictor variables explain the variation in the outcome variable, Y. When we put together our final equation, we are looking at whether or not the variables explain most of the variation, r^2, and if this value of r^2 is statistically significant. We can use technological tools to conduct a hypothesis test, testing the significance of this value of r^2, and construct confidence intervals around these results.

Hypothesis Testing

When we conduct a hypothesis test, we test the null hypothesis that the multiple r-value in the population equals zero, or $H_0 : r_{\text{pop}} = 0$. Under this scenario, the predicted values, or fitted values, would all be very close to the mean, and the deviations, $\hat{Y} - \bar{Y}$, and the sum of the squares would be close to 0. Therefore, we want to calculate a test statistic (in this case, the F-*statistic*) that measures the correlation between the predictor variables. If this test statistic is beyond the critical values and the null hypothesis is rejected, we can conclude that there is a nonzero relationship between the criterion variable, Y, and the predictor variables. When we reject the null hypothesis, we can say something like, "The probability that r^2 having the value obtained would have occurred by chance if the null hypothesis were true is less than 0.05 (or whatever the significance level happens to be)." As mentioned, we can use computer programs to determine the F-statistic and its significance.

Let's take a look at the example above and interpret the F-statistic. We see that we have a very high value of r^2 of 0.99, which means that almost all of the variance in the outcome variable (water consumption) can be explained by the predictor variables (practice time and temperature). Our ANOVA (ANalysis Of VAriance) table tells us that we have a calculated F-statistic of 313.17, which has an associated probability value of 4.03e-05. This means that the probability that 99 percent of the variance would have occurred by chance if the null hypothesis were true (i.e., none of the variance was explained) is 0.0000403. In other words, it is highly unlikely that this large level of variance was by chance. F-distributions will be discussed in greater detail in a later chapter.

Standard Error of a Coefficient and Testing for Significance

In addition to performing a test to assess the probability of the regression line occurring by chance, we can also test the significance of individual coefficients. This is helpful in determining whether or not the variable significantly contributes to the regression. For example, if we find that a variable does not significantly contribute to the regression, we may choose not to include it in the final regression equation. Again, we can use computer programs to determine the standard error, the test statistic, and its level of significance.

Example: Looking at our example above, we see that Excel has calculated the standard error and the test statistic (in this case, the *t*-statistic) for each of the predictor variables. We see that temperature has a *t*-statistic of 24.88 and a corresponding *P*-value of 1.55e-05. We also see that practice time has a *t*-statistic of 6.48 and a corresponding *P*-value of 0.002918. For this situation, we will set α equal to 0.05. Since the *P*-values for both variables are less than $\alpha = 0.05$, we can determine that both of these variables significantly contribute to the variance of the outcome variable and should be included in the regression equation.

Calculating the Confidence Interval for a Coefficient

We can also use technological tools to build a confidence interval around our regression coefficients. Remember, earlier in the chapter we calculated confidence intervals around certain values in linear regression models. However, this concept is a bit different when we work with multiple regression models.

For a predictor variable in multiple regression, the confidence interval is based on a *t*-test and is the range around the observed sample regression coefficient within which we can be 95% (or any other predetermined level) confident that the real regression coefficient for the population lies. In this example, we can say that we are 95% confident that the population regression coefficient for temperature is between 1.34 (the Lower 95% entry) and 1.68 (the Upper 95% entry). In addition, we are 95% confident that the population regression coefficient for practice time is between 7.16 and 17.90.

Lesson Summary

In multiple linear regression, scores for the criterion variable are predicted using multiple predictor variables. The regression equation we use for two predictor variables, X_1 and X_2, is as follows:

$$\hat{Y} = b_1 X_1 + b_2 X_2 + a$$

When calculating the different parts of the multiple regression equation, we can use a number of computer programs, such as Microsoft Excel, SPSS, and SAS.

These programs calculate the multiple regression coefficients, the combined value of r^2, and the confidence intervals for the regression coefficients.

On the Web

www.wku.edu/~david.neal/web1.html

Manuals by a professor at Western Kentucky University for use in statistics, plus TI-83/84 programs for multiple regression that are available for download.

http://education.ti.com/educationportal/activityexchange/activity_list.do

Texas Instrument Website that includes supplemental activities and practice problems using the TI-83 calculator.

Review Questions

1. A lead English teacher is trying to determine the relationship between three tests given throughout the semester and the final exam. She decides to conduct a mini-study on this relationship and collects the test data (scores for Test 1, Test 2, Test 3, and the final exam) for 50 students in freshman English. She enters these data into Microsoft Excel and arrives at the following summary statistics:

Multiple R	0.6859
R Square	0.4707
Adjusted R Square	0.4369
Standard Error	7.5718
Observations	50

TABLE 9.15: ANOVA

	Df	SS	MS	F	Significance F
Regression	3	2342.7228	780.9076	13.621	0.0000
Residual	46	2637.2772	57.3321		
Total	49	4980.0000			

TABLE 9.16:

	Coefficients	**Standard Error**	t **Stat**	P **-value**
Intercept	10.7592	7.6268		
Test 1	0.0506	0.1720	0.2941	0.7700
Test 2	0.5560	0.1431	3.885	0.0003
Test 3	0.2128	0.1782	1.194	0.2387

(a) How many predictor variables are there in this scenario? What are the names of these predictor variables?

(b) What does the regression coefficient for Test 2 tell us?

(c) What is the regression model for this analysis?

(d) What is the value of r^2, and what does it indicate?

(e) Determine whether the multiple r-value is statistically significant.

(f) Which of the predictor variables are statistically significant? What is the reasoning behind this decision?

(g) Given this information, would you include all three predictor variables in the multiple regression model? Why or why not?

Keywords

Bivariate data

Coefficient of determination

Conditional distribution

Correlation

Correlation coefficient

Criterion variable

Curvilinear relationship

e

F-statistic

Homoscedasticity

Least squares line

Line of best fit

Linear regression

Linear regression line

Magnitude

Method of least squares

Multiple regression

Near-zero correlation

Negative correlation

Outcome variable

Outlier

Pearson product-moment correlation coefficient

Perfect correlation

Positive correlation

Predictor variable

r

r^2

Regression coefficient

Regression constant

Residual values

Scatterplots

Transformation

Zero correlation

CHAPTER **10**

Chi-Square

Chapter Outline

10.1 The Goodness-of-Fit Test

Learning Objectives

- Understand the difference between the chi-square distribution and Student's *t*-distribution.
- Identify the conditions which must be satisfied when using the chi-square test.
- Understand the features of experiments that allow goodness-of-fit tests to be used.
- 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.

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?

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	Expected Frequency
Salad	21	25

TABLE 10.1: (continued)

Type of Lunch	Observed Frequency	Expected Frequency
Sub Sandwich	29	25
Daily Special	14	25
Brought Own Lunch	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, 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.

Multimedia Links

For a discussion on *P*-value and an example of a chi-square goodness of fit test **(7.0)(14.0)(18.0)(19.0)**, see APUS 07, Example of a Chi-Square Goodness-of-Fit Test (8:45).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1055

Follow this link to a table of chi-square values: http://tinyurl.com/3ypvj2h

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 two 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

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 100 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	53
Play	18
Science Museum	29

(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) What would be the null hypothesis for the situation above?

(i) There is no preference between the types of field trips that students prefer.

(ii) There is a preference between the types of field trips that students prefer.

(d) What would be the chi-square statistic for the research question above?

(e) If the estimated chi-square level of significance was 5.99, would you reject or fail to reject the null hypothesis?

On the Web

http://onlinestatbook.com/stat_sim/chisq_theor/index.html Explore what happens when you are using the chi-square statistic when the underlying population from which you are sampling does not follow a normal distribution.

10.2 Test of Independence

Learning Objectives

- Understand how to draw data needed to perform calculations when running the chi-square test from contingency tables.
- Run the test of independence to determine whether two variables are independent or not.
- Use 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 $76 \bullet \frac{84}{138} = 46.26$.

The expected frequency for female Republican outcome is $76 \bullet \frac{54}{138} = 29.74$.

The expected frequency for male Democratic outcome is $62 \bullet \frac{84}{138} = 37.74$.

The expected frequency for male Republican outcome is $62 \bullet \frac{54}{138} = 24.26$.

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.

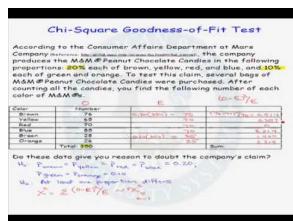O is the observed frequency value for each event.

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 Obs. Freq.	Democratic Exp. Freq.	Democratic $\frac{(O-E)^2}{E}$	Republican Obs. Freq.	Republican Exp. Freq.	Republican $\frac{(O-E)^2}{E}$
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 from the number of Columns ("C") and the number of Rows ("R") as follows:

$$df = (C-1)(R-1)$$
$$= (2-1)(2-1) = 1$$

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.

On the Web

http://tinyurl.com/39lhc3y A chi-square applet demonstrating the test of independence.

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.

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

$H_0 : O = E$ (The probabilities are the same for both dice.)

Alternative Hypothesis

$H_a : O \neq E$ (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 parenthesis) 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 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:
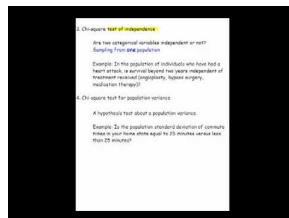
$$\text{Expected Frequency} = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$$

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 equal with respect to certain characteristics.

Multimedia Links

For a discussion of the four different scenarios for use of the chi-square test **(19.0)**, see American Public University, Test Requiring the Chi-Square Distribution (4:13).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1056
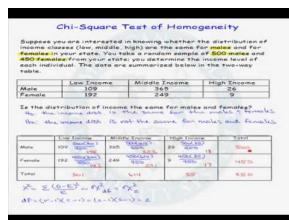
For an example of a chi-square test for homogenity **(19.0)**, see APUS07, Example of a Chi-Square Test of Homogenity (7:57).

MEDIA
Click image to the left or use the URL below.
URL: http://www.ck12.org/flx/render/embeddedobject/1057

For an example of a chi-square test for independence with the TI-83/84 Calculator **(19.0)**, see APUS07, Example of a Chi-Square Test of Independence Using a Calculator (3:29).

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 $= \dfrac{\text{(Row Total)(Column Total)}}{\text{Total Number of Observations}}$
   b. Expected Frequency $= \dfrac{\text{(Total Observations)(Column Total)}}{\text{Row Total}}$
   c. Expected Frequency $= \dfrac{\text{(Total Observations)(Row Total)}}{\text{Column Total}}$
4. Use the table below to answer the following review questions.

TABLE 10.9: Research Question: Are females at UC Berkeley more likely to study abroad than males?

	Studied Abroad	Did Not Study Abroad
Females	322	460
Males	128	152

(a) What is the total number of females in the sample?

450

280

612

782

(b) What is the total number of observations in the sample?

782

533

1,062

612

(c) What is the expected frequency for the number of males who did not study abroad?

161

208

111

129

(d) How many degrees of freedom are in this example?

1

2

3

4

(e) True or False: Our null hypothesis would be that females are as likely as males to study abroad.

(f) What is the chi-square statistic for this example?

