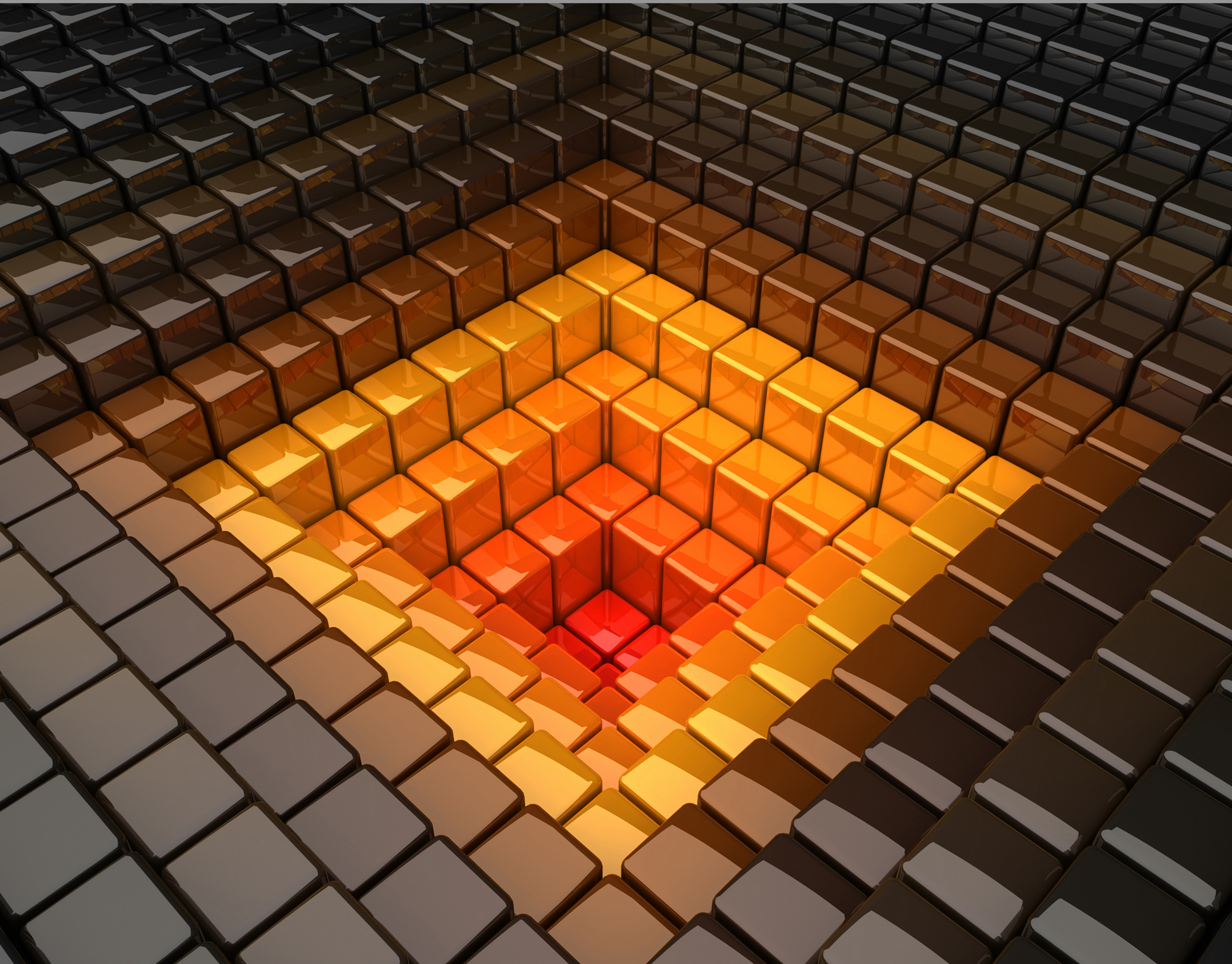
CK-12

flexbook
next generation textbooks

Probability and Statistics (Advanced)



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.

flexbook
next generation textbooks

Contents

Chapter 1

An Introduction to Analyzing Statistical Data

1.1 Definitions of Statistical Terminology

Learning Objectives

- Distinguish between quantitative and categorical variables.
- Distinguish between continuous and discrete 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, students will be introduced to some basic statistical 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 19^{th} 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

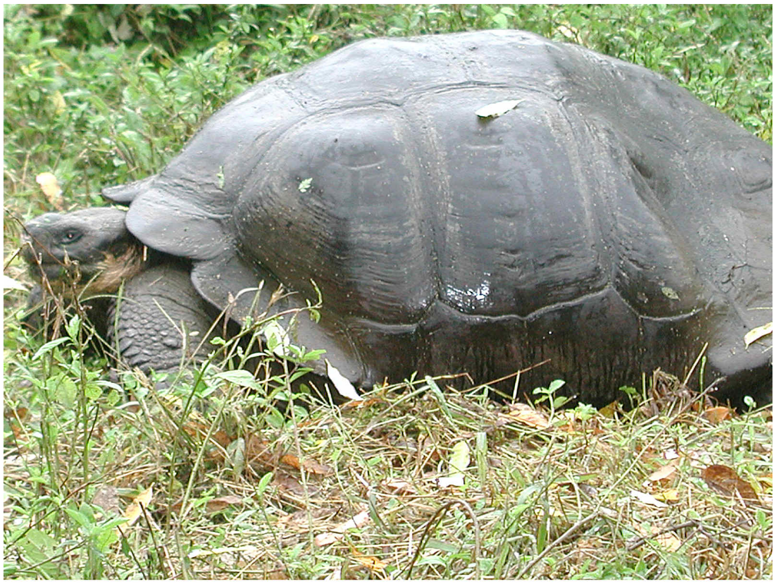


Figure 1.1: Galapagos Tortoise on Santa Cruz. (1)

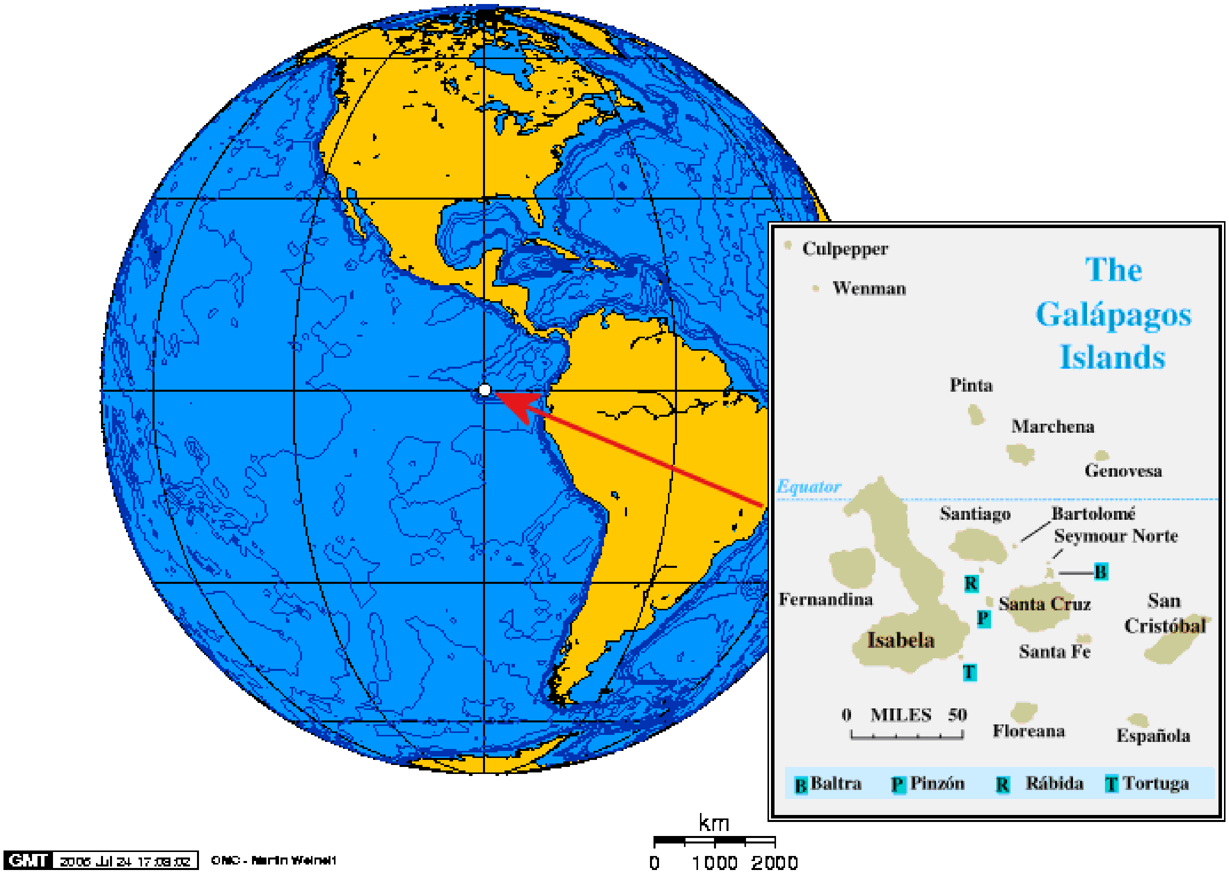


Figure 1.2: Galapagos Map. (5)

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. Settlers brought in domesticated animals like goats and pigs that destroyed the tortoise's 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.

Table 1.1: **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.**

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	$1,139$	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	2574	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

Figure 1.3: Tortoise With Dome-shaped Shell on Santa Cruz Island. (2)

Classifying Variables

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

A researcher studying Galapagos Tortoises would be interested in collecting information about different characteristics of the tortoises. Those characteristics are called **variables**. Each column of the previous figure contains a variable. In the first two columns, the tortoises are grouped according to the island (or volcano) where they live and the scientific names for each 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** (some statisticians use the word qualitative).

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 **numerical** (or quantitative). Note that *repatriation* is the process of raising tortoises and releasing them into the wild when grown to avoid local predators that prey on hatchlings. The figure below explains the remaining variables in the previous figure and labels them as categorical or numerical.

Table 1.2: **Explanation of Remaining Variables.**

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 tortoise 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	The number of tortoises that were captured and marked by scientists to study their health and assist in estimating the total population.	Numerical
Number of Individuals Repatriated	There are two tortoise breeding centers on the islands. Through those programs, many tortoises have been raised and then reintroduced into the wild.	Numerical

Variables can be further classified as either **discrete** or **continuous**. A **discrete** numerical variable can only have values at specific values. For example, the number of tortoises reintroduced into the wild must be a whole number. (How would you introduce half of a tortoise?!) But don't get the wrong idea! It is possible for a variable to have fractional values and still be discrete. Shoe sizes, for example, are discrete as their values occur at set increments: $7, 7\frac{1}{2}, 8, 8\frac{1}{2}$ etc... You should also know that **all** categorical variables are discrete.

On the other hand, the population density, which means the average number of tortoises per square kilometer, could be any positive number. This is an example of a **continuous variable**. Even though the numbers in the table have been rounded, the number of square kilometers can, in theory, be any value depending on the size of the habitat. The average (or mean) rainfall in a city is a continuous variable. Within a reasonable range of values, all amounts of rainfall are possible. However, someone measuring that rainfall may only measure to the nearest centimeter, and it might then be considered discrete. Practically speaking, anytime you measure a variable that can only be measured in discrete values, you are effectively using a variable that is not truly continuous.

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, how would you count ALL the tortoises of one species? 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 (and maybe even impossible!!) 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 (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. When a researcher wanted to find an estimate for the population of a species of tortoise, she would go into the field and locate and mark a number of tortoises. She would then use statistical techniques that we will discover 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**. The number of tortoises in the sample, or any other number that describes the individuals in the sample (like their length, or weight, or age), is called a **statistic**. In general, each **statistic** is an estimate of a **parameter**, whose value is not known exactly.

In the **Table** 1.3, are the actual data from the species of tortoise found on the Volcano Darwin, on Isabela Island. (**Note:** the word "data" is the plural of the word "datum", which means the result of a single measurement.) The number of captured individuals is a statistic as it deals with the sample. The actual population is a parameter that we are trying to estimate.

Table 1.3: **Tortoise Data for Darwin Volcano, Isabela Island.**

Number of Individuals Captured	Population Estimate	Population Estimate Interval
160	818	561 – 1075

Errors in Sampling

Unfortunately, there is a downside to using sampling. We have to accept that estimates using a sample have a chance of being inaccurate or even downright wrong! This cannot be avoided unless we sample the entire population. You can see this in the next figure. The actual data not only includes an estimate, but also an interval of the likely true values for the population parameter. The researcher has to accept that there could be variations in the sample due to chance which lead to changes in the population estimate. A statistician would not say that the parameter is a specific number like 915, but would most likely report something like the following:

"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. In general, the potential difference between the true parameter and the statistic obtained from using a sample is called **sampling error**. It is also possible that the researchers made mistakes in their sampling methods in a way that led to a sample that does not accurately represent the true population. For example, they 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 generally are of two types, **numerical** or **quantitative**, and **categorical** or **qualitative**.

Quantitative variables can be further categorized as those that can only have set, integral values, or **discrete variables**, and those that can be a range of values, or **continuous variables**.

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 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

1. How do we summarize, display, and compare categorical and numerical data differently?
2. What are the best ways to display categorical and numerical data?
3. Is it possible for a variable to be considered both categorical and numerical?

4. How can you compare the effects of one categorical variable on another or one quantitative variable on another?

Review Questions

1. In each of the following situations, identify the population, the units, each variable, and tell if the variable is categorical or quantitative. If it is quantitative, then identify it further as either discrete or continuous.
   (a) A quality control worker with Sweet-tooth Candy weighs every 100^{th} candy bar to make sure it is very close to the published weight.
      i. POPULATION:
      ii. UNITS:
      iii. VARIABLE:
      iv. TYPE:
   (b) Doris decides to clean her sock drawer out and sorts her socks into piles by color.
      i. POPULATION:
      ii. UNITS:
      iii. VARIABLE:
      iv. TYPE:
   (c) A researcher is studying the effect of a new drug treatment for diabetes patients. She performs an experiment on 200 randomly chosen individuals with Type II diabetes. Because she believes that men and women may respond differently, she records each person's gender, as well as their change in sugar level after taking the drug for a month.
      i. POPULATION:
      ii. UNITS:
      iii. VARIABLE 1:
      iv. TYPE:
      v. VARIABLE 2:
      vi. TYPE:
2. In Physical Education class, the teacher has them count off by two's to divide them into teams. Is this a categorical or quantitative variable?
3. A school is studying their students' test scores by grade. Explain how the characteristic "grade" could be considered either a categorical or a numerical variable.

Review Answers

1. (a) i. POPULATION: All candy bars made by the company
      ii. UNITS: each individual candy bar
      iii. VARIABLE: weight of the candy bars

    iv. TYPE: Quantitative. It is continuous. The weights could be any weight reasonably close to the desired weight due to variation in the number and weight of individual candies. Note: if the worker decided to sort the candy bars as acceptable, too light, or too heavy, the same scenario could include a categorical variable.

(b) i. POPULATION: All of Doris' socks
    ii. UNITS: each sock
    iii. VARIABLE: color of socks
    iv. TYPE: Categorical

(c) i. POPULATION: All diabetes sufferers
    ii. UNITS: each individual diabetes patient
    iii. VARIABLE 1: change in sugar level ($+$ or $-$)
    iv. TYPE: Quantitative, continuous
    v. VARIABLE 2: gender
    vi. TYPE: Categorical

2. An argument could be made that by definition, it could be a discrete quantitative variable, but this is really a categorical variable. Students are either on one team or another. The use of the digits "1" and "2" to put the students in groups has no significant numerical meaning. The teacher could have just as easily had the students say "blue" and "red."
3. This variable could be easily described as categorical, as students are in one of the four classes (Freshman, Sophomore, Junior, Senior), but it could also be appropriate to think of those classes as grades $9-12$. The numbers do signify order and therefore could be considered to have numerical significance. If so, it would be a discrete numerical variable.

Further Reading

- onlinestatbook.com/
- en.wikipedia.org/wiki/Gal%C3%A1pagos_tortoise
- pes.ucf.k12.pa.us/Themes/Endangered%20Animals/pages/gtortoise5.htm
- Charles Darwin Research Center and Foundation: www.darwinfoundation.org

1.2 An Overview of Data

Learning Objectives

- Understand the difference between the levels of measurement: nominal, ordinal, interval, and ratio.
- Identify the general elements that characterize a study.

- Understand the fundamentals of experimental design.
- Understand the basic concept of measures of center and variation and their uses for statistical analysis.

Introduction

This lesson is an overview of the basic considerations involved with collecting and analyzing data. All of these concepts will be examined in greater detail in later chapters, but it is important that students are familiar with the ideas before examining them in greater detail.

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 when examining the information that is collected for a variable, called the **levels of measurement**. This widely accepted (though not universally used) theory was first proposed by the American psychologist, Stanley Smith Stevens in 1946 (see links at end of this section). According to Stevens' theory, the four levels of measurement are:

- nominal
- ordinal
- interval
- ratio

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

Nominal Measurement

It is easiest to think of nominal measurement in terms of discrete, categorical variables. This is the type of measurement in which the values of the variable are names, and not numerical at all. The names of the different species of Galapagos tortoises would be a nominal measurement.

Ordinal Measurement

This type of measurement involves collecting information in which the order is somehow significant. The name of this level is derived from the use of ordinal numbers for ranking ($1^{st}, 2^{nd}, 3^{rd}$, 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 1^{st} and 2^{nd} largest tortoise populations by species may differ by a few thousand individuals, while the 7^{th} and 8^{th} may only differ by a few hundred.

Interval Measurement

In interval measurement, we add to the ranking of ordinal measurement by collecting data in which there is significance to the distance between any two values. An example commonly cited for interval measurement is temperature (either Celsius or Fahrenheit degrees). A change of 1 degree is the same if the temperature goes from $0^\circ C$ to $1^\circ C$, as it is when the temperature goes from $40^\circ C$ to $41^\circ C$. Additionally, there is meaning to the values between the ordinal numbers (i.e. $\frac{1}{2}$ a degree can be interpreted)

Ratio Measurement

Ratio measurement gets its name from the fact that a meaningful fraction (or ratio) can be constructed with a ratio variable. Ratio is the deepest, most meaningful level of measurement, and consequently, the most useful. 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. In the temperature scale of the previous example, $0^\circ C$ 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 $40^\circ C$ for example, is not really "twice" as hot as $20^\circ C$. 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 3300 really is approximately three times as large as one with a population near 1100.