1.60

2.45

3.32

3.98

5. If the chi-square critical value at 0.05 and 1 degree of freedom is 3.81, and we have a calculated chi-square statistic of 2.22, we would:

   a. reject the null hypothesis
   b. fail to reject the null hypothesis

6. True or False: We use the test of homogeneity to evaluate the equality of several samples of certain variables.
7. The test of homogeneity is carried out the exact same way as:

   a. the goodness-of-fit test
   b. the test of independence

10.3 Testing One Variance

Learning Objectives

- Test a hypothesis about a single variance using the chi-square distribution.
- Calculate a confidence interval for a population variance based on a sample standard deviation.

Introduction

In the previous lesson, we learned how the chi-square test can help us assess the relationships between two variables. In addition to assessing these relationships, the chi-square test can also help us test hypotheses surrounding variance, which is the measure of the variation, or scattering, of scores in a distribution. There are several different tests that we can use to assess the variance of a sample. The most common tests used to assess variance are the chi-square test for one variance, the *F*-test, and the Analysis of Variance (ANOVA). Both the chi-square test and the *F*-test are extremely sensitive to non-normality (or when the populations do not have a normal distribution), so the ANOVA test is used most often for this analysis. However, in this section, we will examine in greater detail the testing of a single variance using the chi-square test.

Testing a Single Variance Hypothesis Using the Chi-Square Test

Suppose that we want to test two samples to determine if they belong to the same population. The test of variance between samples is used quite frequently in the manufacturing of food, parts, and medications, since it is necessary for individual products of each of these types to be very similar in size and chemical make-up. This test is called the *test for one variance*.

To perform the test for one variance using the chi-square distribution, we need several pieces of information. First, as mentioned, we should check to make sure that the population has a normal distribution. Next, we need to determine the number of observations in the sample. The remaining pieces of information that we need are the standard deviation and the hypothetical population variance. For the purposes of this exercise, we will assume that we will be provided with the standard deviation and the population variance.

Using these key pieces of information, we use the following formula to calculate the chi-square value to test a hypothesis surrounding single variance:

$$\chi^2 = \frac{df(s^2)}{\sigma^2}$$

where:

χ^2 is the chi-square statistical value.

$df = n - 1$, where n is the size of the sample.

s^2 is the sample variance.

σ^2 is the population variance.

We want to test the hypothesis that the sample comes from a population with a variance greater than the observed variance. Let's take a look at an example to help clarify.

Example: Suppose we have a sample of 41 female gymnasts from Mission High School. We want to know if their heights are truly a random sample of the general high school population with respect to variance. We know from a previous study that the standard deviation of the heights of high school women is 2.2.

To test this question, we first need to generate null and alternative hypotheses. Our null hypothesis states that the sample comes from a population that has a variance of less than or equal to 4.84 (σ^2 is the square of the standard deviation).

Null Hypothesis

$H_0 : \sigma^2 \leq 4.84$ (The variance of the female gymnasts is less than or equal to that of the general female high school population.)

Alternative Hypothesis

$H_a : \sigma^2 > 4.84$ (The variance of the female gymnasts is greater than that of the general female high school population.)

Using the sample of the 41 gymnasts, we compute the standard deviation and find it to be $s = 1.2$. Using the information from above, we calculate our chi-square value and find the following:

$$\chi^2 = \frac{(40)(1.2^2)}{4.84} = 11.9$$

Therefore, since 11.9 is less than 55.758 (the value from the chi-square table given an alpha level of 0.05 and 40 degrees of freedom), we fail to reject the null hypothesis and, therefore, cannot conclude that the female gymnasts have a significantly higher variance in height than the general female high school population.

Calculating a Confidence Interval for a Population Variance

Once we know how to test a hypothesis about a single variance, calculating a confidence interval for a population variance is relatively easy. Again, it is important to remember that this test is dependent on the normality of the population. For non-normal populations, it is best to use the *ANOVA test*, which we will cover in greater detail in another lesson. To construct a confidence interval for the population variance, we need three pieces of information: the number of observations in the sample, the variance of the sample, and the desired confidence interval. With the desired confidence interval, α (most often this is set at 0.10 to reflect a 90% confidence interval or at 0.05 to reflect a 95% confidence interval), we can construct the upper and lower limits around the significance level.

Example: We randomly select 30 containers of Coca Cola and measure the amount of sugar in each container. Using the formula that we learned earlier, we calculate the variance of the sample to be 5.20. Find a 90% confidence interval for the true variance. In other words, assuming that the sample comes from a normal population, what is the range of the population variance?

To construct this 90% confidence interval, we first need to determine our upper and lower limits. The formula to construct this confidence interval and calculate the population variance, σ^2, is as follows:

$$\frac{dfs^2}{\chi^2_{0.05}} \leq \sigma^2 \leq \frac{dfs^2}{\chi^2_{0.95}}$$

Using our standard chi-square distribution table (http://tinyurl.com/3ypvj2h), we can look up the critical χ^2 values for 0.05 and 0.95 at 29 degrees of freedom. According to the χ^2 distribution table, we find that $\chi^2_{0.05} = 42.557$ and

that $\chi^2_{0.95} = 17.708$. Since we know the number of observations and the standard deviation for this sample, we can then solve for σ^2 as shown below:

$$\frac{dfs^2}{42.557} \leq \sigma^2 \leq \frac{dfs^2}{17.708}$$
$$\frac{150.80}{42.557} \leq \sigma^2 \leq \frac{150.80}{17.708}$$
$$3.54 \leq \sigma^2 \leq 8.52$$

In other words, we are 90% confident that the variance of the population from which this sample was taken is between 3.54 and 8.52.

Lesson Summary

We can also use the chi-square distribution to test hypotheses about population variance. Variance is the measure of the variation, or scattering, of scores in a distribution, and we often use this test to assess the likelihood that a population variance is within a certain range.

To perform the test for one variance using the chi-square statistic, we use the following formula:

$$\chi^2 = \frac{df(s^2)}{\sigma^2}$$

where:

χ^2 is the Chi-Square statistical value.

$df = n - 1$, where n is the size of the sample.

s^2 is the sample variance.

σ^2 is the population variance.

This formula gives us a chi-square statistic, which we can compare to values taken from the chi-square distribution table to test our hypothesis.

We can also construct a confidence interval, which is a range of values that includes the population variance with a given level of confidence. To find this interval, we use the formula shown below:

$$\frac{dfs^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{dfs^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

Review Questions

1. We use the chi-square distribution for the:
    a. goodness-of-fit test
    b. test for independence
    c. testing of a hypothesis of single variance

      d. all of the above

2. True or False: We can test a hypothesis about a single variance using the chi-square distribution for a non-normal population.
3. In testing variance around the population mean, our null hypothesis states that the two population means that we are testing are:

      a. equal with respect to variance
      b. not equal
      c. none of the above

4. In the formula for calculating the chi-square statistic for single variance, σ^2 is:

      a. standard deviation
      b. number of observations
      c. hypothesized population variance
      d. chi-square statistic

5. If we knew the number of observations in a sample, the standard deviation of the sample, and the hypothesized variance of the population, what additional information would we need to solve for the chi-square statistic?

      a. the chi-square distribution table
      b. the population size
      c. the standard deviation of the population
      d. no additional information is needed

6. We want to test a hypothesis about a single variance using the chi-square distribution. We weighed 30 bars of Dial soap, and this sample had a standard deviation of 1.1. We want to test if this sample comes from the general factory, which we know from a previous study to have an overall variance of 3.22. What is our null hypothesis?
7. Compute χ^2 for Question 6.
8. Given the information in Questions 6 and 7, would you reject or fail to reject the null hypothesis?
9. Let's assume that our population variance for this problem is unknown. We want to construct a 90% confidence interval around the population variance, σ^2. If our critical values at a 90% confidence interval are 17.71 and 42.56, what is the range for σ^2?
10. What statement would you give surrounding this confidence interval?

Keywords

ANOVA test

Chi-square distribution

Chi-square statistic

Contingency table

Degrees of freedom

Goodness-of-fit test

Test for one variance

Test of homogeneity

Test of independence

CHAPTER **11** # Analysis of Variance and the F-Distribution

Chapter Outline

11.1 The F-Distribution and Testing Two Variances

Learning Objectives

- Understand the differences between the F-distribution and Student's t-distribution.
- Calculate a test statistic as a ratio of values derived from sample variances.
- Use random samples to test hypotheses about multiple independent population variances.
- Understand the limits of inferences derived from these methods.

Introduction

In previous lessons, we learned how to conduct hypothesis tests that examined the relationship between two variables. Most of these tests simply evaluated the relationship of the means of two variables. However, sometimes we also want to test the variance, or the degree to which observations are spread out within a distribution. In the figure below, we see three samples with identical means (the samples in red, green, and blue) but with very different variances:
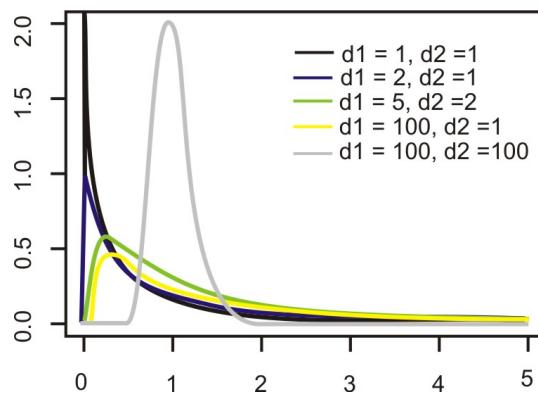
So why would we want to conduct a hypothesis test on variance? Let's consider an example. Suppose a teacher wants to examine the effectiveness of two reading programs. She randomly assigns her students into two groups, uses a different reading program with each group, and gives her students an achievement test. In deciding which reading program is more effective, it would be helpful to not only look at the mean scores of each of the groups, but

also the "spreading out" of the achievement scores. To test hypotheses about variance, we use a statistical tool called the F-distribution.

In this lesson, we will examine the difference between the F-distribution and Student's t-distribution, calculate a test statistic with the F-distribution, and test hypotheses about multiple population variances. In addition, we will look a bit more closely at the limitations of this test.

The

The F-*distribution* is actually a family of distributions. The specific F-distribution for testing two population variances, σ_1^2 and σ_2^2, is based on two values for degrees of freedom (one for each of the populations). Unlike the normal distribution and the t-distribution, F-distributions are not symmetrical and span only non-negative numbers. (Normal distributions and t-distributions are symmetric and have both positive and negative values.) In addition, the shapes of F-distributions vary drastically, especially when the value for degrees of freedom is small. These characteristics make determining the critical values for F-distributions more complicated than for normal distributions and Student's t-distributions. F-distributions for various degrees of freedom are shown below:

We use the F-*ratio test statistic* when testing the hypothesis that there is no difference between population variances. When calculating this ratio, we really just need the variance from each of the samples. It is recommended that the larger sample variance be placed in the numerator of the F-ratio and the smaller sample variance in the denominator. By doing this, the ratio will always be greater than 1.00 and will simplify the hypothesis test.

Example: Suppose a teacher administered two different reading programs to two groups of students and collected the following achievement score data:

Program 1	Program 2
$n_1 = 31$	$n_2 = 41$
$\bar{x}_1 = 43.6$	$\bar{x}_2 = 43.8$
$s_1{}^2 = 105.96$	$s_2{}^2 = 36.42$

What is the F-ratio for these data?

$$F = \frac{s_1{}^2}{s_2{}^2} = \frac{105.96}{36.42} \approx 2.909$$

When we test the hypothesis that two variances of populations from which random samples were selected are equal, $H_0 : \sigma_1^2 = \sigma_2^2$ (or in other words, that the ratio of the variances $\frac{\sigma_1^2}{\sigma_2^2} = 1$), we call this test the *F-Max test*. Since we have a null hypothesis of $H_0 : \sigma_1^2 = \sigma_2^2$, our alternative hypothesis would be $H_a : \sigma_1^2 \neq \sigma_2^2$.

Establishing the critical values in an *F*-test is a bit more complicated than when doing so in other hypothesis tests. Most tables contain multiple *F*-distributions, one for each of the following: 1 percent, 5 percent, 10 percent, and 25 percent of the area in the right-hand tail. (Please see the supplemental link for an example of this type of table.) We also need to use the degrees of freedom from each of the samples to determine the critical values.

On the Web

http://www.statsoft.com/textbook/sttable.html#f01 *F*-distribution tables.

Example: Suppose we are trying to determine the critical values for the scenario in the preceding section, and we set the level of significance to 0.02. Because we have a two-tailed test, we assign 0.01 to the area to the right of the positive critical value. Using the *F*-table for $\alpha = 0.01$, we find the critical value at 2.203, since the numerator has 30 degrees of freedom and the denominator has 40 degrees of freedom.

Once we find our critical values and calculate our test statistic, we perform the hypothesis test the same way we do with the hypothesis tests using the normal distribution and Student's *t*-distribution.

Example: Using our example from the preceding section, suppose a teacher administered two different reading programs to two different groups of students and was interested if one program produced a greater variance in scores. Perform a hypothesis test to answer her question.

For the example, we calculated an *F*-ratio of 2.909 and found a critical value of 2.203. Since the observed test statistic exceeds the critical value, we reject the null hypothesis. Therefore, we can conclude that the observed ratio of the variances from the independent samples would have occurred by chance if the population variances were equal less than 2% of the time. We can conclude that the variance of the student achievement scores for the second sample is less than the variance of the scores for the first sample. We can also see that the achievement test means are practically equal, so the difference in the variances of the student achievement scores may help the teacher in her selection of a program.

The Limits of Using the

The test of the null hypothesis, $H_0; \sigma_1^2 = \sigma_2^2$, using the *F*-distribution is only appropriate when it can safely be assumed that the population is normally distributed. If we are testing the equality of standard deviations between two samples, it is important to remember that the *F*-test is extremely sensitive. Therefore, if the data displays even small departures from the normal distribution, including non-linearity or outliers, the test is unreliable and should not be used. In the next lesson, we will introduce several tests that we can use when the data are not normally distributed.

Lesson Summary

We use the *F*-Max test and the *F*-distribution when testing if two variances from independent samples are equal.

The *F*-distribution differs from the normal distribution and Student's *t*-distribution. Unlike the normal distribution and the *t*-distribution, *F*-distributions are not symmetrical and go from 0 to ∞, not from $-\infty$ to ∞ as the others do.

When testing the variances from independent samples, we calculate the *F*-ratio test statistic, which is the ratio of the variances of the independent samples.

When we reject the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2$, we conclude that the variances of the two populations are not equal.

The test of the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2$, using the F-distribution is only appropriate when it can be safely assumed that the population is normally distributed.

Review Questions

1. We use the F-Max test to examine the differences in the ___ between two independent samples.
2. List two differences between the F-distribution and Student's t-distribution.
3. When we test the differences between the variances of two independent samples, we calculate the ___.
4. When calculating the F-ratio, it is recommended that the sample with the ___ sample variance be placed in the numerator, and the sample with the ___ sample variance be placed in the denominator.
5. Suppose a guidance counselor tested the mean of two student achievement samples from different SAT preparatory courses. She found that the two independent samples had similar means, but also wants to test the variance associated with the samples. She collected the following data:

SAT Prep Course #1   SAT Prep Course #2

$n = 31$ $n = 21$

$s^2 = 42.30$ $s^2 = 18.80$

(a) What are the null and alternative hypotheses for this scenario?

(b) What is the critical value with $\alpha = 0.10$?

(c) Calculate the F-ratio.

(d) Would you reject or fail to reject the null hypothesis? Explain your reasoning.

(e) Interpret the results and determine what the guidance counselor can conclude from this hypothesis test.

6. True or False: The test of the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2$, using the F-distribution is only appropriate when it can be safely assumed that the population is normally distributed.

11.2 The One-Way ANOVA Test
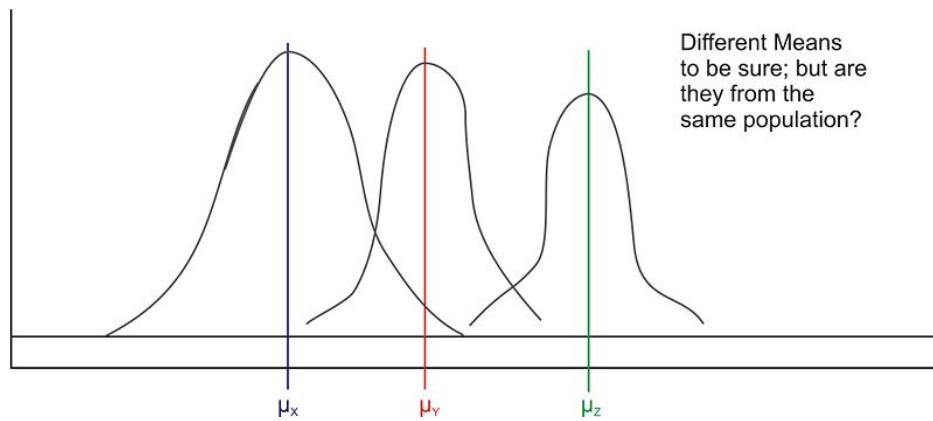
Learning Objectives

- Understand the shortcomings of comparing multiple means as pairs of hypotheses.
- Understand the steps of the ANOVA method and the method's advantages.
- Compare the means of three or more populations using the ANOVA method.
- Calculate pooled standard deviations and confidence intervals as estimates of standard deviations of populations.

Introduction

Previously, we have discussed analyses that allow us to test if the means and variances of two populations are equal. Suppose a teacher is testing multiple reading programs to determine the impact on student achievement. There are five different reading programs, and her 31 students are randomly assigned to one of the five programs. The mean achievement scores and variances for the groups are recorded, along with the means and the variances for all the subjects combined.

We could conduct a series of t-tests to determine if all of the sample means came from the same population. However, this would be tedious and has a major flaw, which we will discuss shortly. Instead, we use something called the Analysis of Variance (ANOVA), which allows us to test the hypothesis that multiple population means and variances of scores are equal. Theoretically, we could test hundreds of population means using this procedure.

Shortcomings of Comparing Multiple Means Using Previously Explained Methods

As mentioned, to test whether pairs of sample means differ by more than we would expect due to chance, we could conduct a series of separate t-tests in order to compare all possible pairs of means. This would be tedious, but we could use a computer or a TI-83/84 calculator to compute these quickly and easily. However, there is a major flaw with this reasoning.

When more than one t-test is run, each at its own level of significance, the probability of making one or more type I errors multiplies exponentially. Recall that a type I error occurs when we reject the null hypothesis when we should not. The level of significance, α, is the probability of a type I error in a single test. When testing more than one pair of samples, the probability of making at least one type I error is $1 - (1 - \alpha)^c$, where α is the level of significance for each t-test and c is the number of independent t-tests. Using the example from the introduction, if our teacher conducted separate t-tests to examine the means of the populations, she would have to conduct 10 separate t-tests. If she performed these tests with $\alpha = 0.05$, the probability of committing a type I error is not 0.05 as one would initially expect. Instead, it would be 0.40, which is extremely high!

The Steps of the ANOVA Method

With the *ANOVA method*, we are actually analyzing the total variation of the scores, including the variation of the scores within the groups and the variation between the group means. Since we are interested in two different types of variation, we first calculate each type of variation independently and then calculate the ratio between the two. We use the F-distribution as our sampling distribution and set our critical values and test our hypothesis accordingly.

When using the ANOVA method, we are testing the null hypothesis that the means and the variances of our samples are equal. When we conduct a hypothesis test, we are testing the probability of obtaining an extreme F-statistic by chance. If we reject the null hypothesis that the means and variances of the samples are equal, and then we are saying that the difference that we see could not have happened just by chance.

To test a hypothesis using the ANOVA method, there are several steps that we need to take. These include:

1. Calculating the *mean squares between groups*, MS_B. The MS_B is the difference between the means of the various samples. If we hypothesize that the group means are equal, then they must also equal the population mean. Under our null hypothesis, we state that the means of the different samples are all equal and come from the same population, but we understand that there may be fluctuations due to sampling error. When we calculate the MS_B, we must first determine the SS_B, which is the sum of the differences between the individual scores and the mean in each group. To calculate this sum, we use the following formula:

$$SS_B = \sum_{k=1}^{m} n_k (\bar{x}_k - \bar{x})^2$$

where:

k is the group number.

n_k is the sample size of group k.

$\bar{x}_k$ is the mean of group k.

$\bar{x}$ is the overall mean of all the observations.

m is the total number of groups.

When simplified, the formula becomes:

$$SS_B = \sum_{k=1}^{m} \frac{T_k^2}{n_k} - \frac{T^2}{n}$$

where:

T_k is the sum of the observations in group k.

T is the sum of all the observations.

n is the total number of observations.

Once we calculate this value, we divide by the number of degrees of freedom, $m - 1$, to arrive at the MS_B. That is, $MS_B = \frac{SS_B}{m-1}$

2. Calculating the *mean squares within groups*, MS_W. The mean squares within groups calculation is also called the *pooled estimate of the population variance*. Remember that when we square the standard deviation of a sample, we are estimating population variance. Therefore, to calculate this figure, we sum the squared deviations within each group and then divide by the sum of the degrees of freedom for each group.

To calculate the MS_W, we first find the SS_W, which is calculated using the following formula:

$$\frac{\sum(x_{i1} - \bar{x}_1)^2 + \sum(x_{i2} - \bar{x}_2)^2 + \ldots + \sum(x_{im} - \bar{x}_m)^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_m - 1)}$$

Simplified, this formula becomes:

$$SS_W = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} \frac{T_k^2}{n_k}$$

where:

T_k is the sum of the observations in group k.

Essentially, this formula sums the squares of each observation and then subtracts the total of the observations squared divided by the number of observations. Finally, we divide this value by the total number of degrees of freedom in the scenario, $n - m$.

$$MS_W = \frac{SS_W}{n - m}$$

3. Calculating the test statistic. The formula for the test statistic is as follows:

$$F = \frac{MS_B}{MS_W}$$

4. Finding the critical value of the F-distribution. As mentioned above, $m - 1$ degrees of freedom are associated with MS_B, and $n - m$ degrees of freedom are associated with MS_W. In a table, the degrees of freedom for MS_B are read across the columns, and the degrees of freedom for MS_W are read across the rows.

5. Interpreting the results of the hypothesis test. In ANOVA, the last step is to decide whether to reject the null hypothesis and then provide clarification about what that decision means.

The primary advantage of using the ANOVA method is that it takes all types of variations into account so that we have an accurate analysis. In addition, we can use technological tools, including computer programs, such as SAS, SPSS, and Microsoft Excel, as well as the TI-83/84 graphing calculator, to easily perform the calculations and test our hypothesis. We use these technological tools quite often when using the ANOVA method.