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 (which they frequently do):

Nominal: We could collect information about the students' gender, the town or sub-division in which they live, race, or political opinions.

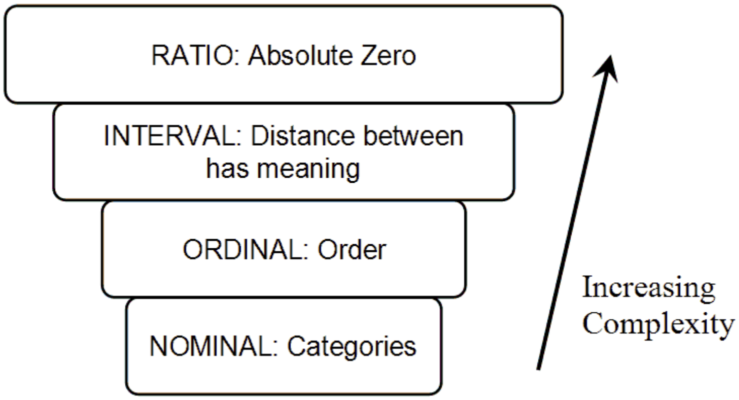
Ordinal: If we collect data about the students' year in school, we are now ordering that data numerically ($9, 10, 11$ or 12^{th} grade).

Interval: If we gather data for students' SAT math scores, we have interval measurement.

There is no absolute 0, as SAT scores are scaled. The ratio between two scores is also meaningless (i.e. a person who scores a 600 did not necessarily do “twice as well” as a student who scored a 300).

Ratio: Data about a student’s age, height, weight, and grades will be measured on the ratio level. In each of these cases there is an absolute zero that has real meaning. Someone who is 18 really 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.



Observational Studies

Some of the other famous residents of the Galapagos that have provided scientists with a wealth of information and opportunities for study are the so-called Darwin’s finches. Each of the numerous species of finches has developed special adaptations that allow it to survive in a particular area. There are ground finches, tree finches, cactus finches, medium-billed, small-billed, and large-billed finches, just to name a few. One particular variety has even learned to use a stick as a tool to dig for bugs. To the untrained observer, it is almost impossible to tell them all apart, and on a visit to the islands you will see them everywhere!

Two researchers from Princeton University, Peter and Rosemary Grant, spent over 30 years studying the adaptations of finches to environmental conditions on a small island in the Galapagos called Daphne Major.

The Grants’ spent up to 6 months a year on this “rock” documenting how species with certain beak size and shape would thrive in years when vegetation that suited those species grew well, and would dramatically decrease in numbers in years when that vegetation was sparse. This type of long-term approach to collecting data by making detailed observations

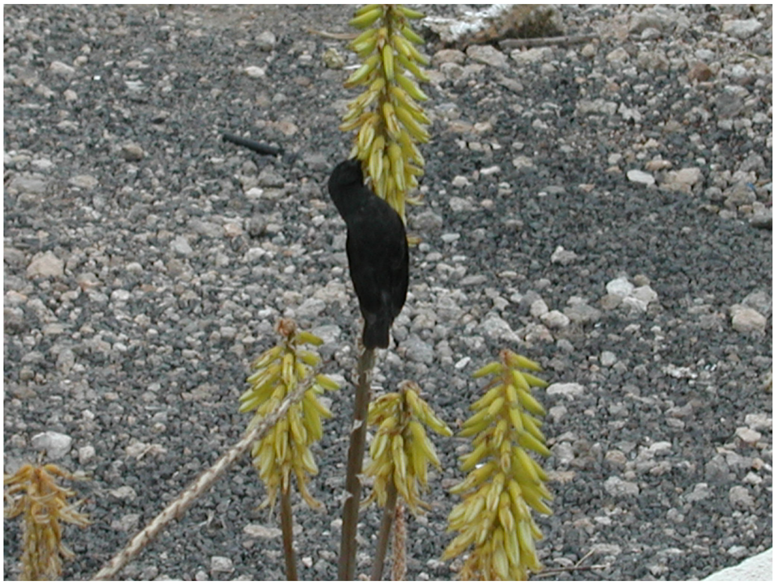


Figure 1.4: Small Ground Finch, Santa Cruz, Galapagos Islands. (7)

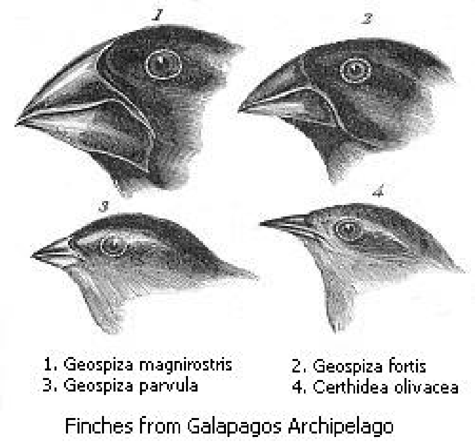


Figure 1.5: Darwin's Finches (6)

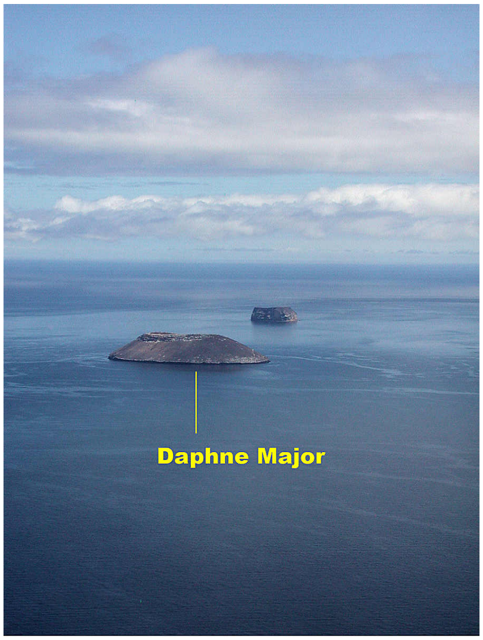


Figure 1.6: Daphne Major, Galapagos Islands. (4)

is called an **observational study**, and is a widely used method of gathering data. In an observational study, the researcher observes the population of interest and records the results without making an attempt to control the outcomes.

Another famous observational study in the United States is the Framingham Heart Study. Researchers have followed the lives of people from the town of Framingham Massachusetts for 60 years, and the information gathered has led to many of the current approaches to treating and preventing heart disease. This type of long-term observational study in which the same group of subjects is observed for very long periods of time is also called a **longitudinal study**.

Experiments

The other widely used method for conducting research is called an **experiment**. In an experiment, the researcher imposes a treatment on a group of subjects in an effort to determine a “cause and effect” relationship between variables. While observational studies could appear to show a relationship between diet and heart disease, for example, there could be another factor that is actually causing an individual’s heart condition. An experiment designed to investigate this relationship might take two groups of similar subjects, impose different diets on each group of those subjects, and then record any differences in the condition of their hearts. What makes this difficult, and in some instances impossible, is that the researcher would then need to make sure that anything else that might have an influence on a subject’s heart health (e.g. exercise, genetics, stress level) is controlled, or exactly the same for each individual in the study. One of the ways that statisticians insure this control is by randomly assigning subjects and treatments, thereby using the laws of probability to help guarantee the validity of the results. Designing experiments can be difficult and costly, but they are the only way to establish meaningful and reliable cause and effect relationships. We will study the elements of designing experiments in more detail in later chapters.

Measures of Center and Spread

Let us assume that you have collected some data on one of the various levels of measurement (nominal, ordinal, interval, or ratio) using a statistically valid procedure (observational study or experiment). How do you summarize this information? One of the most important tools for summarizing data is to display it visually, and the various methods for doing so will be covered in later chapters. If we want to use one number or value to summarize the data, we can look at where the data is centered. Data measured at different levels can be characterized by different summaries. Look back at the Tortoise data. This data was collected through an observational study. The variable "Climate Type" is a categorical variable that has been measured at the nominal level. The easiest way to summarize this variable is to identify the most common value (**mode**), which is "humid." Variables that are measured at the ratio level, like "population density," we might find the **average** (**mean**) or the middle number (**median**) in the data to summarize it.

Another important element of a data set is how it is spread. In the tortoise population estimate data, the numbers per species range from 6320, down to 1, or a spread of approximately $6,000$ tortoises. However, the population of the Alcedo tortoises is much larger than the other species, so this number might not give a true indication of how most of the other populations vary. We have other measures that might help shed some light on the spread of the typical tortoise species,such as the **interquartile range** and the **standard deviation**, which we will cover in detail in the following lessons.

Lesson Summary

Data can be measured at different levels depending on the type of variable and 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 includes an absolute zero value.

Statisticians and researchers use two main techniques to form important conclusions about the relationships between variables. An **observational study** is when a researcher observes the subjects in the real world without manipulating them. An **experiment** is the way to establish true cause-and-effect relationships. It involves the researcher imposing some randomly assigned treatment(s) on the subjects in an effort to isolate the effect of a single variable.

In order to summarize a set of data, we often look to a single quantity to describe where it is centered. There are various measures that are used for this summary, including the **mean**, **median**, and **mode**. These will be covered in detail in later sections, but they are generally referred to as **measures of center**. Similarly, for information about how the data is spread out, we investigate **measures of spread** that include the **range**, **interquartile range**, and **standard deviation**.

Points to Consider

1. How do we summarize, display, and compare data measured at different levels?
2. What are the differences between an observational study and an experiment?
3. What are the advantages/disadvantages of observational studies and experiments?
4. How do you determine which measure of center or spread best describes a particular data set?

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 is their favorite thing to eat.
   (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 3 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 3, then 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. Here is a picture of poor George!

Figure 1.7: *Lonesome George*, the Last Pinta tortoise, Charles Darwin Research Station, Santa Cruz, Galapagos Islands. (3)

4. In each of the following situations, identify if it is an observational study or an experiment.

   (a) In an attempt to determine if students prefer bottled water to tap water, you set up a table in the cafeteria at lunchtime and have students sample some of each and ask them which they prefer.
   (b) Researchers collect data over 15 years about 100 sets of identical twins to see how their personalities develop similar or different characteristics.
   (c) Cloned mice are put into different colored cage environments to see if there is an effect on their temperaments.
   (d) Researchers find that babies who were exposed to lead paint have a high risk of brain damage.

Review Answers

1. (a) Ordinal
   (b) Nominal
   (c) Interval. Even though Celsius has a "0", this is a completely arbitrary decision to set the freezing point of water and not the "absence" of temperature.
   (d) Ratio. The Kelvin scale is based on an absolute zero, the theoretical temperature at which molecules stop moving.
2. (d) The levels of measurement theory is a useful tool to help categorize data, but like much of statistics, it is not an absolute "rule" that applies easily to every situation and several statisticians have pointed out some of the difficulties with the theory. See: http://en.wikipedia.org/wiki/Level_of_measurement
3. (d) Population densities are certainly measured up to the interval level as there is meaning to the values and distance between two observations. To decide if it is measured at the ratio level, we need to establish a meaning for absolute zero. In this case, it would be 0 individuals per km^2. This is possible and indeed represents the extinct populations.
4. (a) This is an experiment as each subject is drinking both waters (the imposed treatment). However, it will have to be designed properly. Students should not know which water is bottled and which is tap (this is called a "blind" experiment) and they should be randomly assigned the order in which they drink the water. Other conditions such as the appearance, amount, and temperature would also need to be tightly controlled.
   (b) Observational study.
   (c) Experiment. The research is imposing a treatment (different color rooms) on the mice.
   (d) Observational Study. It would be unacceptable to intentionally expose a baby to potentially harmful substances. The dangers of lead paint were discovered through years of careful observational studies.

Further Reading

- 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
- Framingham Heart Study: http://en.wikipedia.org/wiki/Framingham_Heart_Study

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 $n-$percent **trimmed mean.**
- Calculate the **midrange**, **weighted mean**, **percentiles**, and **quartiles.**

Introduction

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

Mode, Mean, and Median

In the last lesson, you learned that it makes sense to summarize a data set by identifying a value around which the data is centered. Three commonly used statistics that quantify the idea of center are the mode, median and mean.

Mode

The mode is defined as the most frequently occurring number in a data set. While many elementary school children learn the mode as their first introduction to measures of center, as you delve deeper into statistics, you will most likely encounter it less frequently. The mode really only has significance for data measured at the most basic of levels. The mode is most useful in situations that involve categorical (qualitative) data that is 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 is 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 a large number of students in our class have 2 children in their home.

Notice how careful we are to NOT apply this to a larger population and assume that this will be true for any population other than our class! In a later chapter, you will learn how to correctly select a sample that could represent a broader population.

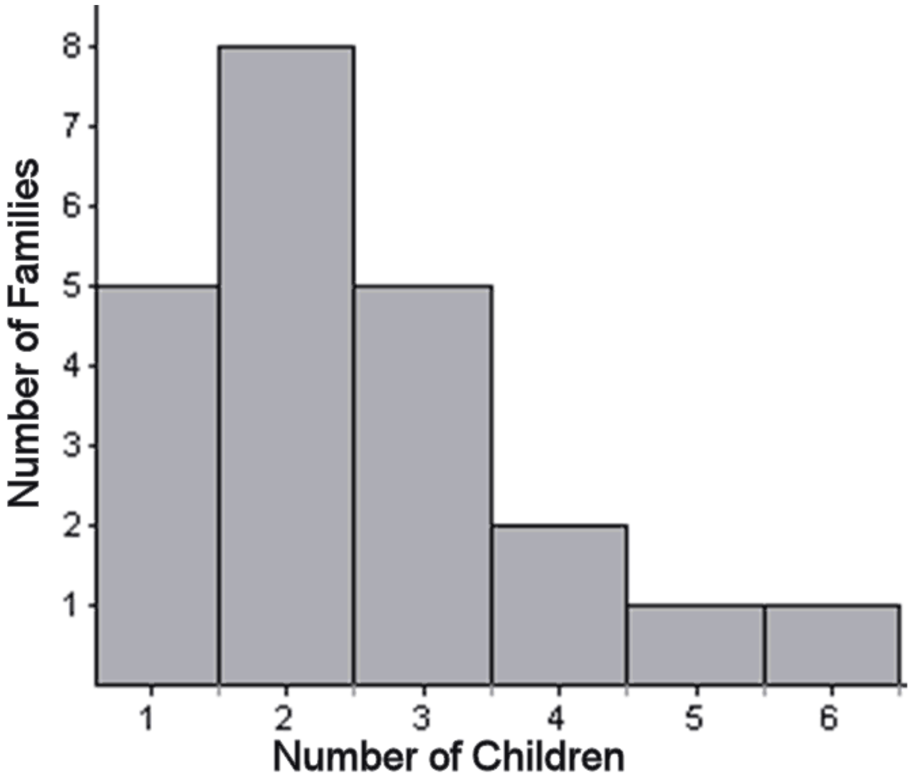
Two Issues with the Mode

1. If there is more than one number that is the most frequent than the mode is usually both of those numbers. For example, if there were seven 3−child households and seven with 2 children, we would say that the mode is, "2 and 3." When data 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 those cases, we would most likely search for a different statistic to describe the center of such data.
2. If each data value occurs an equal number of times, we usually say, "There is no mode." Again, this is a case where the mode is not at all useful in helping us to understand the behavior of the data.

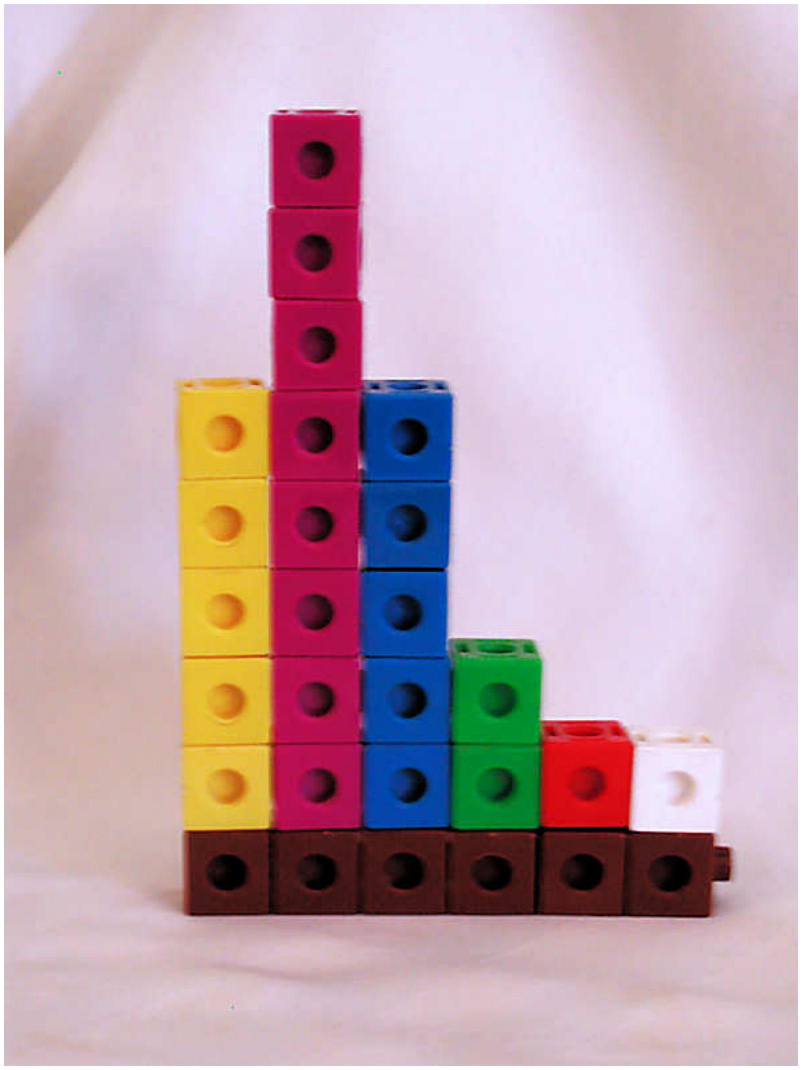
Do You Mean the Average?