Example: Let's go back to the example in the introduction with the teacher who is testing multiple reading programs to determine the impact on student achievement. There are five different reading programs, and her 31 students are randomly assigned to one of the five programs. She collects the following data:

Method

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

Compare the means of these different groups by calculating the mean squares between groups, and use the standard deviations from our samples to calculate the mean squares within groups and the pooled estimate of the population variance.

To solve for SS_B, it is necessary to calculate several summary statistics from the data above:

Number (n_k)		7	6	8	5	5	31
Total (T_k)		22	33	51	38	50	$= 194$
Mean $(\bar{x})$		3.14	5.50	6.38	7.60	10.00	$= 6.26$
Sum of Squared Obs. $\left(\sum_{i=1}^{n_k} x_{ik}^2 \right)$		92	199	345	306	510	$= 1,452$
$\dfrac{\text{Sum of Obs. Squared}}{\text{Number of Obs}} \left(\dfrac{T_k^2}{n_k} \right)$		69.14	181.50	325.13	288.80	500.00	$= 1,364.57$

Using this information, we find that the sum of squares between groups is equal to the following:

$$SS_B = \sum_{k=1}^{m} \frac{T_k^2}{n_k} - \frac{T^2}{N}$$

$$\approx 1,364.57 - \frac{(194)^2}{31} \approx 150.5$$

Since there are four degrees of freedom for this calculation (the number of groups minus one), the mean squares between groups is as shown below:

$$MS_B = \frac{SS_B}{m-1} \approx \frac{150.5}{4} \approx 37.6$$

Next, we calculate the mean squares within groups, MS_W, which is also known as the pooled estimate of the population variance, σ^2.

To calculate the mean squares within groups, we first use the following formula to calculate SS_W:

$$SS_W = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} \frac{T_k^2}{n_k}$$

Using our summary statistics from above, we can calculate SS_W as shown below:

$$SS_W = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} \frac{T_k^2}{n_k}$$
$$\approx 1,452 - 1,364.57$$
$$\approx 87.43$$

This means that we have the following for MS_W:

$$MS_W = \frac{SS_W}{n-m} \approx \frac{87.43}{26} \approx 3.36$$

Therefore, our F-ratio is as shown below:

$$F = \frac{MS_B}{MS_W} \approx \frac{37.6}{3.36} \approx 11.19$$

We would then analyze this test statistic against our critical value. Using an F-distribution table for α=0.01 (equivalent to a two-tailed significance of 0.02), and also the numerator degrees of freedom of m-1=5-1=4 and the denominator degrees of freedom of m-n=31-5=26, we find our critical value equal to 4.140. Since our test statistic of 11.19 exceeds the critical value of 4.140, we reject the null hypothesis. We can conclude, therefore, that not all of the population means of the five programs are equal and that obtaining an F-ratio this extreme by chance is highly improbable.

On the Web

http://www.statsoft.com/textbook/sttable.html#f01 *F*-distribution tables.

Technology Note: Calculating a One-Way ANOVA with Excel

Here is the procedure for performing a one-way ANOVA in Excel using this set of data.

Enter the table above (i.e., with data on how the 31 students divided into five groups performed on reading comprehension) into an empty Excel worksheet.

Click the Data choice on the toolbar, then select 'Data Analysis,' and then choose 'Regression' from the list that appears (Note, if Data Analysis does not appear as a choice on your Data page need to follow the add-in instructions below).

Place the cursor in the 'Input Y range' field and select the third column.

Place the cursor in the 'Input X range' field and select the first and second columns.

Place the cursor in the 'Output Range' field and click somewhere in a blank cell below and to the left of the table.

Click 'Labels' so that the names of the predictor variables will be displayed in the table.

Click 'OK', and the results shown below will be displayed.

Note: In Excel 2007, to add Data Analysis to your Data page, perform the following functions. Click the Microsoft Office Button in the upper left, then click on Excel Options. Click on Add-ins, then highlight the Analysis ToolPak, click Go, make sure the Analysis ToolPak box is checked off, and then click OK. The Data Analysis choice should now appear on your Excel Data page. Follow the remaining instructions above.

Anova: Single Factor

TABLE 11.1: SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	7	22	3.142857	3.809524
Column 2	6	33	5.5	3.5
Column 3	8	51	6.375	2.839286
Column 4	5	38	7.6	4.3
Column 5	6	50	10	2.5

TABLE 11.2: ANOVA

Source of Variation	SS	df	MS	F	P- value	F crit
Between Groups	150.5033	4	37.62584	11.18893	2.05e-05	2.742594
Within Groups	87.43214	26	3.362775			
Total	237.9355	30				

Technology Note: One-Way ANOVA on the TI-83/84 Calculator

Enter raw data from population 1 into **L1**, population 2 into **L2**, population 3 into **L3**, population 4 into **L4**, and so on.

Now press **[STAT]**, scroll right to **TESTS**, scroll down to 'ANOVA(', and press **[ENTER]**. Then enter the lists to produce a command such as 'ANOVA(L1, L2, L3, L4)' and press **[ENTER]**.

Lesson Summary

When testing multiple independent samples to determine if they come from the same population, we could conduct a series of separate t-tests in order to compare all possible pairs of means. However, a more precise and accurate analysis is the Analysis of Variance (ANOVA).

In ANOVA, we analyze the total variation of the scores, including the variation of the scores within the groups, the variation between the group means, and the total mean of all the groups (also known as the *grand mean*).

In this analysis, we calculate the F-ratio, which is the total mean of squares between groups divided by the total mean of squares within groups.

The total mean of squares within groups is also known as the pooled estimate of the population variance. We find this value by analysis of the standard deviations in each of the samples.

Review Questions

1. What does the ANOVA acronym stand for?
2. If we are testing whether pairs of sample means differ by more than we would expect due to chance using multiple t-tests, the probability of making a type I error would ___.
3. In the ANOVA method, we use the ___ distribution.

    a. Student's t-

    b. normal

    c. *F*-

4. In the ANOVA method, we complete a series of steps to evaluate our hypothesis. Put the following steps in chronological order.

    a. Calculate the mean squares between groups and the mean squares within groups.
    b. Determine the critical values in the *F*-distribution.
    c. Evaluate the hypothesis.
    d. Calculate the test statistic.
    e. State the null hypothesis.

5. A school psychologist is interested in whether or not teachers affect the anxiety scores among students taking the AP Statistics exam. The data below are the scores on a standardized anxiety test for students with three different teachers.

TABLE 11.3: Teacher's Name and Anxiety Scores

Ms. Jones	Mr. Smith	Mrs. White
8	23	21
6	11	21
4	17	22
12	16	18
16	6	14
17	14	21
12	15	9
10	19	11
11	10	
13		

(a) State the null hypothesis.

(b) Using the data above, fill out the missing values in the table below.

TABLE 11.4:

	Ms. Jones	Mr. Smith	Mrs. White	Totals
Number (n_k)			8	$=$
Total (T_k)		131		$=$
Mean $(\bar{x})$		14.6		$=$
Sum of Squared Obs. $(\sum_{i=1}^{n_k} x_{ik}^2)$				$=$
Sum of Obs. Squared/Number of Obs. $(\frac{T_k^2}{n_k})$				$=$

(c) What is the value of the mean squares between groups, MS_B?

(d) What is the value of the mean squares within groups, MS_W?

(e) What is the *F*-ratio of these two values?

(f) With $\alpha = 0.05$, use the *F*-distribution to set a critical value.

(g) What decision would you make regarding the null hypothesis? Why?

11.3 The Two-Way ANOVA Test

Learning Objectives

- Understand the differences in situations that allow for one-way or two-way ANOVA methods.
- Know the procedure of two-way ANOVA and its application through technological tools.
- Understand completely randomized and randomized block methods of experimental design and their relation to appropriate ANOVA methods.

Introduction

In the previous section, we discussed the one-way ANOVA method, which is the procedure for testing the null hypothesis that the population means and variances of a single independent variable are equal. Sometimes, however, we are interested in testing the means and variances of more than one independent variable. Say, for example, that a researcher is interested in determining the effects of different dosages of a dietary supplement on the performance of both males and females on a physical endurance test. The three different dosages of the medicine are low, medium, and high, and the genders are male and female. Analyses of situations with two independent variables, like the one just described, are called two-way ANOVA tests.

TABLE 11.5: Mean Scores on a Physical Endurance Test for Varying Dosages and Genders

	Dietary Supplement Dosage	Dietary Supplement Dosage	Dietary Supplement Dosage	
	Low	Medium	High	Average
Female	35.6	49.4	71.8	52.3
Male	55.2	92.2	110.0	85.8
Average	45.4	70.8	90.9	

There are several questions that can be answered by a study like this, such as, "Does the medication improve physical endurance, as measured by the test?" and "Do males and females respond in the same way to the medication?"

While there are similar steps in performing one-way and two-way ANOVA tests, there are also some major differences. In the following sections, we will explore the differences in situations that allow for the one-way or two-way ANOVA methods, the procedure of two-way ANOVA, and the experimental designs associated with this method.

The Differences in Situations that Allow for One-way or Two-Way ANOVA

As mentioned in the previous lesson, ANOVA allows us to examine the effect of a single independent variable on a dependent variable (i.e., the effectiveness of a reading program on student achievement). With *two-way ANOVA*, we are not only able to study the effect of two independent variables (i.e., the effect of dosages and gender on the results of a physical endurance test), but also the interaction between these variables. An example of interaction between the two variables gender and medication is a finding that men and women respond differently to the medication.

351

We could conduct two separate one-way ANOVA tests to study the effect of two independent variables, but there are several advantages to conducting a two-way ANOVA test.

Efficiency. With simultaneous analysis of two independent variables, the ANOVA test is really carrying out two separate research studies at once.

Control. When including an additional independent variable in the study, we are able to control for that variable. For example, say that we included IQ in the earlier example about the effects of a reading program on student achievement. By including this variable, we are able to determine the effects of various reading programs, the effects of IQ, and the possible interaction between the two.

Interaction. With a two-way ANOVA test, it is possible to investigate the interaction of two or more independent variables. In most real-life scenarios, variables do interact with one another. Therefore, the study of the interaction between independent variables may be just as important as studying the interaction between the independent and dependent variables.

When we perform two separate one-way ANOVA tests, we run the risk of losing these advantages.

Two-Way ANOVA Procedures

There are two kinds of variables in all ANOVA procedures-dependent and independent variables. In one-way ANOVA, we were working with one independent variable and one dependent variable. In two-way ANOVA, there are two independent variables and a single dependent variable. Changes in the dependent variables are assumed to be the result of changes in the independent variables.

In one-way ANOVA, we calculated a ratio that measured the variation between the two variables (dependent and independent). In two-way ANOVA, we need to calculate a ratio that measures not only the variation between the dependent and independent variables, but also the interaction between the two independent variables.

Before, when we performed the one-way ANOVA, we calculated the total variation by determining the variation within groups and the variation between groups. Calculating the total variation in two-way ANOVA is similar, but since we have an additional variable, we need to calculate two more types of variation. Determining the total variation in two-way ANOVA includes calculating: variation within the group (within-cell variation), variation in the dependent variable attributed to one independent variable (variation among the row means), variation in the dependent variable attributed to the other independent variable (variation among the column means), and variation between the independent variables (the interaction effect).