You are probably comfortable calculating averages. The average is a measure of center that statisticians call the **mean**. Most students learn early on in their studies that you calculate the mean by adding all of the numbers and dividing by the number of numbers. While you are expected to be able to perform this calculation, most real data sets that statisticians deal with are so large that they very rarely calculate a mean by hand. It is much more critical that you understand ***why*** the mean is such an important measure of center. The mean is actually 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.

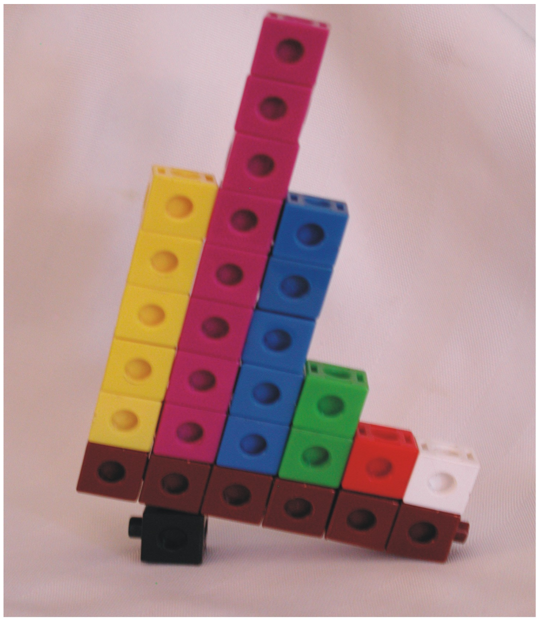


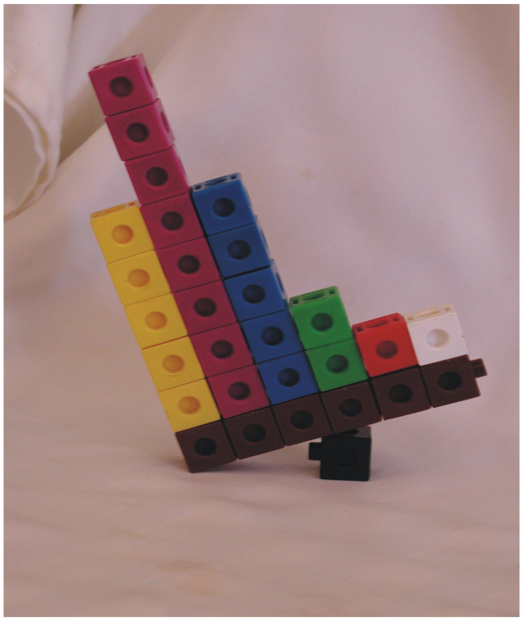
If you have snap cubes like you used to use in elementary school, you can make a physical model of the graph, using one cube to represent each student's family and a row of six cubes at the bottom to hold them together like this:

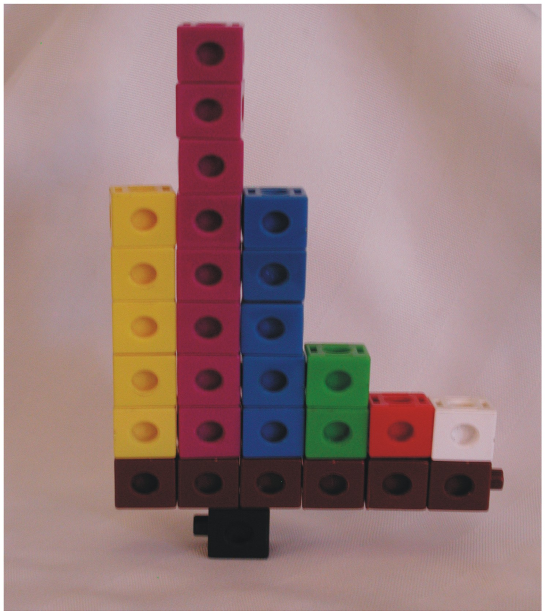


There are 22 students in this class and the total number of children in all of their houses is 55, so the mean of this data is $55 \div 22 = 2.5$. Statisticians use the symbol $\overline{X}$ to represent the mean when X is the symbol for a single measurement. It is pronounced "x bar."

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, and while one placed at 2 causes the graph to tip right. However, if you place the block at about 2.5, it balances perfectly!







Technology Note: Use the TI-83/84, and Mean it!

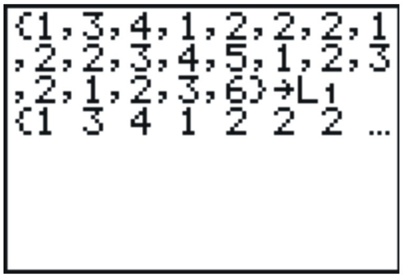
As was already mentioned, once you understand how to calculate a mean, and unless you need practice with your arithmetic skills, you rarely calculate them by hand. Here is how to calculate a mean with the TI-83/4 family of graphing calculators.

Step 1: Entering the data

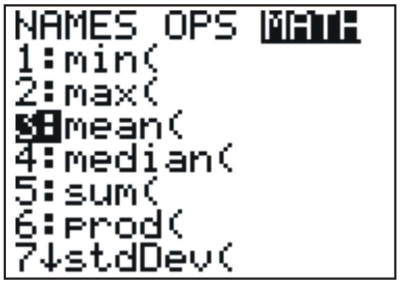
On the home screen, press [**2nd**] [**{**], then enter the data separated by commas. When you have entered all the data, press [**2nd**] [**}**] [**sto**] [**2nd**] [**L1**] [**enter**]. You will see the screen on the left below:

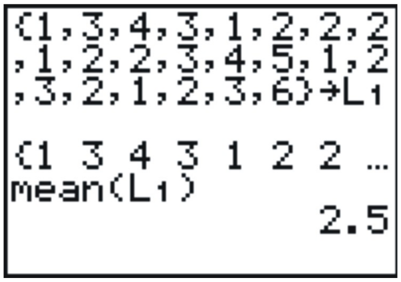
$$1, 3, 4, 3, 1, 2, 2, 2, 1, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 6$$

Step 2: Computing the mean



```
{1,3,4,1,2,2,2,1
,2,2,3,4,5,1,2,3
,2,1,2,3,6}→L1
{1 3 4 1 2 2 2 …
```





```
{1,3,4,3,1,2,2,2
,1,2,2,3,4,5,1,2
,3,2,1,2,3,6}→L1

{1 3 4 3 1 2 2 …
mean(L1)
                2.5
```

On the home screen, press [**2nd**] ***'[LIST]'*** to enter the list menu, press ([**rightarrow**]) once to go to the MATH menu (the middle screen above), and either arrow down or choose 3 for the mean. Finally, press [**2nd**] [**L1**] [)] to insert L1 and press [**enter**] (see the screen on the right above).

Right Down the Middle: The Median

The median is simply the middle number in a set of data. Think of 5 students seated in a row in statistics class:

Aliyah Bob Catalina David Elaine

Which student is sitting in the middle? If there were only four students, what would be the middle of the row? These are the same issues you face when calculating the numeric middle of a data set using the median.

Let's say that Ron has taken five quizzes in his statistics class and received the following grades:

$$80, 94, 75, 90, 96$$

Before finding the median, you must put the data in order. The median is the numeric 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. Notice that just by coincidence, this was also the third quiz that he took, but this will usually not be the case.

Of course, when there is an even number of numbers, there is no true value in the middle. In this case we take the two middle numbers and find their mean. If there are four students sitting in a row, the middle of the row is halfway between the second and third students.

Example

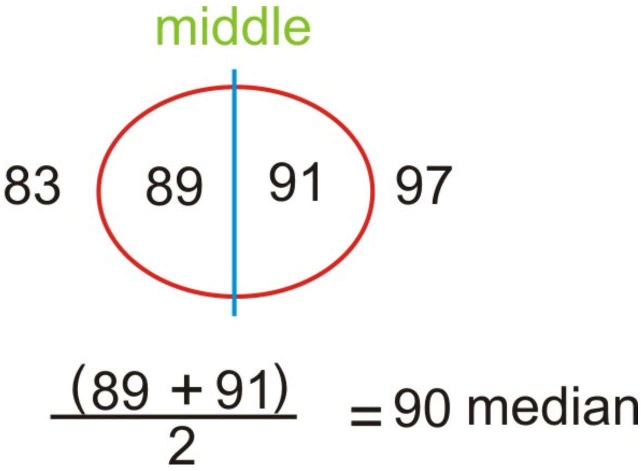
Take Rhonda's quiz grades:

$$91, 83, 97, 89$$

Place them in numeric order:

$$83, 89, 91, 97$$

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



Mean vs. Median

Both the mean and the median are important and widely used measures of center. So you might wonder why we need them both. There is an important difference between them that can be explained by the following example.

Let's say that you get 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 a 64! What would the median be? 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. If you divide a much smaller sum by 3, the mean will also be much smaller.

Outliers and Resistance

So, why are the mean and median so different in this example? It is 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. Remember the balancing point example? If you created another number that was far away, you would be forced to move the block toward it to make it stay balanced.

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. For example, in 2005 the CEO of Yahoo, Terry Semel, was paid almost $231 million, see http://www.forbes.com/static/execpay2005/rank.html. 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.

Population Mean vs. Sample Mean

Now that we understand some basic concepts about the mean, it is important to be able to represent and understand the mean symbolically. When you are calculating the mean as a statistic from a finite sample of data, we call this the sample mean and as we have already mentioned, the symbol for this is $\overline{X}$. Written symbolically then, the formula for a sample mean is:

$$\bar{x} = \frac{\sum(x_1 + x_2 + \cdots + x_n)}{n}$$

You may have remembered seeing the symbol $\sum$ before on a calculator or in another mathematics class. It is called "sigma," the Greek capital S. In mathematics, we use this symbol as a shortcut for "the sum of". So, the formula is the sum of all the data values (x_1, x_2, etc.) divided by the number of observations (n).

Recall that the mean of an entire population is a parameter. The symbol for a population mean is another Greek letter, μ. It is the lowercase Greek m and is called "mu" (pronounced "mew", like the sound a cat makes). In this case the symbolic representation would be:

$$\mu = \frac{\sum(X_1 + X_2 + \cdots + X_n)}{N}$$

The formula is very much the same, because we calculate the mean the same way, but we typically use capital X for the individuals in the population and capital N to represent the size of the population.

In general, statisticians say that $\overline{x}$, the mean of a portion of the population is an estimate of μ, the mean of the population, which is usually unknown. In this course you will learn to determine how good that estimate is.

Other Measures of Center

There are many other lesser-known measures of center that can prove useful in describing certain data sets. We will highlight a few of them in this section.

Midrange

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

In a previous example we used the following data from Ron's grades:

$$75, 80, 90, 94, 96$$

The midrange would be:

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

One of the reasons that the midrange is not commonly used is that it is only based on two values of the data set, and not just any two, but the values that are most likely to be outliers!

It would be like basing your class grade on only two assessments and ignoring all the other work you may have done. Even if it works out as a higher grade for you, much of your accomplishments would be meaningless!

Trimmed Mean

Remember 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 numbers that remain.

Let’s go back to Ron’s grades again:

$$75, 80, 90, 94, 96$$

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

$$\begin{gathered}\cancel{75}, 80, 90, 94, \cancel{96}\\ \frac{(80+90+4}{3} = 88\end{gathered}$$

n% Trimmed Mean

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, you would remove the specified percent of the number of values from the data, half on each end. For example, in a data set that contained 100 numbers, if a researcher wanted to calculate a 10% trimmed mean, she would need to remove 10% of the data, or 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.

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:

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

With 22 values, 10% of them 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 we cannot remove exactly 10% then you either have to remove more or less. 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 when some of the data values are counted frequently. The most common type of weight to use is the **frequency**, which is the number of times each number is observed in the data. The calculation gives the same result as the standard mean, but each observed data point is multiplied by its weight first, then the total sum is calculated and the result is divided by the sum of the weights.

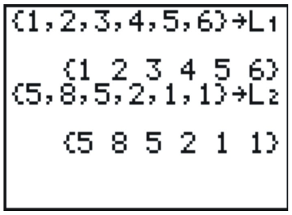
When we calculated the mean for the children living at home, we could have used a weighted mean calculation. The calculation would look like this:

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

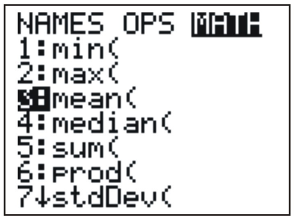
Technology Note: Weighted Means on the TI 83 or 84 Graphing Calculator

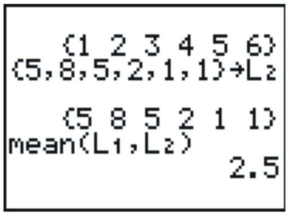
Weighted means are easy to calculate using a graphing calculator. We can use list $L1$ for the number of children, and in list $L2$ we will enter the frequencies, or weights.

Enter the data as shown in the left screen below:



```
{1,2,3,4,5,6}→L1
   {1 2 3 4 5 6}
{5,8,5,2,1,1}→L2
   {5 8 5 2 1 1}
```





```
   {1 2 3 4 5 6}
{5,8,5,2,1,1}→L2
   {5 8 5 2 1 1}
mean(L1,L2)
             2.5
```

For weighted means, we use the same procedure, but enter the two lists in the mean computation. Press [**2nd**] ***’[LIST]***’ to enter the list menu, press the left arrow [**leftarrow**] to go to the math menu (the middle screen above), and either arrow down or choose 3 for the mean. Finally, press [**2nd**] Ω [**comma**] ***’[LIST]***’ [)] [**enter**] and you will see the screen on the right above. Note that the mean is 2.5, as before.

Percentiles and Quartiles

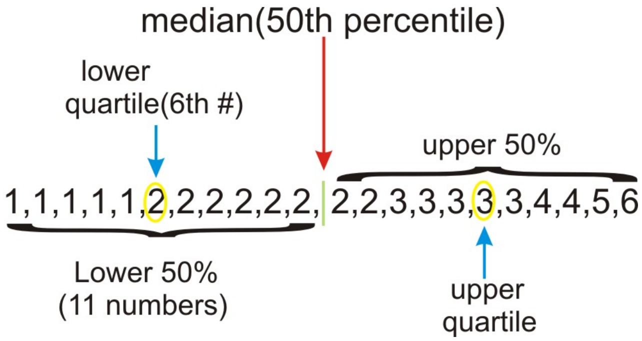
A **percentile** is a statistic that identifies the percentage of the data that is *less* than the given value. The most commonly used percentile is the median. Because it is in the numeric middle of the data, half of the data is below the median. Therefore, we could also call the median the 50^{th} **percentile**. A 40^{th} percentile would be a value in which 40% of the numbers are less than that observation. Your first exposure to percentiles was most likely as a baby! 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 70^{th} percentile is taller than 70% of the children of their same age.

Two very commonly used percentiles are the 25^{th} and 75^{th} percentiles. Because they divide the data into quarters (when taken together with the median), they are referred to as the **lower and upper quartiles**. They are sometimes abbreviated Q_1 and Q_3. A quartile divides the data into 4 approximately equal groups. Technically, the median is a “middle” quartile and is sometimes referred to as Q_2. Some also refer to the minimum value in a data set as Q_0 and the maximum as Q_4.

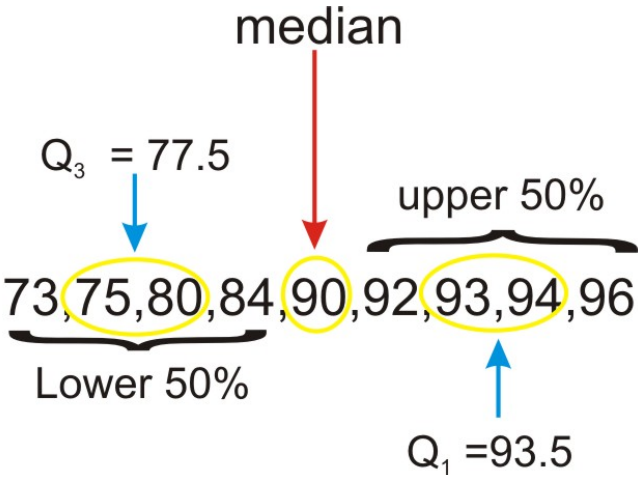
Returning to a previous data set:

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

Recall that the median (50^{th} 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 numbers in each half. If there were an even number of numbers, then we would follow the procedure for medians and average the middle two numbers of each half. Look at the following set of data:

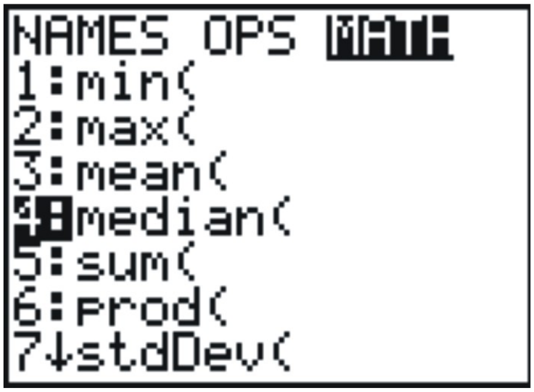


The median in this set is 90. Because it is the middle number, it is not technically part of either the lower or upper halves of the data, so we do not include it when calculating the quartiles. However, not all statisticians agree that this is the proper way to calculate the quartiles in this case. 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 those topics. To read more about some alternate methods for calculating quartiles in certain situations, see the following website:

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

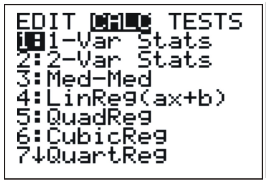
Technology Note: Medians and Quartiles on the Graphing Calculator

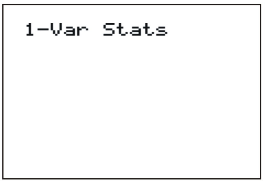
The median and quartiles can also be calculated using the graphing calculator. You may have noticed earlier that median is available in the **MATH** submenu of the [**LIST**] menu (see below).



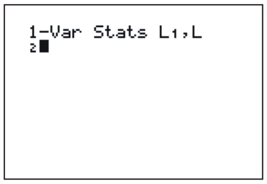
While there is a way to access each quartile individually, we will usually want them both, so we will access them through the one-variable statistics in the [**STAT**] menu.

You should still have the data in [**L1**] and the frequencies or weights in [**L2**], so press [**stat**], then arrow over to [**CALC**] (the left screen below) and choose 1-var Stat, which returns you to the Home Screen (see the middle screen below.). Enter [**2nd**] [**L1**] [**comma**] [**2nd**] [**L2**] 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).

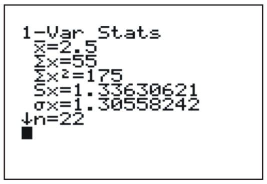




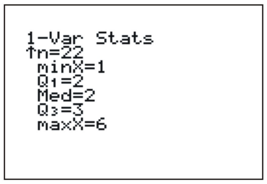
```
1-Var Stats
```



```
1-Var Stats L₁,L
₂■
```



```
1-Var Stats
 x̄=2.5
 Σx=55
 Σx²=175
 Sx=1.33630621
 σx=1.30558242
↓n=22
■
```



```
1-Var Stats
↑n=22
 minX=1
 Q₁=2
 Med=2
 Q₃=3
 maxX=6
```

Remember that Q_1 corresponds to the 25^{th} percentile and Q_3 is the 75^{th} percentile.

Lesson Summary

When examining a set of data, we use descriptive statistics to provide information about where the data is 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 notated as $\overline{x}$. When the entire population is involved, the population mean is μ. The **median** is the numeric middle of a data set. If there are an odd number of numbers, this middle value is easy to find. If there is an even number of data values, however, the median is the mean of the middle two values. The median is **resistant**, that is, it is not affected by the presence of outliers. An **outlier** is a number that has an extreme value when compared with most of the data. 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, is in 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.

Other measures of center include the **midrange**, which is the mean of the maximum and minimum values. In an $n\%$ **trimmed mean**, you remove a certain 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 weights.

A **percentile** is a data value in which the specified percentage of the data is below that value. The median is the 50^{th} percentile. Two well-known percentiles are the 25^{th} percentile, which is called the **lower quartile** (LQ or Q_1), and the 75^{th} percentile, which is called the **upper quartile** (UQ or Q_3)

Points to Consider

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

Review Questions

1. In Lois' $2^{nd}-$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 4^{th} 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 1.4:

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?

Review Answers

1. There is an outlier that is larger than most of the data. This outlier will "pull" the mean towards it while the median tends to stay in the center of the data, clustered somewhere between 45 and 52.
2. His mean for all four quarters would need to be at least 92.5 in order to receive the necessary grade. Multiplying 92.5 by 4, yields 370 as the necessary total. His existing grades total to 273. $370 - 273 = 97$.
3. 10% of 300 is 30, therefore, we would remove 15 numbers from each end.
4. 270
5. (a) 0
   (b) 210
   (c) 299.2
   (d) 222 (10% of 11 data points is really 1.1, so we decided to remove two points, or about 18%)
   (e) 646.5
   (f) $Q_1 : 0, Q_3 : 498$
   (g) 72.7%
6. There is one extreme point, 1293, which causes the mean to be greater than the median.

Further Reading

- 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/

Vocabulary

Mode The most frequently occurring number in a data set.

Mean The average, or the sum of the values in a data set divided by the number of values.

Median The numeric middle of a data set.

Outlier An extreme value in a data set.

Resistance A property of a statistic in which it is not affected by extreme values (outliers).

Midrange The mean of the minimum and maximum values in a data set.

$N\%$ **Trimmed Mean** A mean in which $n\%$ of the original data (equal amounts from either end) is removed before calculating the mean.

Weighted Mean A mean in which some values contribute more to the sum than others. Each value is multiplied by its weight, or frequency and then the sum of those totals is divided by the total of the weights or frequencies.

Percentiles A value in a data set in which the given percentage of the data is below that value.

Quartiles The values that divide a data set roughly into four roughly equal groups. The lower quartile is the 25^{th} percentile, and the upper quartile is the 75^{th} percentile.

Numerical (or Quantitative) Variable A variable in which the count is the attribute of interest.

Discrete Variable A numerical variable that only exhibits a finite set of values at given intervals.

Continuous Variable A numerical variable that can be any of an infinite range of values.

Sample A smaller, representative subset of the population.

Parameter A statistical measure or number that summarizes the entire population.

Statistic A measure or number that summarizes the individuals in a sample.

Sampling Error The inaccuracy that results from estimating using a sample, rather than the entire population.

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 concentrated on statistics that provided information about the way in which a data set is centered. Another important feature that can help us understand more about a data set is the manner in which the data is 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

For most students, their first introduction to a statistic that measures spread is the **range**. The range is simply the difference between the smallest value (minimum) and the largest value (maximum) in the data. Let's return to the data set used in the previous lesson:

$$75, 80, 90, 94, 96$$

Most students find it intuitive to say that the values **range** from 75 to 96. However, the range is a statistic, and as such is a single number. It is therefore more proper to say that the range is 21.

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

Interquartile Range

Similar to the range, the **interquartile range** is the difference between the quartiles. If the range tells us how widely spread the entire data set is, the interquartile range (abbreviated IQR) gives information about how the middle 50% of the data is spread.

Example:

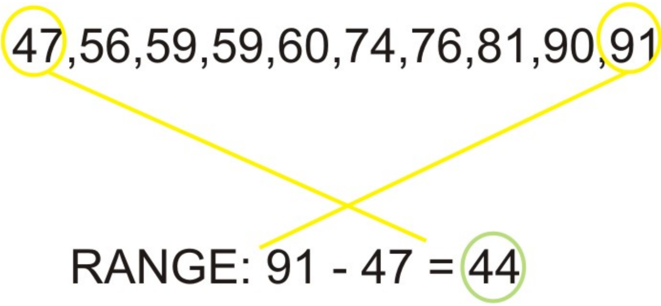
A recent study proclaimed Mobile, Alabama the "wettest" city in America (http://www.livescience.com/environment/070518_rainy_cities.html). The following table lists a measurement of the approximate annual rainfall in Mobile for the last 10 years. Find the Range and **IQR** for this data.

Table 1.5:

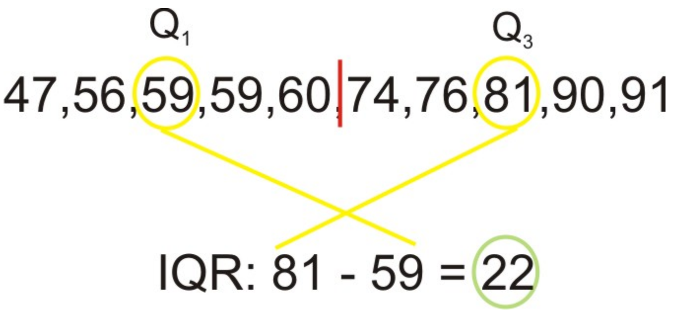
Year	Rainfall (inches)
1998	90
1999	56
2000	60
2001	59
2002	74
2003	76
2004	81
2005	91
2006	47
2007	59

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 subtract $Q3 - Q1$



Even though we are doing easy calculations, statistics is never about meaningless arithmetic and you should always be thinking about what a particular statistical measure means in the real context of the data. 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 is spread is to look at how far away the values are from the mean. The difference between the actual value and the mean is called the **deviation**. Written symbolically it would be:

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

Let's take a simple data set of three randomly selected individuals' shoe sizes:

$9\frac{1}{2}, 11\frac{1}{2}$, and 12

The mean of this data set is 11. The deviations then would be as follows:

Table 1.6: **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 the deviation of a point that is less than the mean is negative. Points that are above the mean have positive deviations.

We need a statistic that can summarize *all* of the deviations. The standard deviation is such a summary. It 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 sum to 0, in effect canceling each other out.

Table 1.7: **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 the deviations→	$-1.5 + 0.5 + 1 = 0$

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

Table 1.8:

Observed Data	Deviations	$(x - \bar{x}^2)$
9.5	-1.5	$-1.5^2 = 2.25$
11.5	0.5	$0.5^2 = 0.25$
12	1	$1^2 = 1$
	Sum of the deviations $= 0$	

Now find the sum of the squared deviations:

Table 1.9:

Observed Data	Deviations	$(x - \bar{x}^2)$
9.5	-1.5	2.25
11.5	0.5	0.25
12	1	1
	Sum of the squared deviations $= 3.5$	

Normally if you were finding a mean, you would now divide by the number of numbers (n). This is the part that puzzles many beginning statistics students. Instead of dividing by n, we divide by $n-1$, which will be explained later in this section. Dividing by 2 gives:

$$\frac{3.5}{2} = 1.75$$

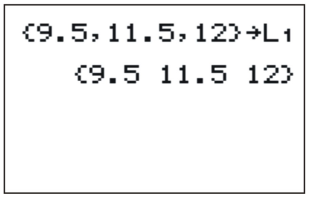
Remember that this number was obtained by squaring the deviations, so the result is much larger than it should be. This quantity is actually called the **variance** and it will be very important in later chapters. The final step is to "unsquare" the variance, or take the square root:

$$\sqrt{1.75} \approx 1.32$$

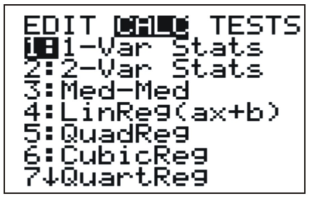
This is the standard deviation! This means that in our sample, the "typical" value is approximately 1.32 units away from the mean.

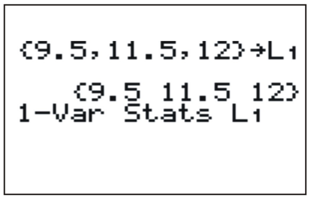
Technology Note: standard deviation on the TI-83 or 84

- Enter the above data in list [**L1**], as you did in the previous lesson (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 **Sx** is the standard deviation.

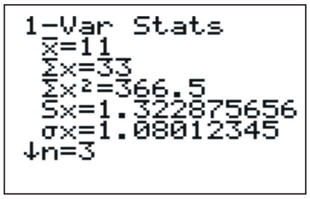


```
{9.5,11.5,12}→L₁
     {9.5 11.5 12}
```





```
{9.5,11.5,12}→L₁
     {9.5 11.5 12}
1-Var Stats L₁
```



```
1-Var Stats
 x̄=11
 Σx=33
 Σx²=366.5
 Sx=1.322875656
 σx=1.08012345
↓n=3
```

Why n-1?

There are several ways to look at the need to divide the sum by $n-1$ when calculating the standard deviation. For now, we will skip some of the more technical explanations that involve things like degrees of freedom that you will cover in later chapters in favor of adjusting for sampling error. 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 the number of numbers (N). But 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.

Sample Standard Deviation:

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

Because the variance is the square of the standard deviation, the variance formulas are as follows:

Variance of a population:

$$\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N}$$

Variance of a sample:

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

Chebyshev's Theorem

Pafnuty Chebyshev was a 19^{th} 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 is as follows:

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

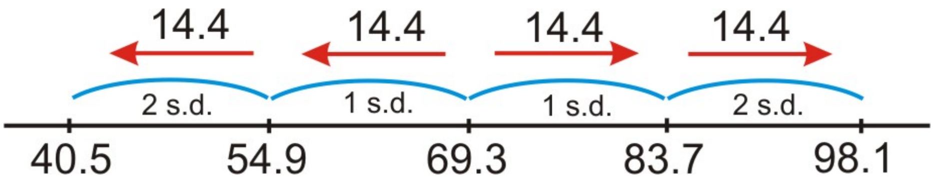
$1 - \frac{1}{k^2}$, where $k > 1$

As an example, let's return to the rainfall data from Mobile. The mean yearly rainfall amount is 69.3 and the sample standard deviation is about 14.4.

Let's investigate the information that Chebyshev's Theorem gives us about the proportion of data within 2 standard deviations of the mean. If we replace k with 2, the result is:

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

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



According to the drawing, Chebyshev's Theorem states that at least 75% of the data is between 40.5 and 98.1. Well, this probably doesn't seem too significant in this example, because all of the data falls within that range. In a later chapter we will learn a more informative rule about standard deviation, but 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 also use descriptive statistics to provide information about how the data is 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 quantity is called the **variance**. Taking the square root of the variance gives the standard deviation. For a population, the standard deviation is notated σ. Because a sample is prone to random variation (sampling error), we adjust the sample standard deviation to make it a little larger by divided 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 a information about the minimum percentage of data that is within a certain number of standard deviations of the mean it applies to any population or sample, regardless of how that data is distributed.

Points to Consider

1. How do you determine which measure of spread best describes a particular data set?
2. What information does the standard deviation tell us about the specific, real data being observed?
3. What are the effects of outliers on the various measures of spread?
4. 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 standard deviation and the variance.

Table 1.10:

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		
		Sum →	

Variance:

Standard Deviation:

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

Table 1.11:

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 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

Review Answers

1. (a) 69.3 inches
   (b)

Table 1.12:

Year	Rainfall (inches)	Deviation	Squared Deviations
1998	90	20.7	428.49
1999	56	−13.3	176.89

Table 1.12: (continued)

Year	Rainfall (inches)	Deviation	Squared Deviations
2000	60	-9.3	86.49
2001	59	-10.3	106.09
2002	74	4.7	22.09
2003	76	6.7	44.89
2004	81	11.7	136.89
2005	91	21.7	470.89
2006	47	-22.3	497.29
2007	59	-10.3	106.09
		Sum →	2076.1

Variance: 230.68
Standard Deviation: 15.19

2. RANGE: 1293 IQR: 498
3. 387.03
4. a
5. b

Further Reading

- 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

Vocabulary

Range The maximum value in a data set minus the minimum value.

Interquartile Range (IQR) The upper quartile in a data set minus the lower quartile.

Deviation The difference of the mean of a data set subtracted from the actual data value.

Standard Deviation A measure of the "typical" distance of all the data points in a set from the mean.

Population Standard Deviation The square root of the result of dividing the sum of the squared deviations by the population size.

Sample Standard Deviation The square root of the result of dividing the sum of the squared deviations by one less than the sample size.

Variance The square of the standard deviation.

1.5 Chapter Review

Part One: Multiple Choice

1. Which of the following is true for any set of data?

   (a) The range is a resistant measure of spread.
   (b) The standard deviation is not resistant.
   (c) The range can be greater than the standard deviation.
   (d) The IQR is always greater than the range.
   (e) The range can be negative.

2. The following shows the mean number of days of precipitation by month in Juneau Alaska:

Table 1.13: **Mean Number of Days With Precipitation > 0.1 inches**

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

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

Which month contains the median number of days of rain?

(a) January

(b) February

(c) June

(d) July

(e) September

3. Given this set of 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. Which of the following represents a true statistical experiment?

(a) Researchers collect temperatures from the Arctic Ocean to determine the rate of climate change.
(b) Researchers collect, tag, and release geese from Siberia to determine their migration patterns.
(c) Researchers select 50 individuals who smoke 1 pack of cigarettes a day and 50 individuals who do not smoke to test their lung capacities.
(d) Researchers select 50 individuals at random. 25 are given a new drug to boost memory function, 25 are given a water pill and told that it is medication that will help their memories (a placebo). The memory function of both groups is then tested and compared.
(e) Researchers select 50 individuals at random and ask them questions about their diet. Each individual's physical fitness is then tested and compared to determine if there is a relationship between diet and health.

8. 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 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.
9. 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 their own experiment that proves their medication is the best.
   (e) When individuals in a sample answer a survey untruthfully.
10. 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 One: Answers

1. b
2. a
3. b
4. e *(Note: while the standard deviation MUST always be smaller than the range, the variance is not always smaller than the range. It is also true that the variance is the square of the standard deviation, but some standard deviations will get smaller when they are squared. Challenge students to find examples of data sets that illustrate these points.)*
5. c
6. c
7. d

8. d
9. b
10. b

Part Two: Open-Ended Questions

11. 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. Determine the following:

(i) mode

(ii) mean

(iii) median

(iv) upper and lower quartiles

(v) midrange

(vi) range

(b) If Erica's 62 quiz was removed from the data, briefly describe (without recalculating) the anticipated effect on the statistics you calculated in part a.

12. Mr. Crunchy's sells small bags of potato chips that are advertised to contain 12 ounces of potato chips. To minimize complaints from their customers, the factory sets the machines to fill bags with an average weight of 13 ounces. For an experiment in his statistics class, Spud goes to 5 different stores, purchases 1 bag from each store and then weighs the contents. The weights of the bags are: $13.18, 12.65, 12.87, 13.32$, and 12.93 grams.

(a) Calculate the sample mean

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

Table 1.14:

Observed Data	Deviations	$(x - \bar{x}^2)$
13.18		

Table 1.14: (continued)

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

(c) Calculate the variance

(d) Calculate the standard deviation

(e) Explain what the standard deviation means in the context of the problem.

13. 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 1.15:

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

Source: http://en.wikipedia.org/wiki/Gal%C3%A1pagos_Islands

(a) Calculate the mode, mean, median, quartiles, range, and standard deviation for this data.

Mode:

Mean:

Median:

Upper Quartile:

Lower Quartile:

Range:

Standard Deviation:

(b) Explain why the mean is so much larger than the median in the context of this data.

(c) Explain why the standard deviation is so large.

14. At http://content.usatoday.com/sports/baseball/salaries/default.aspx, USAToday keeps a data base of major league baseball salaries. You will see a pull-down menu that says, "Choose an MLB Team". Pick a team and find the salary statistics for that team. Next to the current year 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.

(b) Find the other measures of center and record them.

Mean:

Mode:

Midrange:

Lower Quartile:

Upper Quartile:

QR:

(c) Explain the real-world meaning of each measure of center in the context of this data.

Mean:

Median:

Mode:

Midrange:

Lower Quartile:

Upper Quartile:

IQR:

(d) Find the following measures of spread:

Range:

Standard Deviation:

(e) Explain the real-world meaning of each measure of spread in the context of this situation.

(f) Write two sentences commenting on two interesting features about the way the salary data is distributed for this team.

Part Two: Answers

11. (a) i. mode 99 and 87
    ii. mean 89.23
    iii. median 89
    iv. upper and lower quartiles $Q1 = 87, Q3 = 97.5$
    v. midrange 80.5
    vi. range37
    (b) The 62 is an outlier in the data set. This would usually cause the mean to be significantly lower than the median, but the three 99's are balancing this out. When we remove the 62, the mode should not be affected. The mean should increase. The most dramatic changes will occur in the midrange and range. The range should be much smaller and the midrange should increase. We would expect little change to the medians and quartiles as they are resistant measures.