The formulas that we use to calculate these types of variations are very similar to the ones that we used in the one-way ANOVA. For each type of variation, we want to calculate the total sum of squared deviations (also known as the sum of squares) around the grand mean. After we find this total sum of squares, we want to divide it by the number of degrees of freedom to arrive at the mean of squares, which allows us to calculate our final ratio. We could do these calculations by hand, but we have technological tools, such as computer programs like Microsoft Excel and graphing calculators, that can compute these figures much more quickly and accurately than we could manually. In order to perform a two-way ANOVA with a TI-83/84 calculator, you must download a calculator program at the following site: http://www.wku.edu/~david.neal/statistics/ .

The process for determining and evaluating the null hypothesis for the two-way ANOVA is very similar to the same process for the one-way ANOVA. However, for the two-way ANOVA, we have additional hypotheses, due to the additional variables. For two-way ANOVA, we have three null hypotheses:

1. In the population, the means for the rows equal each other. In the example above, we would say that the mean for males equals the mean for females.
2. In the population, the means for the columns equal each other. In the example above, we would say that the means for the three dosages are equal.
3. In the population, the null hypothesis would be that there is no interaction between the two variables. In the example above, we would say that there is no interaction between gender and amount of dosage, or that all

effects equal 0.

Let's take a look at an example of a data set and see how we can interpret the summary tables produced by technological tools to test our hypotheses.

Example: Say that a gym teacher is interested in the effects of the length of an exercise program on the flexibility of male and female students. The teacher randomly selected 48 students (24 males and 24 females) and assigned them to exercise programs of varying lengths (1, 2, or 3 weeks). At the end of the programs, she measured the students' flexibility and recorded the following results. Each cell represents the score of a student:

TABLE 11.6:

		Length of Program 1 Week	Length of Program 2 Weeks	Length of Program 3 Weeks
Gender	Females	32	28	36
		27	31	47
		22	24	42
		19	25	35
		28	26	46
		23	33	39
		25	27	43
		21	25	40
	Males	18	27	24
		22	31	27
		20	27	33
		25	25	25
		16	25	26
		19	32	30
		24	26	32
		31	24	29

Do gender and the length of an exercise program have an effect on the flexibility of students?

Solution:

From these data, we can calculate the following summary statistics:

TABLE 11.7:

			Length of Program 1 Week	Length of Program 2 Weeks	Length of Program 3 Weeks	Total
Gender	Females	n	8	8	8	24
		Mean	24.6	27.4	41.0	31.0
		St. Dev.	4.24	3.16	4.34	8.23
	Males	n	8	8	8	24
		Mean	21.9	27.1	28.3	25.8
		St. Dev.	4.76	2.90	3.28	4.56
	Totals	n	16	16	16	48
		Mean	23.3	27.3	34.6	28.4
		St. Dev.	4.58	2.93	7.56	7.10

As we can see from the tables above, it appears that females have more flexibility than males and that the longer

programs are associated with greater flexibility. Also, we can take a look at the standard deviation of each group to get an idea of the variance within groups. This information is helpful, but it is necessary to calculate the test statistic to more fully understand the effects of the independent variables and the interaction between these two variables.

Technology Note: Calculating a Two-Way ANOVA with Excel

Here is the procedure for performing a two-way ANOVA with Excel using this set of data.

1. Copy and paste the earlier table (with the flexibility data from the 48 students) into an empty Excel worksheet, without the labels 'Length of program' and 'Gender'.
2. Select 'Data Analysis' from the Tools menu and choose 'ANOVA: Two-Factor Without Replication' from the list that appears.
3. Place the cursor in the 'Input Range' field and select the entire table.
4. Place the cursor in the 'Output Range' field and click somewhere in a blank cell below the table.
5. Click 'Labels' only if you have also included the labels in the table. This will cause the names of the predictor variables to be displayed in the table.
6. Click 'OK', and the results shown below will be displayed.

Using technological tools, we can generate the following summary table:

TABLE 11.8:

Source	*SS*	*df*	*MS*	*F*	Critical Value of *F**
Rows (gender)	582.58	15	38.84	1.62	2.015
Columns (length)	1,065.5	2	532.75	22.22	3.32
Error	719.17	30	23.97		
Total	2,367.25	47			

∗Statistically significant at $\alpha = 0.05$.

From this summary table, we can see that all three *F*-ratios exceed their respective critical values.

This means that we can reject all three null hypotheses and conclude that:

In the population, the mean for males differs from the mean of females.

In the population, the means for the three exercise programs differ.

There is an interaction between the length of the exercise program and the student's gender.

Experimental Design and its Relation to the ANOVA Methods

Experimental design is the process of taking the time and the effort to organize an experiment so that the data are readily available to answer the questions that are of most interest to the researcher. When conducting an experiment using the ANOVA method, there are several ways that we can design an experiment. The design that we choose depends on the nature of the questions that we are exploring.

In a totally randomized design, the subjects or objects are assigned to treatment groups completely at random. For example, a teacher might randomly assign students into one of three reading programs to examine the effects of the different reading programs on student achievement. Often, the person conducting the experiment will use a computer to randomly assign subjects.

In a randomized block design, subjects or objects are first divided into homogeneous categories before being randomly assigned to a treatment group. For example, if an athletic director was studying the effect of various

physical fitness programs on males and females, he would first categorize the randomly selected students into homogeneous categories (males and females) before randomly assigning them to one of the physical fitness programs that he was trying to study.

In ANOVA, we use both randomized design and randomized block design experiments. In one-way ANOVA, we typically use a completely randomized design. By using this design, we can assume that the observed changes are caused by changes in the independent variable. In two-way ANOVA, since we are evaluating the effect of two independent variables, we typically use a randomized block design. Since the subjects are assigned to one group and then another, we are able to evaluate the effects of both variables and the interaction between the two.

Lesson Summary

With two-way ANOVA, we are not only able to study the effect of two independent variables, but also the interaction between these variables. There are several advantages to conducting a two-way ANOVA, including efficiency, control of variables, and the ability to study the interaction between variables. Determining the total variation in two-way ANOVA includes calculating the following:

Variation within the group (within-cell variation)

Variation in the dependent variable attributed to one independent variable (variation among the row means)

Variation in the dependent variable attributed to the other independent variable (variation among the column means)

Variation between the independent variables (the interaction effect)

It is easier and more accurate to use technological tools, such as computer programs like Microsoft Excel, to calculate the figures needed to evaluate our hypotheses tests.

Review Questions

1. In two-way ANOVA, we study not only the effect of two independent variables on the dependent variable, but also the ___ between the two independent variables.
2. We could conduct multiple t-tests between pairs of hypotheses, but there are several advantages when we conduct a two-way ANOVA. These include:
    a. Efficiency
    b. Control over additional variables
    c. The study of interaction between variables
    d. All of the above

3. Calculating the total variation in two-way ANOVA includes calculating ___ types of variation.

    a. 1
    b. 2
    c. 3
    d. 4

4. A researcher is interested in determining the effects of different doses of a dietary supplement on the performance of both males and females on a physical endurance test. The three different doses of the medicine are low, medium, and high, and again, the genders are male and female. He assigns 48 people, 24 males and 24 females, to one of the three levels of the supplement dosage and gives a standardized physical endurance test. Using technological tools, he generates the following summary ANOVA table:

TABLE 11.9:

Source	SS	df	MS	F	Critical Value of F
Rows (gender)	14.832	1	14.832	14.94	4.07
Columns (dosage)	17.120	2	8.560	8.62	3.23
Interaction	2.588	2	1.294	1.30	3.23
Within-cell	41.685	42	992		
Total	76,226	47			

$$^*\alpha = 0.05$$

(a) What are the three hypotheses associated with the two-way ANOVA method?

(b) What are the three null hypotheses for this study?

(c) What are the critical values for each of the three hypotheses? What do these tell us?

(d) Would you reject the null hypotheses? Why or why not?

(e) In your own words, describe what these results tell us about this experiment.

On the Web

http://www.ruf.rice.edu/~lane/stat_sim/two_way/index.html Two-way ANOVA applet that shows how the sums of square total is divided between factors A and B, the interaction of A and B, and the error.

http://tinyurl.com/32qaufs Shows partitioning of sums of squares in a one-way analysis of variance.

http://tinyurl.com/djob5t Understanding ANOVA visually. There are no numbers or formulas.

Keywords

ANOVA method

Experimental design

F-distribution

F-Max test

F-ratio test statistic

Grand mean

Mean squares between groups

Mean squares within groups

Pooled estimate of the population variance

SS_B

SS_W

Two-way ANOVA

CHAPTER **12** **Non-Parametric Statistics**

Chapter Outline

12.1 Introduction to Non-Parametric Statistics

Learning Objectives

- Understand situations in which non-parametric analytical methods should be used and the advantages and disadvantages of each of these methods.
- Understand situations in which the sign test can be used and calculate z-scores for evaluating a hypothesis using matched pair data sets.
- Use the sign test to evaluate a hypothesis about the median of a population.
- Examine a categorical data set to evaluate a hypothesis using the sign test.
- Understand the signed-ranks test as a more precise alternative to the sign test when evaluating a hypothesis.

Introduction

In previous lessons, we discussed the use of the normal distribution, Student's t-distribution, and the F-distribution in testing various hypotheses. With each of these distributions, we made certain assumptions about the populations from which our samples were drawn. Specifically, we made assumptions that the underlying populations were normally distributed and that there was homogeneity of variance within the populations. But what do we do when we have data that are not normally distributed or not homogeneous with respect to variance? In these situations, we use something called non-parametric tests.

These tests include tests such as the sign test, the sign-ranks test, the ranks-sum test, the Kruskal-Wallis test, and the runs test. While parametric tests are preferred, since they are more powerful, they are not always applicable. The following sections will examine situations in which we would use non-parametric methods and the advantages and disadvantages of using these methods.

Situations Where We Use Non-Parametric Tests

If *non-parametric tests* have fewer assumptions and can be used with a broader range of data types, why don't we use them all the time? The reason is because there are several advantages of using parametric tests. They are more robust and have greater power, which means that they have a greater chance of rejecting the null hypothesis relative to the sample size when the null hypothesis is false.

However, parametric tests demand that the data meet stringent requirements, such as normality and homogeneity of variance. For example, a one-sample t-test requires that the sample be drawn from a normally distributed population. When testing two independent samples, not only is it required that both samples be drawn from normally distributed populations, but it is also required that the standard deviations of the populations be equal. If either of these conditions is not met, our results are not valid.

As mentioned, an advantage of non-parametric tests is that they do not require the data to be normally distributed. In addition, although they test the same concepts, non-parametric tests sometimes have fewer calculations than their parametric counterparts. Non-parametric tests are often used to test different types of questions and allow us to perform analysis with categorical and rank data. The table below lists the parametric tests, their non-parametric counterparts, and the purpose of each test.

Commonly Used Parametric and Non-parametric Tests

TABLE 12.1:

Parametric Test (Normal Distributions)	Non-parametric Test (Non-normal Distributions)	Purpose of Test
t-test for independent samples	Rank sum test	Compares means of two independent samples
Paired t-test	Sign test	Examines a set of differences of means
Pearson correlation coefficient	Rank correlation test	Assesses the linear association between two variables.
One-way analysis of variance (F-test)	Kruskal-Wallis test	Compares three or more groups
Two-way analysis of variance	Runs test	Compares groups classified by two different factors

The Sign Test

One of the simplest non-parametric tests is the *sign test*. The sign test examines the difference in the medians of matched data sets. It is important to note that we use the sign test only when testing if there is a difference between the matched pairs of observations. This test does not measure the magnitude of the relationship-it simply tests whether the differences between the observations in the matched pairs are equally likely to be positive or negative. Many times, this test is used in place of a paired t-test.

For example, we would use the sign test when assessing if a certain drug or treatment had an impact on a population or if a certain program made a difference in behavior. We first determine whether there is a positive or negative difference between each of the matched pairs. To determine this, we arrange the data in such a way that it is easy to identify what type of difference that we have. Let's take a look at an example to help clarify this concept.

Example: Suppose we have a school psychologist who is interested in whether or not a behavior intervention program is working. He examines 8 middle school classrooms and records the number of referrals written per month both before and after the intervention program. Below are his observations:

TABLE 12.2:

Observation Number	Referrals Before Program	Referrals After Program
1	8	5
2	10	8
3	2	3
4	4	1
5	6	4
6	4	1
7	5	7
8	9	6

Since we need to determine the number of observations where there is a positive difference and the number of observations where there is a negative difference, it is helpful to add an additional column to the table to classify each observation as such (see below). We ignore all zero or equal observations.

<div align="center">**TABLE 12.3:**</div>

Observation Number	Referrals Before Program	Referrals After Program	Change
1	8	5	−
2	10	8	−
3	2	3	+
4	4	1	−
5	6	4	−
6	4	1	−
7	5	7	+
8	9	6	−

The test statistic we use is $\dfrac{|\text{number of positive changes – number of negative changes}|-1}{\sqrt{n}}$.

If the sample has fewer than 30 observations, we use the t-distribution to determine a critical value and make a decision. If the sample has more than 30 observations, we use the normal distribution.

Our example has only 8 observations, so we calculate our t-score as shown below:

$$t = \frac{|2-6|-1}{\sqrt{8}} = 1.06$$

Similar to other hypothesis tests using standard scores, we establish null and alternative hypotheses about the population and use the test statistic to assess these hypotheses. As mentioned, this test is used with paired data and examines whether the medians of the two data sets are equal. When we conduct a pre-test and a post-test using matched data, our null hypothesis is that the difference between the data sets will be zero. In other words, under our null hypothesis, we would expect there to be some fluctuations between the pre-test and post-test, but nothing of significance. Therefore, our null and alternative hypotheses would be as follows:

$$H_0 : m = 0$$
$$H_a : m \neq 0$$

With the sign test, we set criterion for rejecting the null hypothesis in the same way as we did when we were testing hypotheses using parametric tests. For the example above, if we set $\alpha = 0.05$, we would have critical values at 2.36 standard scores above and below the mean. Since our standard score of 1.06 is less than the critical value of 2.36, we would fail to reject the null hypothesis and cannot conclude that there is a significant difference between the pre-test and post-test scores.

When we use the sign test to evaluate a hypothesis about the median of a population, we are estimating the likelihood, or the probability, that the number of successes would occur by chance if there was no difference between pre-test and post-test data. When working with small samples, the sign test is actually the binomial test, with the null hypothesis being that the proportion of successes will equal 0.5.

Example: Suppose a physical education teacher is interested in the effect of a certain weight-training program on students' strength. She measures the number of times students are able to lift a dumbbell of a certain weight before the program and then again after the program. Below are her results:

TABLE 12.4:

Before Program	After Program	Change
12	21	+
9	16	+
11	14	+
21	36	+
17	28	+
22	20	−
18	29	+
11	22	+

If the program had no effect, then the proportion of students with increased strength would equal 0.5. Looking at the data above, we see that 7 of the 8 students had increased strength after the program. But is this statistically significant? To answer this question, we use the binomial formula, which is as follows:

$$P(r) = \frac{N!}{r!\,(N-r)!} p^r (1-p)^{N-r}$$

Using this formula, we need to determine the probability of having either 7 or 8 successes as shown below:

$$P(7) = \frac{8!}{7!\,(8-7)!} 0.5^7 (1-0.5)^{8-7} = (8)(00391) = 0.03125$$

$$P(8) = \frac{8!}{8!\,(8-8)!} 0.5^8 (1-0.5)^{8-8} = 0.00391$$

To determine the probability of having either 7 or 8 successes, we add the two probabilities together and get $0.03125 + 0.00391 = 0.0352$. This means that if the program had no effect on the matched data set, we have a 0.0352 likelihood of obtaining the number of successes that we did by chance.

Using the Sign Test to Examine Categorical Data

We can also use the sign test to examine differences and evaluate hypotheses with categorical data sets. Recall that we typically use the chi-square distribution to assess categorical data. We could use the sign test when determining if one categorical variable is really more than another. For example, we could use this test if we were interested in determining if there were equal numbers of students with brown eyes and blue eyes. In addition, we could use this test to determine if equal numbers of males and females get accepted to a four-year college.

When using the sign test to examine a categorical data set and evaluate a hypothesis, we use the same formulas and methods as if we were using nominal data. The only major difference is that instead of labeling the observations as positives or negatives, we would label the observations with whatever dichotomy we want to use (male/female, brown/blue, etc.) and calculate the test statistic, or probability, accordingly. Again, we would not count zero or equal observations.

Example: The UC admissions committee is interested in determining if the numbers of males and females who are accepted into four-year colleges differ significantly. They take a random sample of 200 graduating high school seniors who have been accepted to four-year colleges. Out of these 200 students, they find that there are 134 females and 66 males. Do the numbers of males and females accepted into colleges differ significantly? Since we have a large sample, calculate the z-score and use $\alpha = 0.05$.

To answer this question using the sign test, we would first establish our null and alternative hypotheses:

$$H_0 : m = 0$$
$$H_a : m \neq 0$$

This null hypothesis states that the median numbers of males and females accepted into UC schools are equal.

Next, we use $\alpha = 0.05$ to establish our critical values. Using the normal distribution table, we find that our critical values are equal to 1.96 standard scores above and below the mean.

To calculate our test statistic, we use the following formula:

$$\frac{|\text{number of positive changes } - \text{ number of negative changes}| - 1}{\sqrt{n}}$$

However, instead of the numbers of positive and negative observations, we substitute the number of females and the number of males. Because we are calculating the absolute value of the difference, the order of the variables does not matter. Therefore, our z-score can be calculated as shown:

$$z = \frac{|134 - 66| - 1}{\sqrt{200}} = 4.74$$

With a calculated test statistic of 4.74, we can reject the null hypothesis and conclude that there is a difference between the number of graduating males and the number of graduating females accepted into the UC schools.

The Benefit of Using the Sign Rank Test

As previously mentioned, the sign test is a quick and easy way to test if there is a difference between pre-test and post-test matched data. When we use the sign test, we simply analyze the number of observations in which there is a difference. However, the sign test does not assess the magnitude of these differences.

A more useful test that assesses the difference in size between the observations in a matched pair is the *sign rank test*. The sign rank test (also known as the *Wilcoxon sign rank test*) resembles the sign test, but it is much more sensitive. Similar to the sign test, the sign rank test is also a nonparametric alternative to the paired Student's t-test. When we perform this test with large samples, it is almost as sensitive as Student's t-test, and when we perform this test with small samples, it is actually more sensitive than Student's t-test.

The main difference with the sign rank test is that under this test, the hypothesis states that the difference between observations in each data pair (pre-test and post-test) is equal to zero. Essentially, the null hypothesis states that the two variables have identical distributions. The sign rank test is much more sensitive than the sign test, since it measures the difference between matched data sets. Therefore, it is important to note that the results from the sign and the sign rank test could be different for the same data set.

To conduct the sign rank test, we first rank the differences between the observations in each matched pair, without regard to the sign of the difference. After this initial ranking, we affix the original sign to the rank numbers. All equal observations get the same rank and are ranked with the mean of the rank numbers that would have been assigned if they had varied. After this ranking, we sum the ranks in each sample and then determine the total number of observations. Finally, the one sample z-statistic is calculated from the signed ranks. For large samples, the z-statistic is compared to percentiles of the standard normal distribution.

It is important to remember that the sign rank test is more precise and sensitive than the sign test. However, since we are ranking the nominal differences between variables, we are not able to use the sign rank test to examine the differences between categorical variables. In addition, this test can be a bit more time consuming to conduct, since the figures cannot be calculated directly in Excel or with a calculator.

Lesson Summary

We use non-parametric tests when the assumptions of normality and homogeneity of variance are not met.

There are several different non-parametric tests that we can use in lieu of their parametric counterparts. These tests include the sign test, the sign rank test, the rank-sum test, the Kruskal-Wallis test, and the runs test.

The sign test examines the difference in the medians of matched data sets. When testing hypotheses using the sign test, we can calculate the standard z-score when working with large samples or use the binomial formula when working with small samples.

We can also use the sign test to examine differences and evaluate hypotheses with categorical data sets.

A more precise test that assesses the difference in size between the observations in a matched pair is the sign rank test.

12.2 The Rank Sum Test and Rank Correlation

Learning Objectives

- Understand the conditions for use of the rank sum test to evaluate a hypothesis about non-paired data.
- Calculate the mean and the standard deviation of rank from two non-paired samples and use these values to calculate a *z*-score.
- Determine the correlation between two variables using the rank correlation test for situations that meet the appropriate criteria, using the appropriate test statistic formula.

Introduction

In the previous lesson, we explored the concept of nonparametric tests. We explored two tests-the sign test and the sign rank test. We use these tests when analyzing matched data pairs or categorical data samples. In both of these tests, our null hypothesis states that there is no difference between the medians of these variables. As mentioned, the sign rank test is a more precise test of this question, but the test statistic can be more difficult to calculate.

But what happens if we want to test if two samples come from the same non-normal distribution? For this type of question, we use the rank sum test (also known as the *Mann-Whitney v-test*). This test is sensitive to both the median and the distribution of the sample and population.

In this section, we will learn how to conduct hypothesis tests using the Mann-Whitney *v*-test and the situations in which it is appropriate to do so. In addition, we will also explore how to determine the correlation between two variables from non-normal distributions using the rank correlation test for situations that meet the appropriate criteria.

Conditions for Use of the Rank Sum Test to Evaluate Hypotheses about Non-Paired Data

The *rank sum test* tests the hypothesis that two independent samples are drawn from the same population. Recall that we use this test when we are not sure if the assumptions of normality or homogeneity of variance are met. Essentially, this test compares the medians and the distributions of the two independent samples. This test is considered stronger than other nonparametric tests that simply assess median values. For example, in the image below, we see that the two samples have the same median, but very different distributions. If we were assessing just the median value, we would not realize that these samples actually have distributions that are very distinct.

When performing the rank sum test, there are several different conditions that need to be met. These include the following:

- Although the populations need not be normally distributed or have homogeneity of variance, the observations must be continuously distributed.
- The samples drawn from the population must be independent of one another.
- The samples must have 5 or more observations. The samples do not need to have the same number of observations.
- The observations must be on a numeric or ordinal scale. They cannot be categorical variables.

Since the rank sum test evaluates both the medians and the distributions of two independent samples, we establish two null hypotheses. Our null hypotheses state that the two medians and the two standard deviations of the independent samples are equal. Symbolically, we could say $H_0 : m_1 = m_2$ and $\sigma_1 = \sigma_2$. The alternative hypotheses state that there is a difference in the medians and the standard deviations of the samples.