12. (a) $\bar{x} = 12.99$
    (b) $s = 0.264$
    (c) $s^2 = 0.07$
    (d) The standard deviation tells you that the "typical" or "average" bag of chips in this sample is within 0.07 grams of the mean weight. Based on our sample, we would not have reason to believe that the company is selling unusually light or heavy bags of chips. Their quality control department appears to be doing a good job! (Note: this answer is very subjective for now, but it is important to start thinking in this manner. In later chapters, we will examine more precise measures and conclusions for this process.)
13. (a) Mode: 60 km^2
    Mean: 439.3 km^2
    Median: 42 km^2
    Upper Quartile: 558 km^2

Lower Quartile: 4.9 km^2
Range: 2639.67 km^2
Standard Deviation: 1088.69 km^2

(b) There is one **very** extreme outlier. Isabela is by far the largest island. In addition to that, there are many points in the lower half of the data that are very closely grouped together. Many of these islands are volcanic rock that barely poke above the surface of the ocean. The upper 50% of the data is much more spread out. This creates a situation in which the median stays very small, but the mean will be strongly pulled towards the larger numbers because it is not resistant.

(c) The standard deviation is a statistic that is based on the mean. Therefore, if the mean is not resistant, the standard deviation is not, and it will also be influenced by the larger numbers. If it is a measure of the "typical" distance from the mean, then the larger points will have a disproportionate influence on the calculation. On a more intuitive level, if the upper 50% of the data is very widely spread, the standard deviation reflects that extreme variation.

14. (a) Will vary
(b) Will vary
(c) Mean: the average salary of the players on this team in 2007.
Median: the salary at which half the players on the team make more than that, and half the players make less than that.
Mode: the salary that more players make than any other individual salary. Usually, this is a league minimum salary that many players make.
Midrange: The mean of just the highest paid and lowest paid players.
Lower Quartile: The salary at which only 25% of the players on the team make less.
Upper Quartile: The salary at which 75% of the players make less, or the salary at which only one quarter of the team makes more.
IQR: The middle 50% of the players varies by this amount.
(d) e.
Range: The gap in salary between the highest- and lowest-paid players.
Standard Deviation: the amount by which a typical player's salary varies from the mean salary.
(f) Answers will vary, but students should comment on spread in one sentence and center in the other. Since many baseball teams have a few star players who make much higher salaries, most examples should give the students an opportunity to comment on the presence of outliers and their affect on the statistical measures of center and spread.

Image Sources

(1) .

(2) .

(3) .

(4) . CC-BY-SA.

(5) *Galapagos Map.*. GNU Free Documentation License.

(6) *Darwin's Finches*. Public Domain.

(7) . CC-BY-SA.

Chapter 2

Visualizations of Data

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

In chapter 1, we focused on describing data using summary statistics. While this is very useful in analyzing and learning important characteristics of a data set, it is also very important and informative to represent data in some visual format. This is in fact the form in which most people are used to encountering data while engaged in such things as reading newspapers, magazines, food labels, or watching television. 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

A Real Context: Recycling Issues

The earth has seemed so large in scope for thousands of years that it is only recently that many have begun to take seriously the idea that we live on a planet of limited and dwindling resources that is in a sense, and island in the middle of space. This is something that residents of the Galapagos Islands are also beginning to understand on a much more dramatic level. 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, as well as other island cultures. 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 not 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 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 data in this chapter is focused around the issues and consequences of our recycling habits, or lack thereof!

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 and the small fraction of it 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.

One of the reasons for the large increase of bottled beverages has been an increased focus on health and fitness and it has spilled over into all aspects of life. Ask your teacher if they ever had water bottles in their classes when they were students?

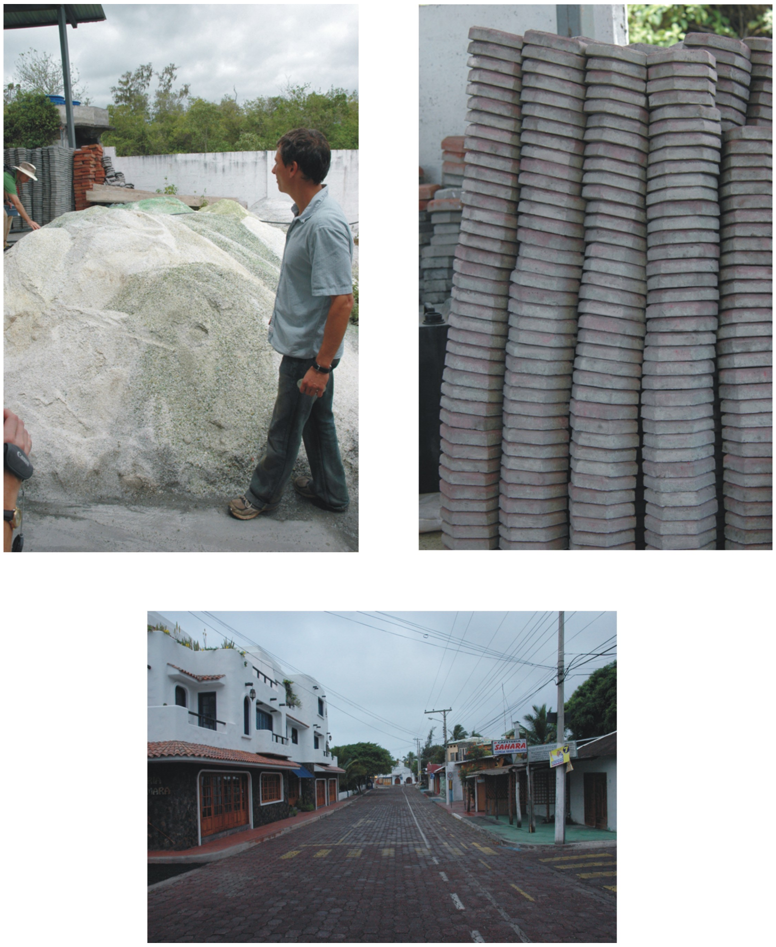


Figure 2.1: The Recycling Center on Santa Cruz in the Galapagos turns all the recycled glass into pavers that are used for the streets in Puerto Ayora. (1)

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 in some way that makes it 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 consumed each week. You could use your class data, but let's use an imaginary class. Here is the raw data:

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

Because the data is only limited to the numbers 1 through 8, it is very simple to create a frequency table using those values. For example, here is a table you could use to collect data from your classmates:

Table 2.1:

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

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

Figure: Imaginary Class Data on Water Bottle Usage

Table 2.2: **Completed Frequency Table for Water Bottle Data**

Number of Plastic Beverage Bottles per Week	Frequency
1	1
2	1
3	3
4	4

Table 2.2: (continued)

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

While this data set is rather simple and small, you can see how much easier it is to interpret the data in this form. One caution about translating raw data into a more helpful visual form is that it is very easy to make a mistake, especially with a larger data set. In this case, it is often helpful to use tally marks as a running total to help construct the table and avoid missing a value or over-representing another.

Table 2.3: **Frequency table using tally marks**

Number of Plastic Beverage Bottles per Week	Tally	Frequency
1	𝍷	1
2	𝍷	1
3	𝍷𝍷𝍷	3
4	𝍷𝍷𝍷𝍷	4
5	𝍸 𝍷	6
6	𝍸 𝍷𝍷𝍷	8
7	𝍸 𝍷𝍷	7
8	𝍷𝍷	2

This data set could almost be considered categorical and was easy to translate into a frequency table. In many situations, you will need to create your own categories, or classifications. The following data set shows the countries in the world that consume the most bottled water per person per year.

Table 2.4:

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

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

This data has been measured at the ratio level (see levels of measurement in chapter one), so there is some flexibility required in order to create meaningful and useful categories for a frequency table. The values range from 80.3 liters, up to 183 liters. By examining the data, it might seem appropriate for us to create frequency table by 10s (80s, 90s, etc.) We will skip the tally marks in this case because the data is already in numerical order and it is easy to see how many are in each classification.

Table 2.5:

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

Table 2.5: (continued)

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

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

Notice the mathematical notation used for each classification. A bracket [or] indicates that the endpoint of the interval is included in the class. A parentheses (or) indicates that the endpoint is not included. What do you do with a number that is in between two classifications? For example, it is unlikely, but possible that a country consumed *exactly* 90 liters of bottled water per person. It is intuitive to include this in the90 s, not the 80 s, but how would we label the categories? If you wrote $80 - 90$ and $90 - 100$, it would seem as if 90 belongs in both classes. But if you wrote$80 - 89$, what would you do with 89.5? It is common practice in statistics to include a number that borders two classes in the **larger** of the two. So, $[80 - 90)$ means this classification includes everything from 80 that gets infinitely close to, but not equal to 90. Even if the bracket notation is not used, you should always place such values in the higher classification.

Histograms, Not Bar Graphs!

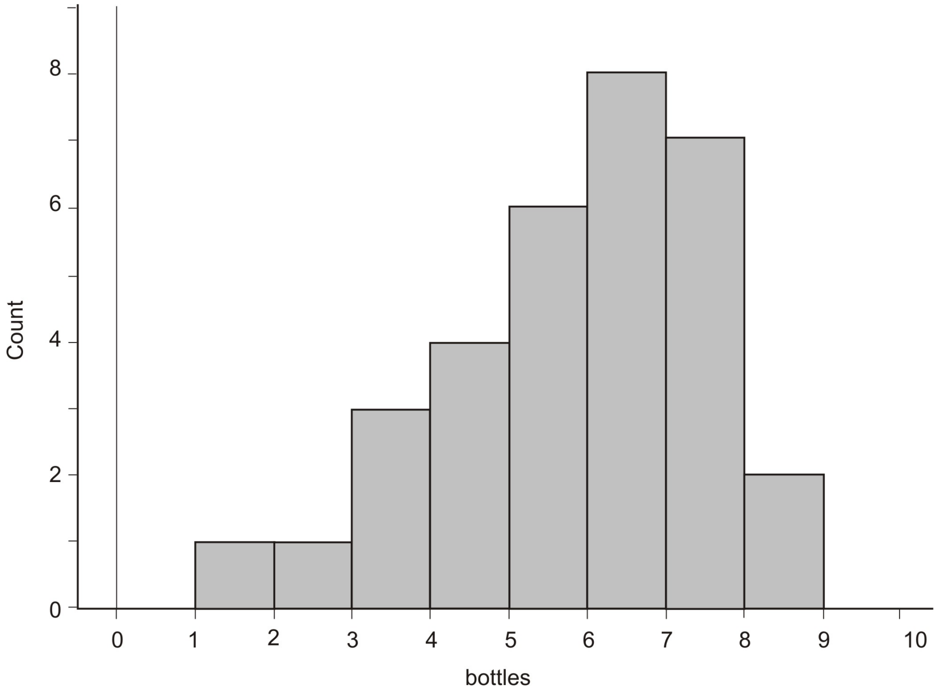
Once you can create a frequency table, you are ready to create our first graphical representation, called a histogram. Let's revisit our data about student bottled beverage habits.

Table 2.6:

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

Figure: Completed Frequency Table for Water Bottle Data

Here is the same data in a histogram:



In this case the horizontal axis represents the variable (number of plastic bottles) and the vertical axis is the frequency or count. Each vertical bar represents the number of people in each class of ranges of bottles (e.g. $[1-2), [2-3)$, etc.) 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 Not a Bar Graph!

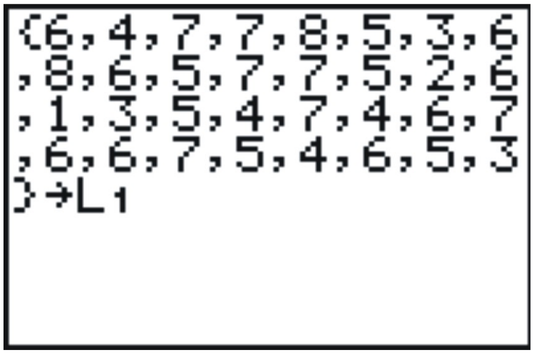
Please avoid a mistake of beginning statistics students and **do not** call this a bar graph! As you will learn later, bar graphs are only for categorical data. A histogram is for numerical data and most often will describe continuous data. With histograms, the different sections are referred to as "bins" rather than "bars." Think of the column, or "bin,", as a vertical container that collects all the data for that range of values.

Just like the frequency table, 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. When students are drawing histograms, they sometimes make the error of looking at the last value in the data and stop their horizontal axis at this 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 midpoints of each bin unless you specify otherwise.

Histograms on the Graphing Calculator

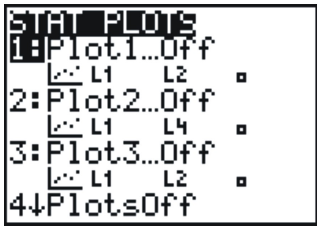
To draw a histogram on your TI-83-family graphing calculator, you must first enter the data in a list. In chapter 1 you used the List Editor. Here is another way to enter data into a list:

In the home screen press **2ND** and then enter the data separated by commas (see the screen below). When all the data has been entered, press **2ND** [**STO**] then **2ND** [**L1**].

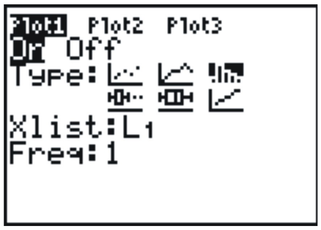


```
{6,4,7,7,8,5,3,6
,8,6,5,7,7,5,2,6
,1,3,5,4,7,4,6,7
,6,6,7,5,4,6,5,3
}→L1
```

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 Plot 1. 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 name 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.

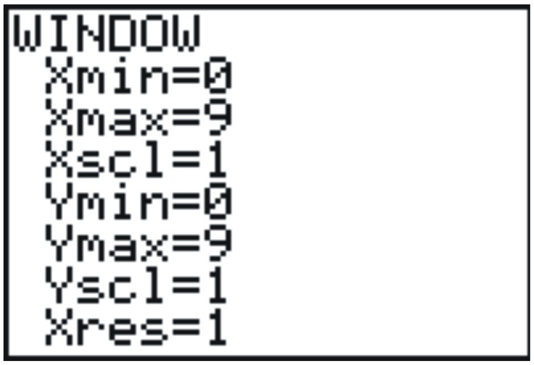


```
STAT PLOTS
1:Plot1…Off
    L1   L2
2:Plot2…Off
    L1   L4
3:Plot3…Off
    L1   L2
4↓PlotsOff
```



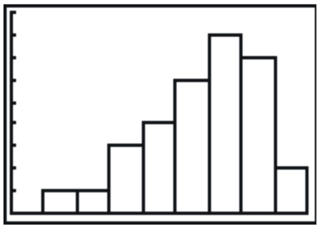
```
Plot1 Plot2 Plot3
On Off
Type:

Xlist:L1
Freq:1
```

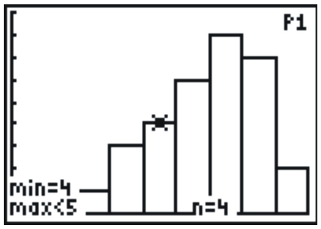
Finally, we need to set a window. Press [**WINDOW**] and enter an appropriate window to display the plot. In this case XSCL is what determines the bin width. Also notice that the maximum x value needs to go up to 9 to show the last bin, even though the data stops at 8.



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

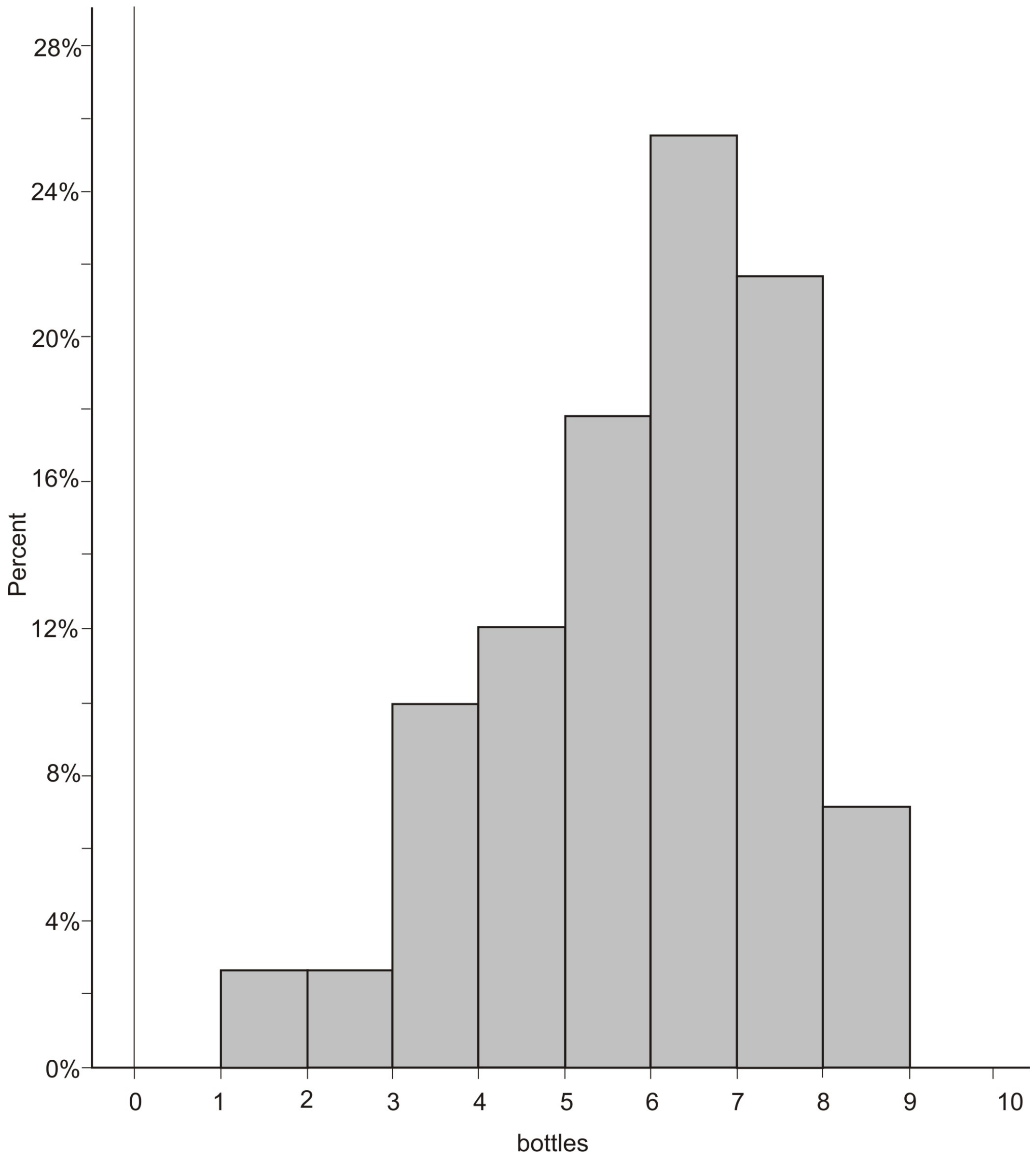
Press [**GRAPH**] to display the histogram. If you press [**TRACE**] and then use the left or right arrows to trace along the graph, notice how the calculator uses the notation to properly represent the values in each bin.