Calculating the Mean and the Standard Deviation of Rank to Calculate a

When performing the rank sum test, we need to calculate a figure known as the U-statistic. This statistic takes both the median and the total distribution of the two samples into account. The U-statistic actually has its own distribution, which we use when working with small samples. (In this test, a small sample is defined as a sample less than 20 observations.) This distribution is used in the same way that we would use the t-distribution and the chi-square distribution. Similar to the t-distribution, the U-distribution approaches the normal distribution as the sizes of both samples grow. When we have samples of 20 or more, we do not use the U-distribution. Instead, we use the U-statistic to calculate the standard z-score.

To calculate the U-statistic, we must first arrange and rank the data from our two independent samples. First, we must rank all values from both samples from low to high, without regard to which sample each value belongs to. If two values are the same, then they both get the average of the two ranks for which they tie. The smallest number gets a rank of 1, and the largest number gets a rank of n, where n is the total number of values in the two groups. After we arrange and rank the data in each of the samples, we sum the ranks assigned to the observations. We record both the sum of these ranks and the number of observations in each of the samples. After we have this information, we can use the following formulas to determine the U-statistic:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

where:

n_1 is the number of observations in sample 1.

n_2 is the number of observations in sample 2.

R_1 is the sum of the ranks assigned to sample 1.

R_2 is the sum of the ranks assigned to sample 2.

We use the smaller of the two calculated test statistics (i.e., the lesser of U_1 and U_2) to evaluate our hypotheses in smaller samples or to calculate the z-score when working with larger samples.

When working with larger samples, we need to calculate two additional pieces of information: the mean of the sampling distribution, μ_U, and the standard deviation of the sampling distribution, σ_U. These calculations are relatively straightforward when we know the numbers of observations in each of the samples. To calculate these figures, we use the following formulas:

$$\mu_U = \frac{n_1 n_2}{2} \text{ and } \sigma_U = \sqrt{\frac{n_1(n_2)(n_1 + n_2 + 1)}{12}}$$

Finally, we use the general formula for the test statistic to test our null hypothesis:

$$z = \frac{U - \mu_U}{\sigma_U}$$

Example: Suppose we are interested in determining the attitudes on the current status of the economy from women who work outside the home and from women who do not work outside the home. We take a sample of 20 women who work outside the home (sample 1) and a sample of 20 women who do not work outside the home (sample 2) and administer a questionnaire that measures their attitudes about the economy. These data are found in the tables below:

TABLE 12.5:

Women Working Outside the Home	Women Working Outside the Home
Score	Rank
9	1
12	3
13	4
19	8
21	9
27	13
31	16
33	17
34	18
35	19
39	21
40	22
44	25
46	26
49	29
58	33
61	34
63	35
64	36
70	39
	$R_1 = 408$

TABLE 12.6:

Women Not Working Outside the Home	Women Not Working Outside the Home
Score	Rank
10	2
15	5
17	6
18	7
23	10

<div align="center">**TABLE 12.6:** (continued)</div>

Women Not Working Outside the Home	Women Not Working Outside the Home
24	11
25	12
28	14
30	15
37	20
41	23
42	24
47	27
48	28
52	30
55	31
56	32
65	37
69	38
71	40
	$R_2 = 412$

Do these two groups of women have significantly different views on the issue?

Since each of our samples has 20 observations, we need to calculate the standard z-score to test the hypothesis that these independent samples came from the same population. To calculate the z-score, we need to first calculate U, μ_U, and σ_U. The U-statistic for each of the samples is calculated as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (20)(20) + \frac{(20)(20+1)}{2} - 408 = 202$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (20)(20) + \frac{(20)(20+1)}{2} - 412 = 198$$

Since we use the smaller of the two U-statistics, we set $U = 198$. When calculating the other two figures, we find the following:

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(20)(20)}{2} = 200$$

and

$$\sigma_U = \sqrt{\frac{n_1(n_2)(n_1+n_2+1)}{12}} = \sqrt{\frac{(20)(20)(20+20+1)}{12}} = \sqrt{\frac{(400)(41)}{12}} = 36.97$$

Thus, we calculate the z-statistic as shown below:

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{198 - 200}{36.97} = -0.05$$

If we set $\alpha = 0.05$, we would find that the calculated test statistic does not exceed the critical value of -1.96. Therefore, we fail to reject the null hypothesis and conclude that these two samples come from the same population.

We can use this *z*-score to evaluate our hypotheses just like we would with any other hypothesis test. When interpreting the results from the rank sum test, it is important to remember that we are really asking whether or not the populations have the same median and variance. In addition, we are assessing the chance that random sampling would result in medians and variances as far apart (or as close together) as observed in the test. If the *z*-score is large (meaning that we would have a small *P*-value), we can reject the idea that the difference is a coincidence. If the *z*-score is small, like in the example above (meaning that we would have a large *P*-value), we do not have any reason to conclude that the medians of the populations differ and, therefore, conclude that the samples likely came from the same population.

Determining the Correlation between Two Variables Using the Rank Correlation Test

It is possible to determine the correlation between two variables by calculating the Pearson product-moment correlation coefficient (more commonly known as the linear correlation coefficient, or *r*). The correlation coefficient helps us determine the strength, magnitude, and direction of the relationship between two variables with normal distributions.

We also use the *Spearman rank correlation coefficient* (also known simply as the *rank correlation coefficient*, ρ, or 'rho') to measure the strength, magnitude, and direction of the relationship between two variables. This test statistic is the nonparametric alternative to the correlation coefficient, and we use it when the data do not meet the assumptions of normality. The Spearman rank correlation coefficient, used as part of the *rank correlation test*, can also be used when one or both of the variables consist of ranks. The Spearman rank correlation coefficient is defined by the following formula:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where *d* is the difference in statistical rank of corresponding observations.

The test works by converting each of the observations to ranks, just like we learned about with the rank sum test. Therefore, if we were doing a rank correlation of scores on a final exam versus SAT scores, the lowest final exam score would get a rank of 1, the second lowest a rank of 2, and so on. Likewise, the lowest SAT score would get a rank of 1, the second lowest a rank of 2, and so on. Similar to the rank sum test, if two observations are equal, the average rank is used for both of the observations. Once the observations are converted to ranks, a correlation analysis is performed on the ranks. (Note: This analysis is not performed on the observations themselves.) The Spearman correlation coefficient is then calculated from the columns of ranks. However, because the distributions are non-normal, a regression line is rarely used, and we do not calculate a non-parametric equivalent of the regression line. It is easy to use a statistical programming package, such as SAS or SPSS, to calculate the Spearman rank correlation coefficient. However, for the purposes of this example, we will perform this test by hand as shown in the example below.

Example: The head of a math department is interested in the correlation between scores on a final math exam and math SAT scores. She took a random sample of 15 students and recorded each student's final exam score and math SAT score. Since SAT scores are designed to be normally distributed, the Spearman rank correlation test may be an especially effective tool for this comparison. Use the Spearman rank correlation test to determine the correlation coefficient. The data for this example are recorded below:

TABLE 12.7:

Math SAT Score	Final Exam Score
595	68
520	55
715	65

<div align="center">**TABLE 12.7:** (continued)</div>

Math SAT Score	Final Exam Score
405	42
680	64
490	45
565	56
580	59
615	56
435	42
440	38
515	50
380	37
510	42
565	53

To calculate the Spearman rank correlation coefficient, we determine the ranks of each of the variables in the data set, calculate the difference for each of these ranks, and then calculate the squared difference.

<div align="center">**TABLE 12.8:**</div>

Math SAT Score (X)	Final Exam Score (Y)	X Rank	Y Rank	d	d^2
595	68	4	1	3	9
520	55	8	7	1	1
715	65	1	2	−1	1
405	42	14	12	2	4
680	64	2	3	−1	1
490	45	11	10	1	1
565	56	6.5	5.5	1	1
580	59	5	4	1	1
615	56	3	5.5	−2.5	6.25
435	42	13	12	1	1
440	38	12	14	−2	4
515	50	9	9	0	0
380	37	15	15	0	0
510	42	10	12	−2	4
565	53	6.5	8	−1.5	2.25
Sum				0	36.50

Using the formula for the Spearman correlation coefficient, we find the following:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{(6)(36.50)}{(15)(225 - 1)} = 0.9348$$

We interpret this rank correlation coefficient in the same way as we interpret the linear correlation coefficient. This coefficient states that there is a strong, positive correlation between the two variables.

Lesson Summary

We use the rank sum test (also known as the Mann-Whitney *v*-test) to assess whether two samples come from the same distribution. This test is sensitive to both the median and the distribution of the samples.

When performing the rank sum test, there are several different conditions that need to be met, including the population not being normally distributed, continuously distributed observations, independence of samples, the samples having greater than 5 observations, and the observations being on a numeric or ordinal scale.

When performing the rank sum test, we need to calculate a figure known as the *U*-statistic. This statistic takes both the median and the total distribution of both samples into account and is derived from the ranks of the observations in both samples.

When performing our hypotheses tests, we calculate the standard score, which is defined as follows:

$$z = \frac{U - \mu_U}{\sigma_U}$$

We use the Spearman rank correlation coefficient (also known simply as the rank correlation coefficient) to measure the strength, magnitude, and direction of the relationship between two variables from non-normal distributions. This coefficient is calculated as shown:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

12.3 The Kruskal-Wallis Test and the Runs Test

Learning Objectives

- Evaluate a hypothesis for several populations that are not normally distributed using multiple randomly selected independent samples with the Kruskal-Wallis Test.
- Determine the randomness of a sample using the runs test to access the number of data sequences and compute a test statistic using the appropriate formula.

Introduction

In the previous sections, we learned how to conduct nonparametric tests, including the sign test, the sign rank test, the rank sum test, and the rank correlation test. These tests allowed us to test hypotheses using data that did not meet the assumptions of being normally distributed or having homogeneity with respect to variance. In addition, each of these non-parametric tests had parametric counterparts.

In this last section, we will examine another nonparametric test-the Kruskal-Wallis one-way analysis of variance (also known simply as the Kruskal-Wallis test). This test is similar to the ANOVA test, and the calculation of the test statistic is similar to that of the rank sum test. In addition, we will also explore something known as the runs test, which can be used to help decide if sequences observed within a data set are random.

Evaluating Hypotheses Using the Kruskal-Wallis Test

The *Kruskal-Wallis test* is the analog of the one-way ANOVA and is used when our data set does not meet the assumptions of normality or homogeneity of variance. However, this test has its own requirements: it is essential that the data set has identically shaped and scaled distributions for each group.

As we learned in Chapter 11, when performing the one-way ANOVA test, we establish the null hypothesis that there is no difference between the means of the populations from which our samples were selected. However, we express the null hypothesis in more general terms when using the Kruskal-Wallis test. In this test, we state that there is no difference in the distributions of scores of the populations. Another way of stating this null hypothesis is that the average of the ranks of the random samples is expected to be the same.

The test statistic for this test is the non-parametric alternative to the F-statistic. This test statistic is defined by the following formula:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1)$$

where:

$N = \sum n_k$.

n_k is number of observations in the kth sample.

R_k is the sum of the ranks in the kth sample.

m is the number of samples.

Like most nonparametric tests, the Kruskal-Wallis test relies on the use of ranked data to calculate a test statistic. In this test, the measurement observations from all the samples are converted to their ranks in the overall data set. The smallest observation is assigned a rank of 1, the next smallest is assigned a rank of 2, and so on. Similar to this procedure in the rank sum test, if two observations have the same value, we assign both of them the same rank.