It's All Relative!!

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. This way the numbers reflect the amount **relative** to the entire data set.

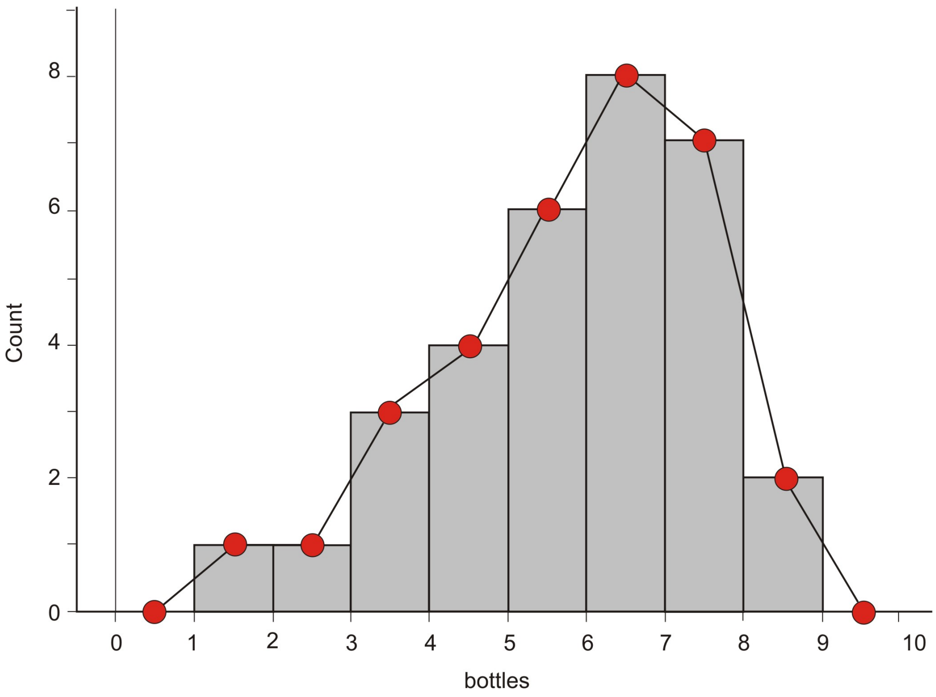


Frequency Polygons

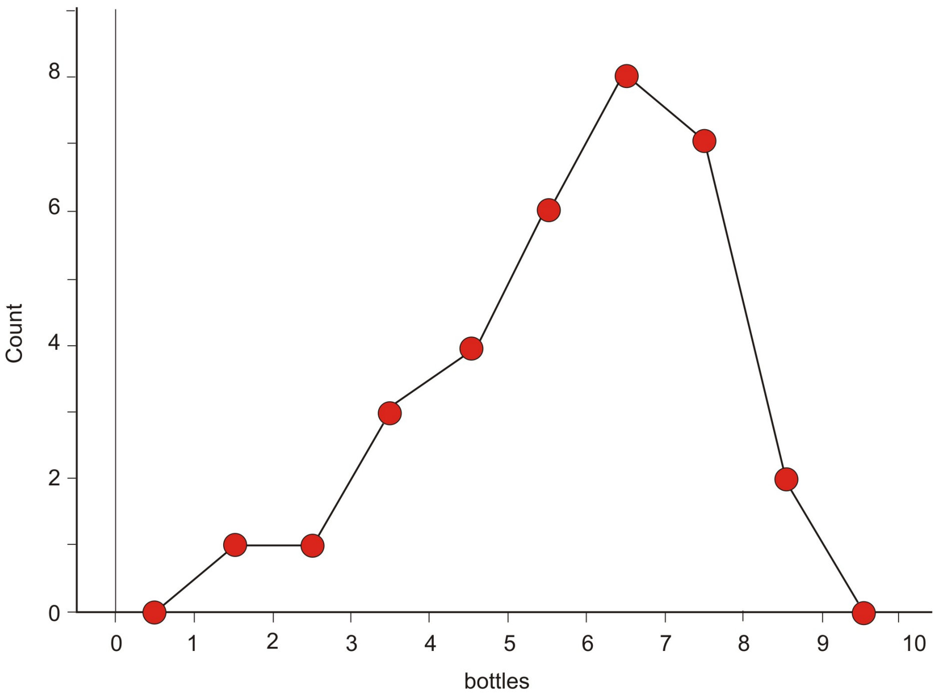
A **frequency polygon** is similar to a histogram, but instead of using bins, a polygon is created by plotting the frequencies and connecting those points with a series of line segments.

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 the frequency polygon constructed directly from the histogram.

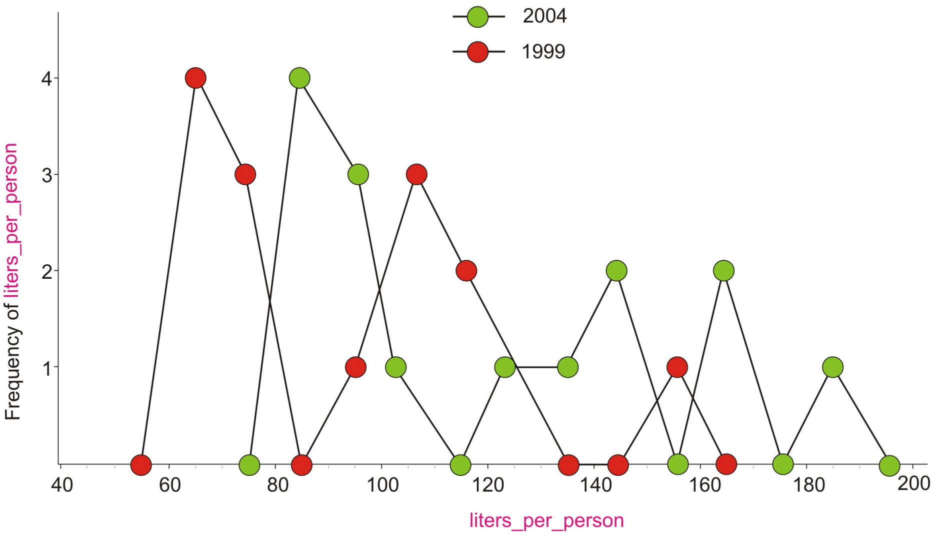


And here is the frequency polygon in finished form.



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!

For example, we looked at the bottled water consumption of the leading countries in the year 2004 and you will work with the data from 1999 at the end of the lesson, but it would be nice to be able to compare the two distributions of data. A frequency polygon would help give an overall picture of how these results are similar and different. In the following graph, the two frequency polygons are overlaid, 1999 in red, and 2004 in green. Can you see any important differences in the way the graph is shaped?



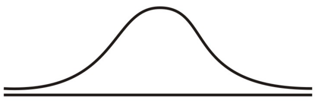
First of all, it appears as if there was a shift to the right in all the data, which is explained by realizing that all of the countries have significantly increased their consumption. The first peak in the lower consuming countries is almost identical but has increased by 20 liters per person. In 1999 there was a middle peak, but that group showed an even more dramatic increase in 2004 and has shifted significantly to the right (by between 40 and 60 liters per person). The frequency polygons is the first type of graph we have learned that make this type of comparison easier and we will learn others in later lessons.

The Mantra of Descriptive Statistics: Shape, Center, Spread

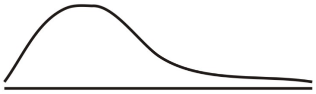
In the first chapter we introduced measures of center and spread as important indicators of a data set. We now have the tools to include the shape of a distribution of data as being very important as well. The “big three”: **Shape, Center, and Spread** should always be your starting point when describing a data set. If a statistician had to wear a uniform, it should probably say: shape, center, and spread.

If you look back at our imaginary student poll on using plastic beverage containers, A first glance would allow us to conclude that the data is **spread** out from 0 up to 9. The graph illustrates this concept, and we have a statistic that we used in the first chapter to quantify it: the range. Notice also that there is a larger 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 that area. We also used statistical measures to quantify this concept such as the mean and the median, but it is important that you "see" the idea of the center of the distribution as being near the large concentration of data.

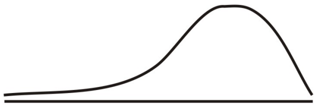
Shape is harder to describe with a single statistical measure, so we will describe it in less quantitative terms. A very important feature of this data set, as well as many that you will encounter is that it has a single large concentration of data that appears like a mountain. Data that is shaped in 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.** Though the true definition of a density curve will come in a later chapter, we should start to get used to the correct terminology now. The most important feature of the first density curve is symmetry. A concise description of the shape of this distribution therefore, would be **symmetric and mound shaped.** Notice in 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.** Be careful!! Many beginning statistics students think it would intuitively make sense to refer to the side with the concentration of the data as the direction of the skewing. Instead, 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 3^{rd} curve, the left tail of the distribution is stretched out, so this distribution is **skewed left**. Our student bottle data has this skewed left shape.

Cumulative Frequency Histograms and Ogive Plots

Very often it is helpful to know how much of the data accumulates 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** that class.

Table 2.7:

Number of Plastic Beverage Bottles per Week	Frequency	Cumulative Frequency
1	1	1
2	1	2

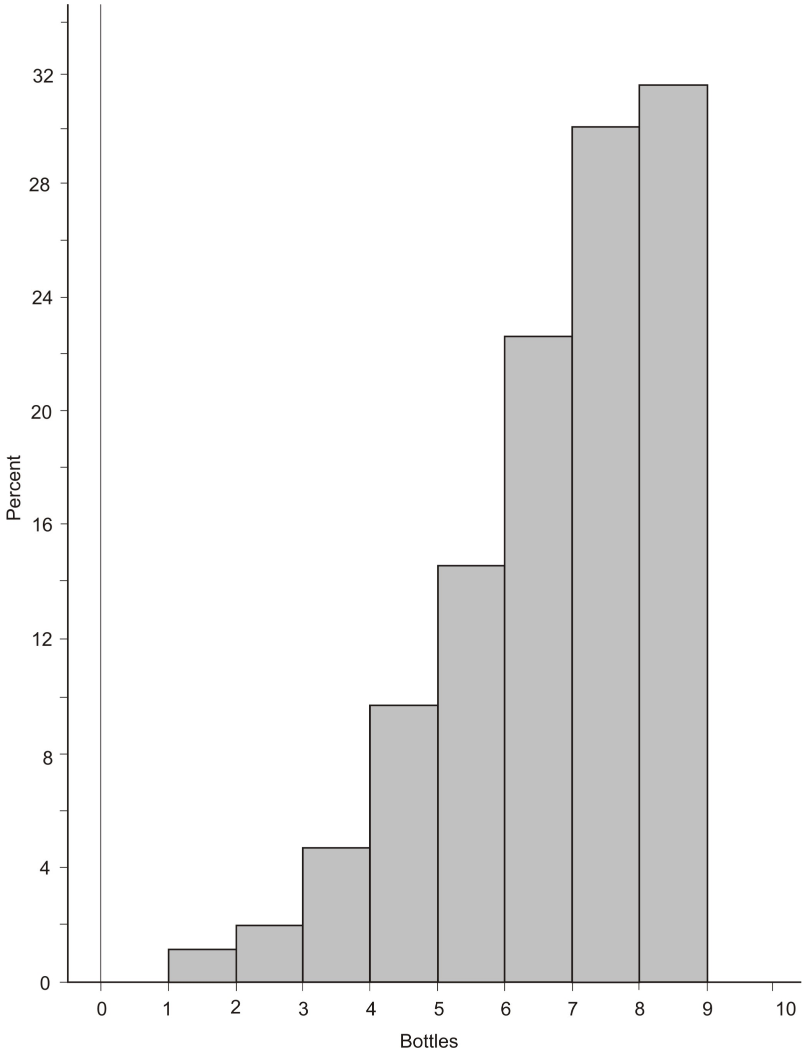
Table 2.7: (continued)

Number of Plastic Beverage Bottles per Week	Frequency	Cumulative Frequency
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 plot, only using the relative frequencies:

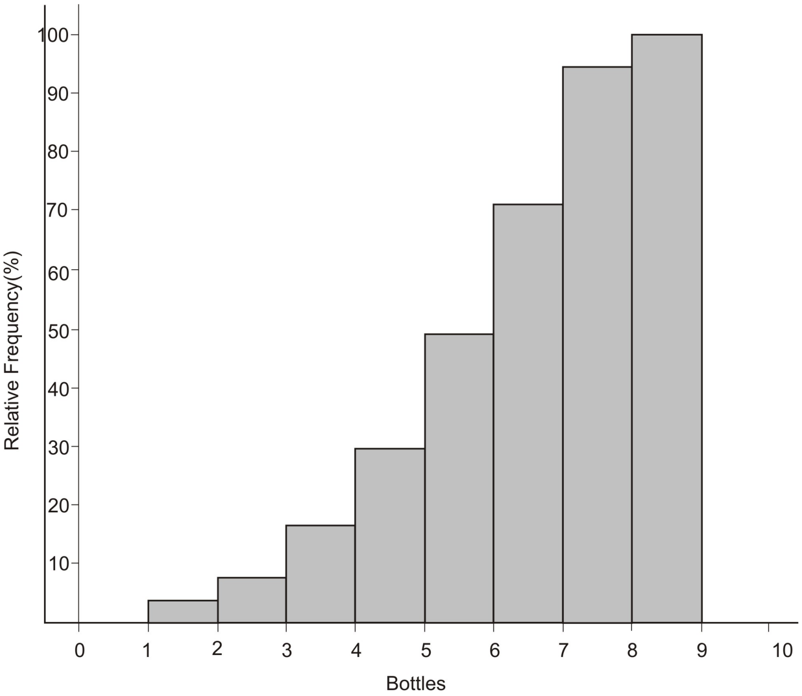
Table 2.8:

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

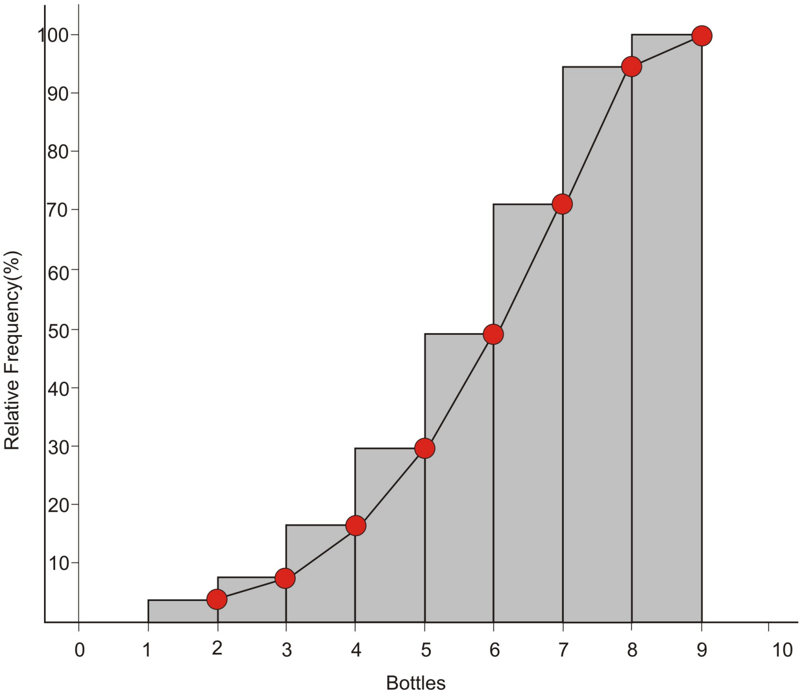
Table 2.8: (continued)

Number of Plastic Beverage Bottles per Week	Frequency	Cumulative Frequency	Relative Cumulative Frequency(%)
7	7	30	93.8
8	2	32	100

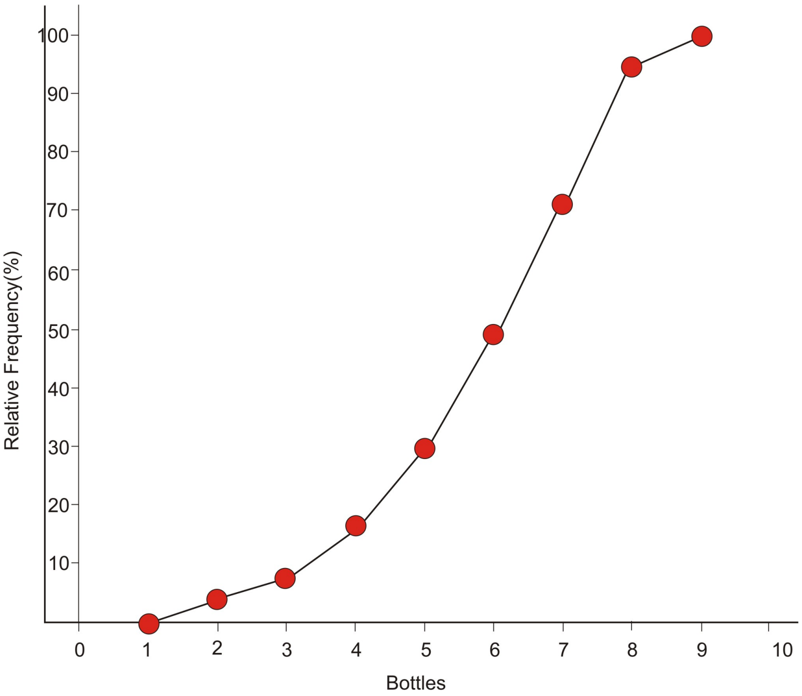
Figure: Cumulative Frequency Table for Bottle Data



Remembering what we did with a frequency polygon, we can remove the bins to create a new type of plot. In the frequency polygon, we used the midpoint of the bin width. It is slightly different for a relative cumulative frequency plot. This time we will plot the points 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 that value, just like 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 greater than or equal to 3 and less than 4. 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 and 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.