Once the observations in all of the samples are converted to ranks, we calculate the test statistic, H, using the ranks and not the observations themselves. Similar to the other parametric and non-parametric tests, we use the test statistic to evaluate our hypothesis. For this test, the sampling distribution for H is the chi-square distribution with $m-1$ degrees of freedom, where m is the number of samples.

It is easy to use Microsoft Excel or a statistical programming package, such as SAS or SPSS, to calculate this test statistic and evaluate our hypothesis. However, for the purposes of this example, we will perform this test by hand.

Example: Suppose that a principal is interested in the differences among final exam scores from Mr. Red, Ms. White, and Mrs. Blue's algebra classes. The principal takes random samples of students from each of these classes and records their final exam scores as shown:

TABLE 12.9:

Mr. Red	Ms. White	Mrs. Blue
52	66	63
46	49	65
62	64	58
48	53	70
57	68	71
54		73

Determine if there is a difference between the final exam scores of the three teachers.

Our hypothesis for the Kruskal-Wallis test is that there is no difference in the distributions of the scores of these three populations. Our alternative hypothesis is that at least two of the three populations differ. For this example, we will set our level of significance at $\alpha = 0.05$.

To test this hypothesis, we need to calculate our test statistic. To calculate this statistic, it is necessary to assign and sum the ranks for each of the scores in the table above as follows:

TABLE 12.10:

Mr. Red	Overall Rank	Ms. White	Overall Rank	Mrs. Blue	Overall Rank
52	4	66	13	63	10
46	1	49	3	65	12
62	9	64	11	58	8
48	2	53	5	70	15
57	7	68	14	71	16
54	6			73	17
Rank Sum	29		46		78

Using this information, we can calculate our test statistic as shown:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1) = \frac{12}{(17)(18)} \left(\frac{29^2}{6} + \frac{46^2}{5} + \frac{78^2}{6} \right) - (3)(17+1) = 7.86$$

Using the chi-square distribution, we determine that with $3 - 1 = 2$ degrees of freedom, our critical value at $\alpha = 0.05$ is 5.991. Since our test statistic of 7.86 exceeds the critical value, we can reject the null hypothesis that stated there is no difference in the final exam scores among students from the three different classes.

Determining the Randomness of a Sample Using the Runs Test

The *runs test* (also known as the *Wald-Wolfowitz test*) is another nonparametric test that is used to test the hypothesis that the samples taken from a population are independent of one another. We also say that the runs test checks the randomness of data when we are working with two variables. A *run* is essentially a grouping or a pattern of observations. For example, the sequence $++--++--++--$ has six runs. Three of these runs are designated by two positive signs, and three of the runs are designated by two negative signs.

We often use the runs test in studies where measurements are made according to a ranking in either time or space. In these types of scenarios, one of the questions we are trying to answer is whether or not the average value of the measurement is different at different points in the sequence. For example, suppose that we are conducting a longitudinal study on the number of referrals that different teachers give throughout the year. After several months, we notice that the number of referrals appears to increase around the time that standardized tests are given. We could formally test this observation using the runs test.

Using the laws of probability, it is possible to estimate the number of runs that one would expect by chance, given the proportion of the population in each of the categories and the sample size. Since we are dealing with proportions and probabilities between discrete variables, we consider the binomial distribution as the foundation of this test. When conducting a runs test, we establish the null hypothesis that the data samples are independent of one another and are random. On the contrary, our alternative hypothesis states that the data samples are not random and/or not independent of one another.

The runs test can be used with either nominal or categorical data. When working with nominal data, the first step in conducting the test is to compute the mean of the data and then designate each observation as being either above the mean (i.e., +) or below the mean (i.e., -). Next, regardless of whether or not we are working with nominal or categorical data, we compute the number of runs within the data set. As mentioned, a run is a grouping of the variables. For example, in the following sequence, we would have 5 runs. We could also say that the sequence of the data switched five times.

$$++----+++-+$$

After determining the number of runs, we also need to record each time a certain variable occurs and the total number of observations. In the example above, we have 11 observations in total, with 6 positives ($n_1 = 6$) and 5 negatives ($n_2 = 5$). With this information, we are able to calculate our test statistic using the following formulas:

$$z = \frac{\text{number of observed runs} - \mu}{\sigma}$$

$$\mu = \text{expected number of runs} = 1 + \frac{2n_1 n_2}{n_1 + n_2}$$

$$\sigma^2 = \text{variance of the number of runs} = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

When conducting the runs test, we calculate the standard z-score and evaluate our hypotheses, just like we do with other parametric and non-parametric tests.

Example: A teacher is interested in assessing if the seating arrangement of males and females in his classroom is random. He observes the seating pattern of his students and records the following sequence:

MFMMFFFFMMMFMFMMMMFFMFFMFFFF

Is the seating arrangement random? Use $\alpha = 0.05$.

To answer this question, we first generate the null hypothesis that the seating arrangement is random and independent. Our alternative hypothesis states that the seating arrangement is not random or independent. With $\alpha = 0.05$, we set our critical values at 1.96 standard scores above and below the mean.

To calculate the test statistic, we first record the number of runs and the number of each type of observation as shown:

$$R = 14 \quad M : n_1 = 13 \quad F : n_2 = 15$$

With these data, we can easily compute the test statistic as follows:

$$\mu = \text{expected number of runs} = 1 + \frac{(2)(13)(15)}{13 + 15} = 1 + \frac{390}{28} = 14.9$$

$$\sigma^2 = \text{variance of the number of runs} = \frac{(2)(13)(15)[(2)(13)(15) - 13 - 15]}{(13 + 15)^2(13 + 15 - 1)} = \frac{(390)(362)}{(784)(27)} = 6.67$$

$$\sigma = 2.58$$

$$z = \frac{\text{number of observed runs} - \mu}{\sigma} = \frac{14 - 14.9}{2.58} = -0.35$$

Since the calculated test statistic is not less than $z = -1.96$, our critical value, we fail to reject the null hypothesis and conclude that the seating arrangement of males and females is random.

Lesson Summary

The Kruskal-Wallis test is used when we are assessing the one-way variance of a specific variable in non-normal distributions.

The test statistic for the Kruskal-Wallis test is the non-parametric alternative to the F-statistic. This test statistic is defined by the following formula:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{m} \frac{R_k^2}{n_k} - 3(N+1)$$

The runs test (also known as the Wald-Wolfowitz test) is another non-parametric test that is used to test the hypothesis that the samples taken from a population are independent of one another. We use the z-statistic to evaluate this hypothesis.

On the Web

http://tinyurl.com/334e5to Good explanations of and examples of different nonparametric tests.

http://tinyurl.com/33s4h3o Allows you to enter data and then performs the Wilcoxon sign rank test.

http://tinyurl.com/33s4h3o Allows you to enter data and performs the Mann Whitney Test.

Keywords

Kruskal-Wallis test

Mann-Whitney *v*-test

Non-parametric tests

Rank correlation coefficient

Rank correlation test

Rank sum test

Run

Runs test

Sign rank test

Sign test

Spearman rank correlation coefficient

U-distribution

U-statistic

Wald-Wolfowitz test

Wilcoxon sign rank test

CHAPTER **13** **Advanced Probability and Statistics - Second Edition Resources**

Chapter Outline

13.1 Resources on the Web for Creating Examples and Activities

Disclaimer: All links here worked when this document was written.

In the Current News: Surveys, Observational Studies, and Randomized Experiments

- http://www.gallup.com The Gallup Organization's site. Frequent updating with current polls and good archive of polls conducted in last few years.
- http://www.washingtonpost.com/wp-srv/politics/polls/datadir.htm A set of links to all of the major polls (USA Today, CNN, NY Times, ABC, Gallup Poll , etc.). Maintained by *The Washington Post* .
- http://www.usatoday.com/news/health/healthindex.htm USA Today Health Index. An archive of past health stories reported in *USA Today*.
- http://www.publicagenda.org Recent survey results for hot public issues (abortion, crime, etc.).
- http://www.pollingreport.com A collection of recent poll results on business, politics, and society from many different sources.
- http://sda.berkeley.edu/ SDA is a set of programs for the documentation and web-based analysis of survey data.

Resources by Teachers for Teachers

- http://www.herkimershideaway.org/ Herkimer's Hideaway, by Sanderson Smith, Department of Mathematics, Cate School, Carpinteria, California. Click on AP Statistics. Contains many ideas for projects and activities.
- http://www.causeweb.org/repository/StarLibrary/activities/ Statistics Teaching and Resource Library. Started in Summer, 2001. The STAR Library collection is peer-reviewed by an editorial board. Mission is "to provide a peer-reviewed journal of resources for introductory statistics teachers that is free of cost, readily available, and easy to customize for the use of the teacher."
- http://www.dartmouth.edu/~chance/chance_news/news.html Chance News: A newsletter of recent (mostly United States) media items useful for class discussion.
- http://exploringdata.net/ This website provides curriculum support materials for teachers of Introductory Statistics.
- http://mathforum.org/workshops/usi/dataproject/usi.genwebsites.html This website has a variety of links to data sets and websites that provide support, ideas, and activities for teachers of statistics.

Survey Methodology

- http://www.publicagenda.org Nice discussions of issues connected to surveys on hot public issues (abortion, crime, etc.). In particular, click on 'Red Flags' for each issue to see examples of how question wording, survey timing, and so on affect survey results.
- http://whyfiles.org/009poll/math_primer.html University of Wisconsin Why Files on Polling. Discusses basic polling principles.

Data Sets

- http://lib.stat.cmu.edu/DASL/ Carnegie Mellon Data and Story Library (DASL). Data sets are cross-indexed by statistical application and research discipline.
- http://sda.berkeley.edu/archive.htm General Social Survey archive and on-line data analysis program at the University of California at Berkeley.
- http://www.cdc.gov/nchs/fastats/default.htm FedStats Home Page. "The gateway to statistics from over 100 U.S. Federal agencies."

- http://www.lib.umich.edu/govdocs/stats.html University of Michigan Statistical Resources Center. Huge set of links to government data sources.
- http://dir.yahoo.com/Social_Science/Social_Research/Data_Collections/ Yahoo!'s directory of social science data collections.
- http://dir.yahoo.com/Reference/Statistics/ Yahoo!'s directory of statistical data collections.

Miscellaneous Case Studies and Data Resources

- http://www.flmnh.ufl.edu/fish/Sharks/ISAF/ISAF.htm Shark Attacks-International Shark Attack File. Shark attack statistics, including special sections for the great white shark and shark attacks on divers. (Thanks to Tom Hettmansperger of Penn State for pointing out this site.)
- http://www.DrugAbuseStatistics.samhsa.gov/ Drug Abuse Statistics from Substance Abuse and Mental Health Services Administration, Office of Applied Statistics.

Java and JavaScript Activities

- http://onlinestatbook.com/rvls.html The Rice University Virtual Lab in Statistics (David Lane). Includes simulations, activities, case studies, and many interesting links.
- http://www-stat.stanford.edu/~susan/surprise/ Probability applets. One illustrates the birthday problem in a fun way.

Advanced Placement Statistics Listserve Archives

- http://mathforum.org/kb/forum.jspa?forumID=67 Searchable archive of thousands of e-mail messages contributed by high school and college statistics teachers about topics as diverse as studies in the news to where to find test questions.

Journal of Statistics Education

- http://www.amstat.org/publications/jse/ Free online journal sponsored by the American Statistical Association. Includes articles about teaching statistics, interesting data sets, and current articles in the news for discussion.