”Relative cumulative frequency plot” is quite a mouthful! This plot is commonly referred to as an Ogive Plot. The name ogive comes from a particular shaped 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.

TU ERES EL CRISTO EL HIJO DE DIOS VIVO

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

Lesson Summary

A **frequency table** is useful to organize data into classes according to the number of occurrences in each class, or frequency. **Relative frequency** shows the percentage of data in each class. A graphical representation of a frequency table (either actual or relative frequencies) that uses bins to show the amount in each class is called a **histogram.** Though it looks very similar, a bar graph is only used for categorical variables. A **frequency polygon** is created by plotting the midpoints of each bin at their frequencies 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.** Data 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 a tail spread out to the right. An **ogive plot,** or relative cumulative frequency plot shows how the data accumulates across the different values of the variable.

Points to Consider

1. What characteristics of a data set make it easier or harder to represent it using frequency tables, histograms, or frequency polygons?
2. What characteristics of a data set make representing it using frequency tables, histograms, frequency polygons, or ogives more or less useful?
3. What effects does the shape of a data set have on the statistical measures of center and spread?
4. How do you determine the most appropriate classification to use for a frequency table or 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 2.9:

Number of Plastic Beverage Bottles per Week	Tally	Frequency
1	𝍷𝍷	
2		
3	𝍷𝍷𝍷	
4	𝍷𝍷	
5	𝍷𝍷𝍷	
6	𝍸 𝍷𝍷	
7	𝍸 𝍷	
8	𝍷	

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

Table 2.10:

Class	Frequency
$[0 - 5)$	4
$[5 - 10)$	0

Table 2.10: (continued)

Class	Frequency
[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 2.11:

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 2.12:

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	1
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.1 , pg 82-83)

(a) Complete the frequency table below including the actual frequency, the relative frequency(round to the nearest tenth of a percent), and the relative cumulative frequency.

(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?

Review Answers

1. There are 24 tally marks, which means that the remaining 6 students must have been "2"s.

Table 2.13:

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

2. (a) 10 is the only possible value present. Because any multiple of ten is included in thee larger class, the only possible intervals are $10-15, 20-25$, and $30-35$. Because two of those classes contain no data, it must be one of the two values in the interval from $10-15$.

3. (a)

Table 2.14:

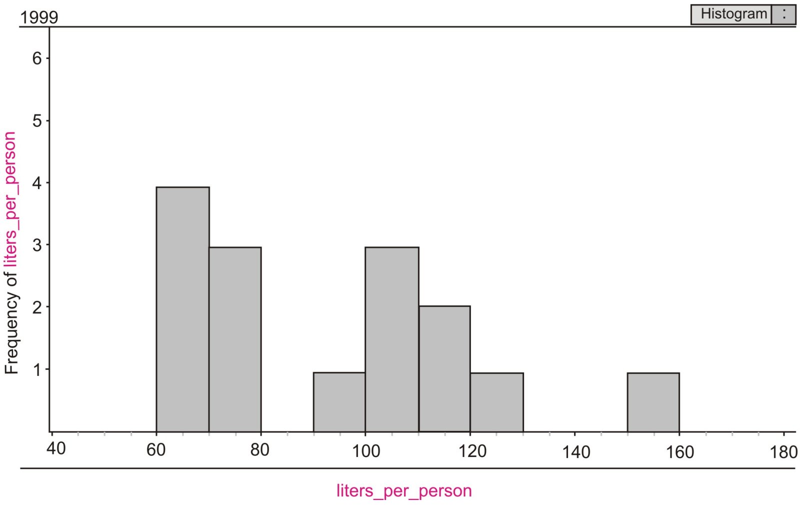
Liters of Water per Person	Frequency
$[60-70)$	4
$[70-80)$	3

Table 2.14: (continued)

Liters of Water per Person	Frequency
[80 – 90)	0
[90 – 100)	1
[100 – 110)	3
[110 – 120)	2
[120 – 130)	1
[130 – 140)	0
[140 – 150)	0
[150 – 160)	1

Completed Frequency Table for World Bottled Water Consumption Data(1999)

(b)



Student answers may vary if they choose a different bin width for their histogram.

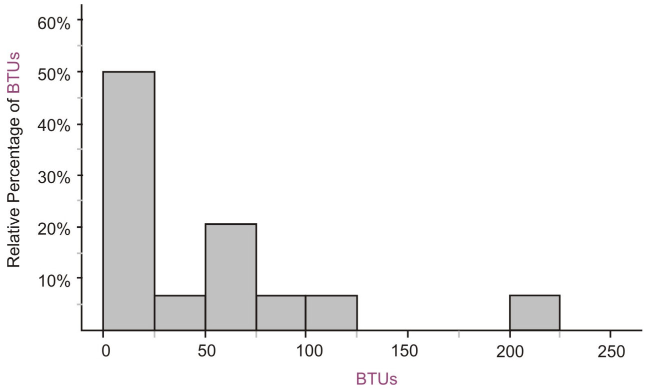
(c) This data set does appear to be have some characteristics of being skewed right. There also appears to be two distinct mounds. This shape is called "bimodal".

4. (a)

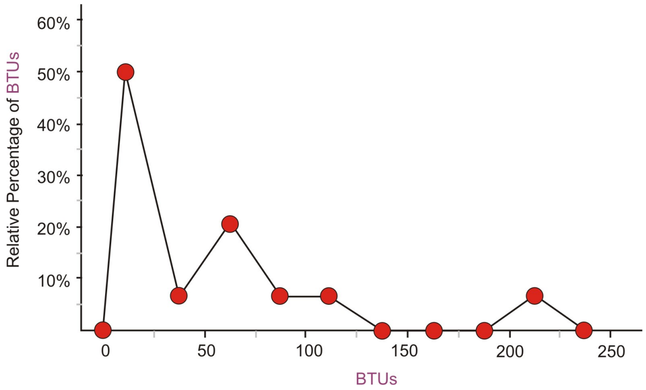
Table 2.15:

Class	Frequency	Relative Frequency(%)	Cumulative Frequency	Relative Cumulative Frequency(%)
[0 – 25)	7	50	7	50
[25 – 50)	1	7.1	8	57.1
[50 – 75)	3	21.4	11	78.6
[75 – 100)	1	7.1	12	85.7
[100 – 125)	1	7.1	13	92.9
[125 – 150)	0	0	13	92.9
[150 – 175)	0	0	13	92.9
[175 – 200)	0	0	13	92.9
[200 – 225)	1	7.1	14	100

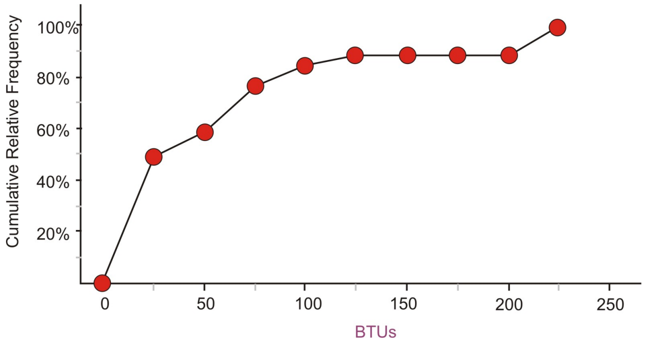
(b)



(c)



(d)



(e) This distribution is skewed to the right, which means that most of the materials are concentrated in the area of saving up to 75 million BTU's by using recycled materials and there are just a few materials (copper wire, carpet, and aluminum cans) that use inordinately large amounts of energy to create from raw materials.

(f) 99.8% >. The total should be **all** of the data, or 100%. The reason for the difference is rounding error.

(g) The horizontal portion of the ogive is where there is no data present, so the amount of accumulated data does not change.

(h) Because the ogive shows the increase in the percentage of data, the steepest section (in this case between 0 and 50%) is where **most** of the data is located and the accumulation of data is therefore changing at the most rapid pace.

References

- http://www.earth-policy.org/Updates/2006/Update51_data.htm
- http://en.wikipedia.org/wiki/Ogive

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 additional ways to represent single variable numerical variables, we will also cover a couple of methods for display categorical variables and an introduction to using a scatterplot and line graph to show the relationship between two variables. Continued emphasis will be placed on what can be learned about the data by describing the shape, center and spread of the distributions. We will also begin to compare the different graphical representations in terms of what additional information each can or cannot give about a data set.

Categorical Variables: Bar Graphs and Pie Graphs

E-Waste and Bar Graphs

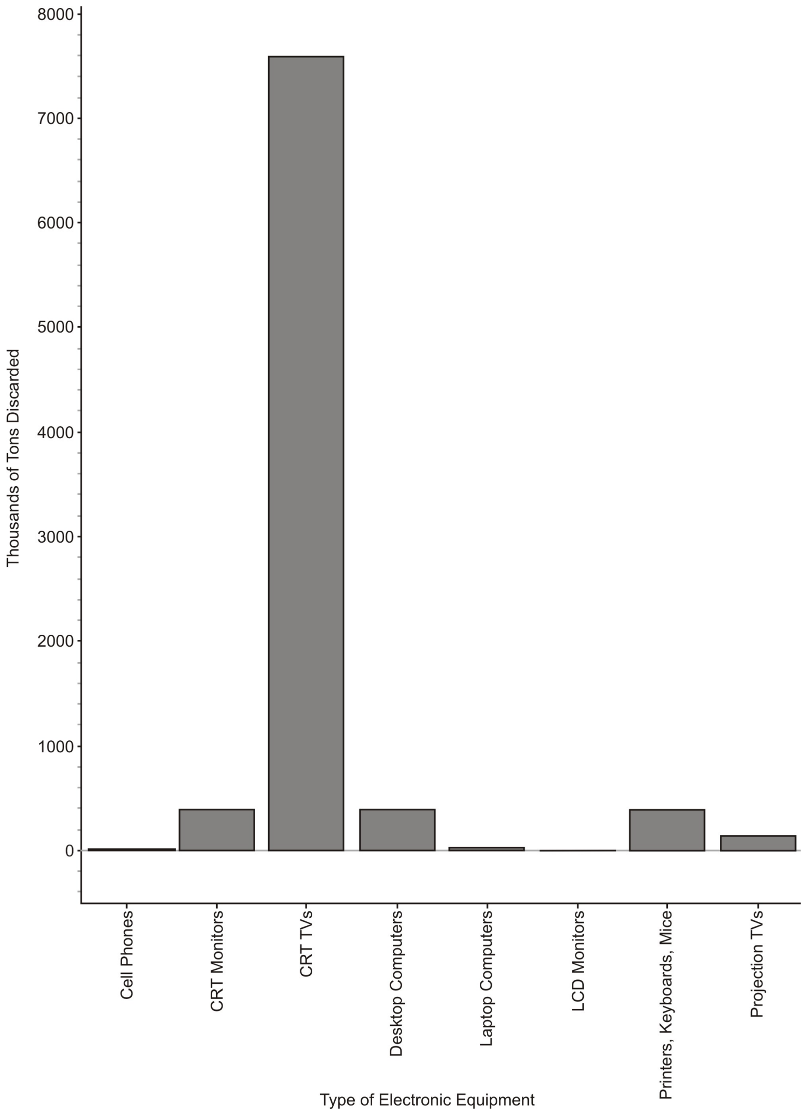
We live in an age of unprecedented access to increasingly sophisticated and affordable personal technology. Cell phones, computers, and televisions now improve so rapidly that, while they may still be in working condition, the drive to make use of the latest technological breakthroughs leads many to 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 2.16:

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. Histograms are used to show a range of continuous, numerical data. Even if we "pushed" the bars together, the space between them has no meaning, the graph is just a series of disjoint categories.

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

Pie Graphs

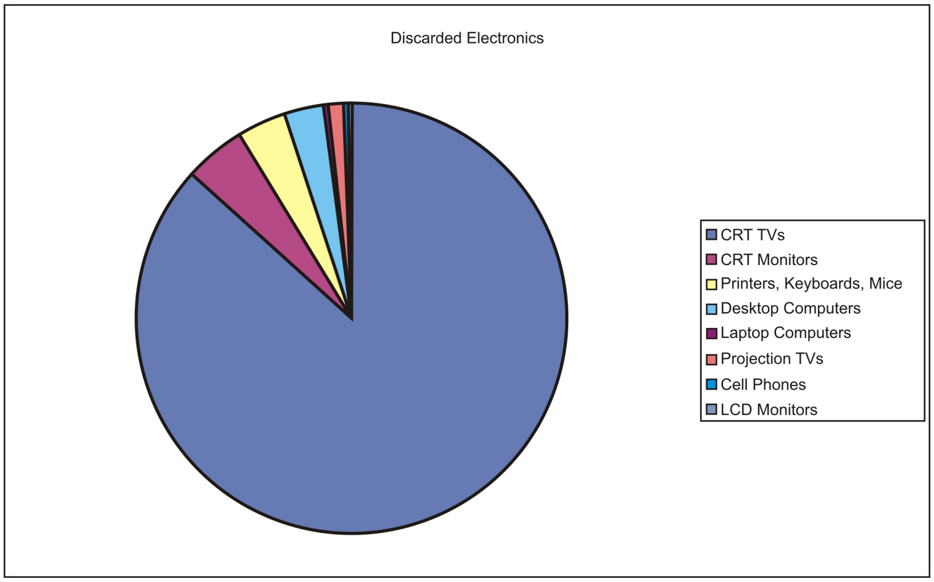
Usually, data that can be represented in a bar graph can also be shown using a pie graph (also commonly called a circle graph or pie chart). In this representation, we convert the count into a percentage so we can show each category relative to the total. Each percentage is then converted into a proportionate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

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

Table 2.17:

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	16.0
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:



Numerical Variables: 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.

For example, here is data from the percentage of paper packaging manufactured from recycled materials for a select group of countries.

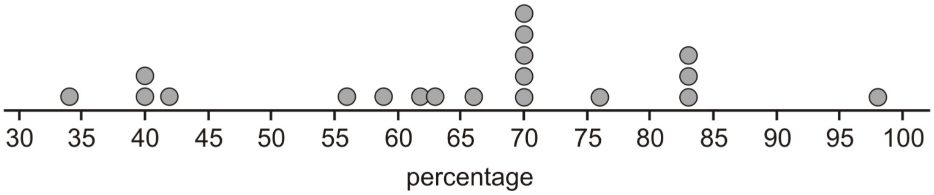
Table 2.18: **Percentage of the paper packaging used in a country that is recycled. National Geographic, January 2008. Volume 213 No.1 , pg 86-87.**

Country	% of Paper Packaging Recycled
Estonia	34
New Zealand	40
Poland	40
Cyprus	42
Portugal	56
United States	59
Italy	62
Spain	63
Australia	66
Greece	70
Finland	70
Ireland	70

Table 2.18: (continued)

Country	% of Paper Packaging Recycled
Netherlands	70
Sweden	70
France	76
Germany	83
Austria	83
Belgium	83
Japan	98

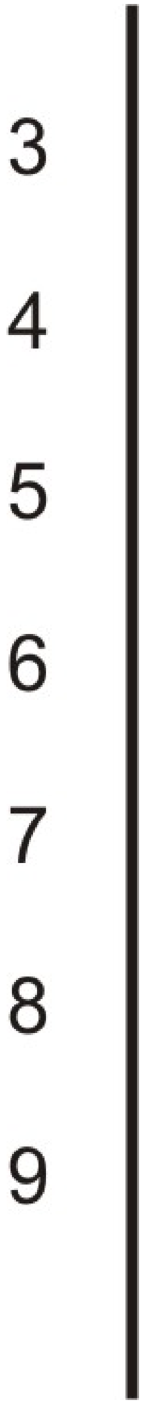
The dot plot for this data would look like this:



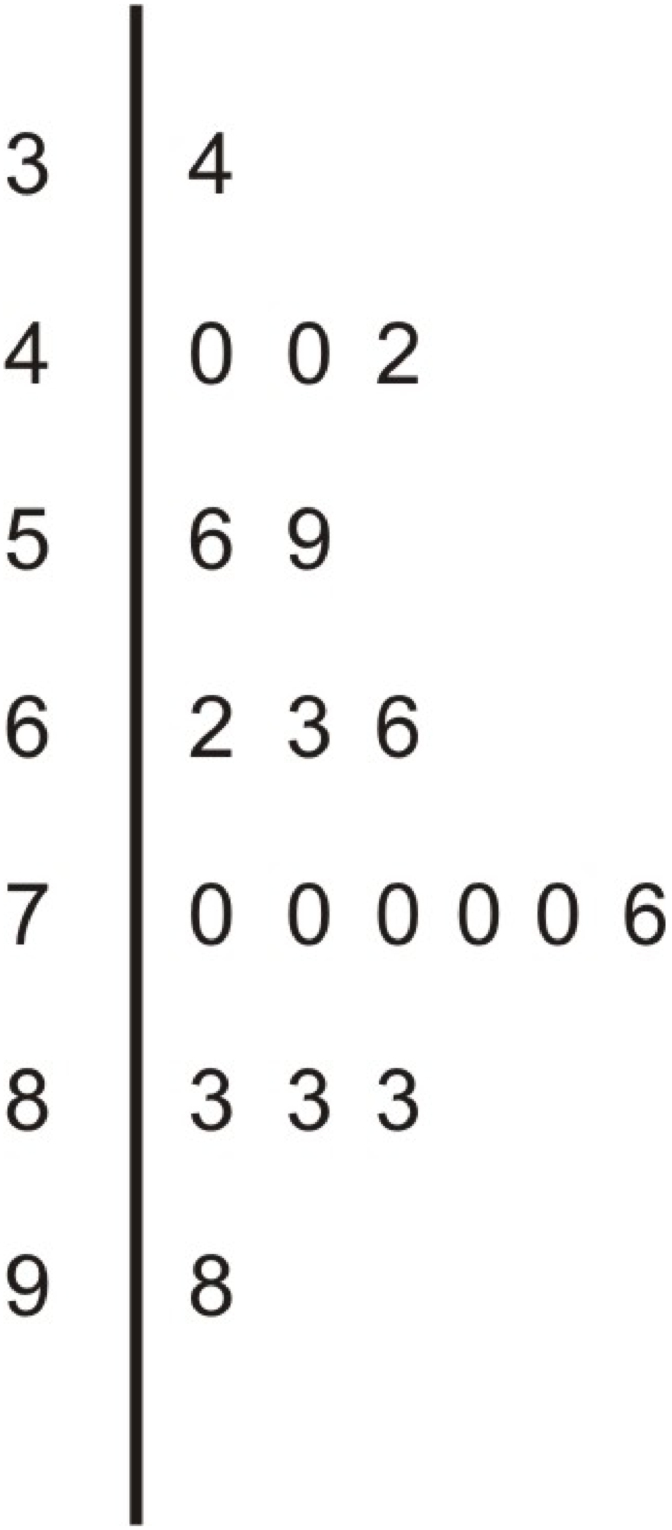
Notice that this data is centered around a manufacturing rate using recycled materials of between 65 and 70 percent. It is spread from 34% up to 98%, and appear very roughly symmetric, perhaps even slightly skewed left. Dot plots have the advantage of showing all the data points and giving a quick and 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.

Numerical Variables: Stem-and-Leaf Plots

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



Once the stems are decided, the leaves representing the one's digit and are listed in numerical order from left to right.

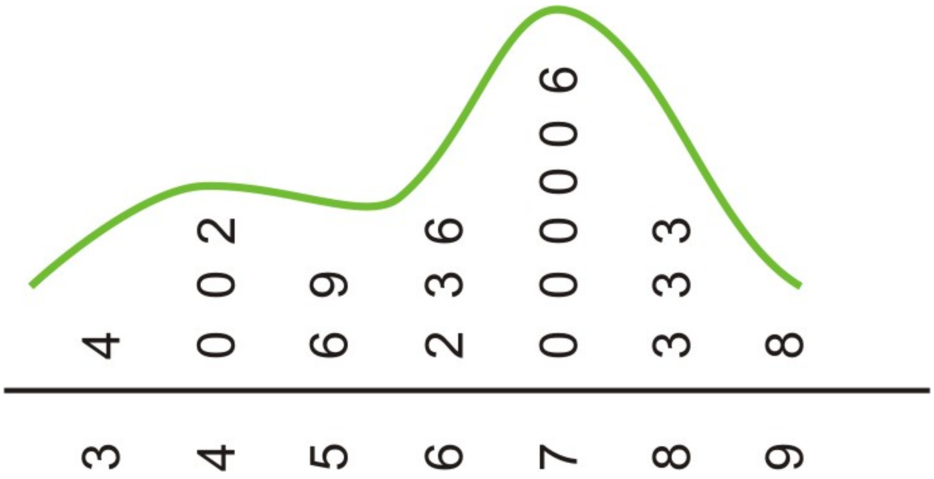


3	4
4	0 0 2
5	6 9
6	2 3 6
7	0 0 0 0 0 6
8	3 3 3
9	8

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 contains more than two digits, you will need to remove some of the information by rounding. Data that has large gaps between values can also make the stem plot hard to create and less useful when interpreting the data.



Back-to-Back Stem Plots

Stem plots can also be a useful tool for comparing two distributions when placed next to each other or what is commonly called **”back-to-back”**.

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

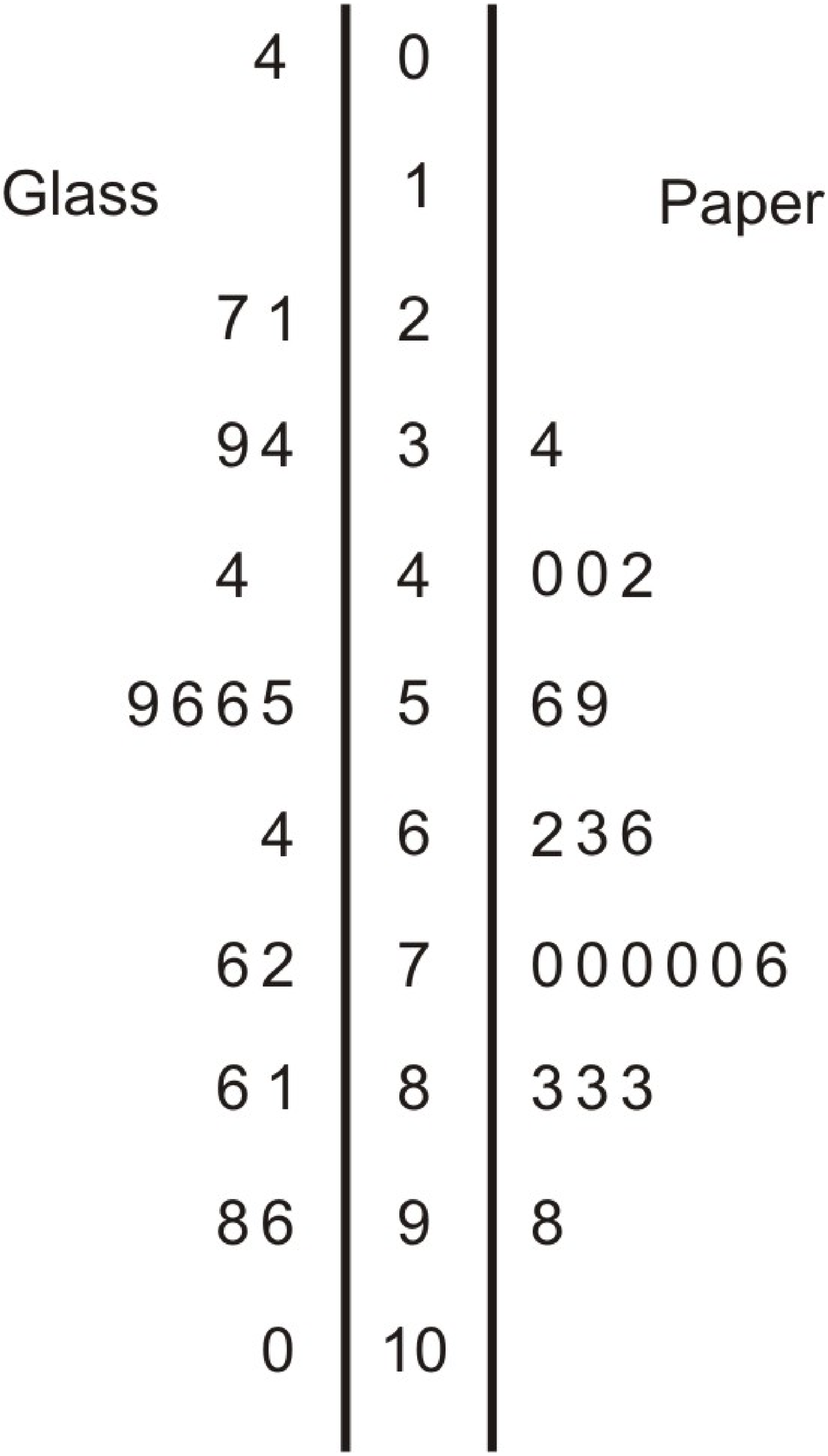
Table 2.19: **Percentage of the paper packaging used in a country that is recycled. National Geographic, January 2008. Volume 213 No.1 , pg 86-87.**

Country	% of Glass Packaging Recycled
Cyprus	4
United States	21
Poland	27

Table 2.19: (continued)

Country	% of Glass Packaging Recycled
Greece	34
Portugal	39
Spain	41
Australia	44
Ireland	55
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 is no data in a stem, you must include it to preserve the spacing or you will not get an accurate picture of the shape and spread.



Glass		Paper
4	0	
	1	
7 1	2	
9 4	3	4
4	4	0 0 2
9 6 6 5	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 had 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 their effective percentage is actually more than 100!

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. In the previous example, we 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. Here is a data table that includes both percentages.

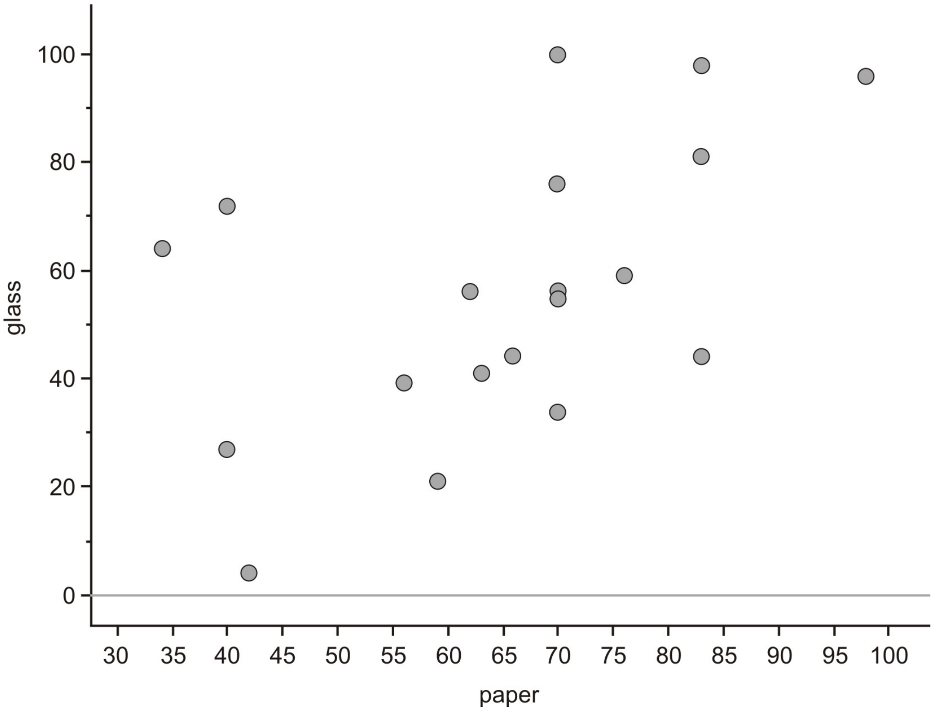
Table 2.20:

Country	% of Paper Packaging Recycled	% of Glass Packaging Recycled
Estonia	34	64
New Zealand	40	72
Poland	40	27
Cyprus	42	4
Portugal	56	39
United States	59	21
Italy	62	56
Spain	63	41
Australia	66	44
Greece	70	34
Finland	70	56
Ireland	70	55
Netherlands	70	76
Sweden	70	100
France	76	59
Germany	83	81
Austria	83	44
Belgium	83	98
Japan	98	96

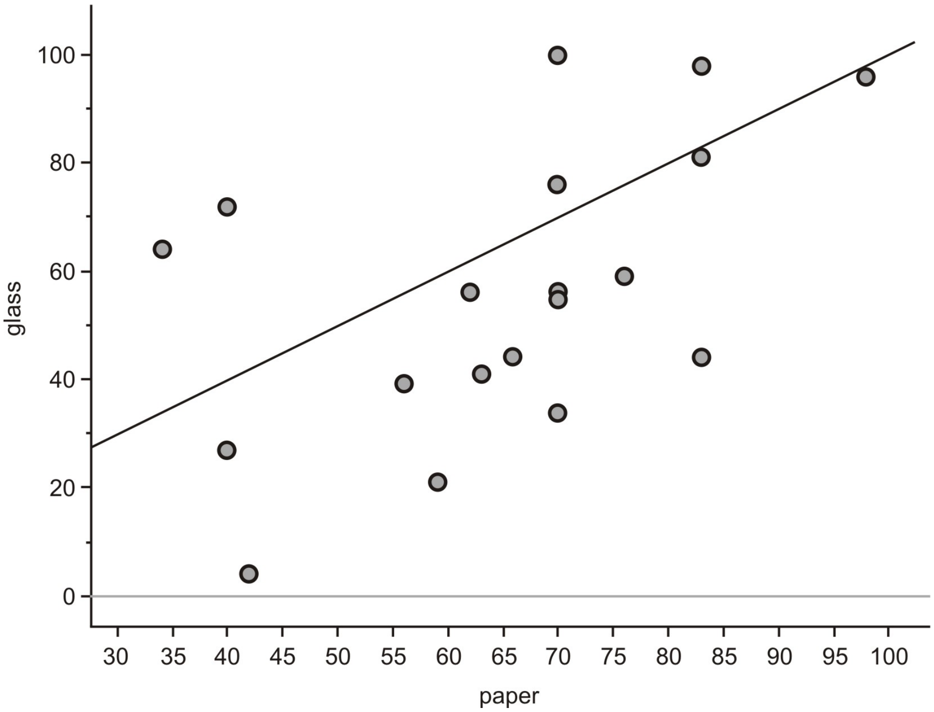
Figure: Paper and Glass Packaging Recycling Rates for 19 countries

Scatterplots

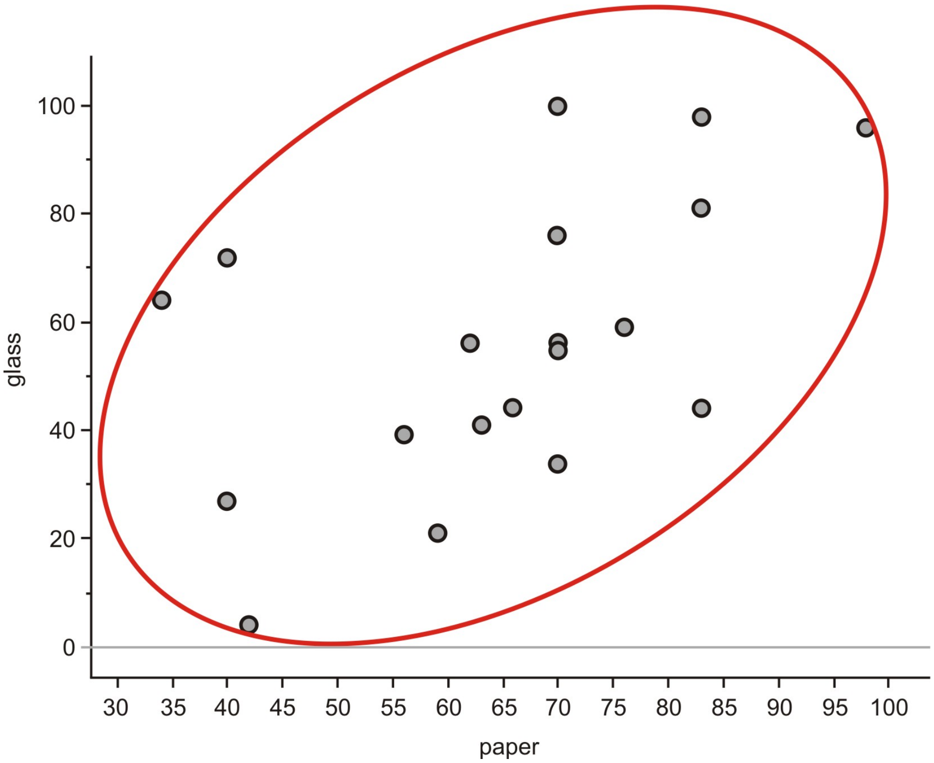
We will place the paper recycling rates on the horizontal axis, and the glass on the vertical axis. Next, plot a point that shows each country's rate of recycling for the two materials. This series of disconnected points is referred to as a **scatterplot.**



What can we learn from plotting the data in this manner? Remember 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 paper and recycling rates being equal. If all the countries had the same 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.

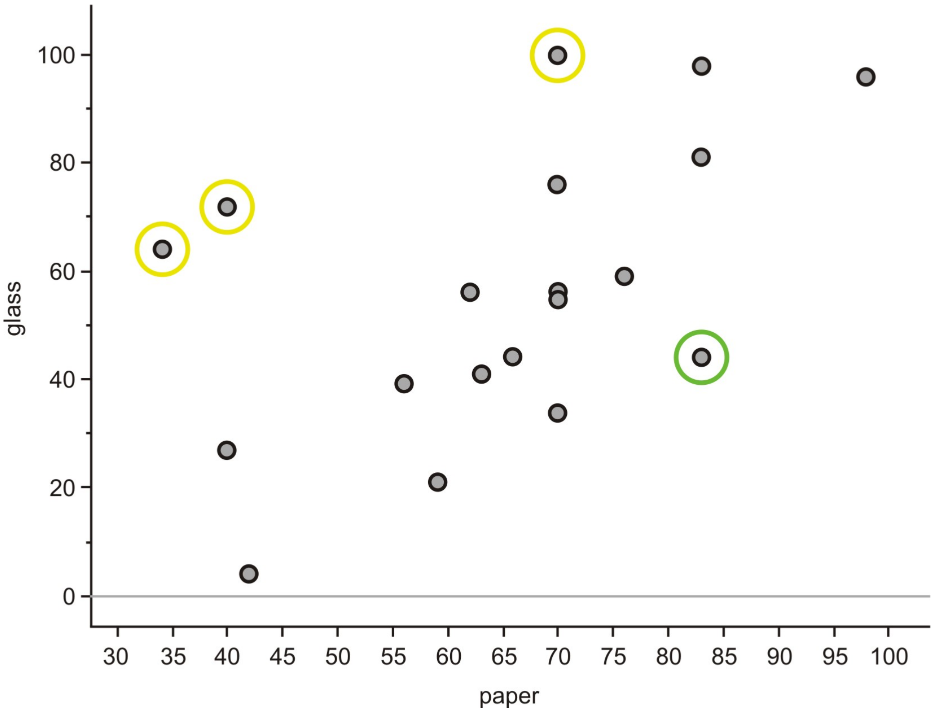


In univariate data, we are interested primarily in the ideas of shape, center, and spread to initially characterize a data set. For bivariate data, we will also discuss three important characteristics that are slightly different; **shape, direction, and strength,** to inform us about the association between the two variables. We will save formal discussions of these ideas, as well as statistics to quantify them, for a later chapter, but direction and strength are easy to introduce in this example. 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 is oriented in this manner is 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 country's with higher paper recycling rates have higher glass recycling rates. This is similar to a concept of slope in Algebra. 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 was 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 glass and paper recycling rates, the cloud of data would be much longer than it is wide. Long and narrow ellipses mean strong linear association, shorter and wider one's show a weaker linear relationship. In this example, there are some countries in 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 rates, and Austria (circled in green) is an example of a country with a much lower glass rate than their paper rate. These data points are spread away from the rest of the data enough to make the ellipse much wider, therefore weakening the association between the variables.

Explanatory and Response Variables

In this example, there was really no compelling reason to put paper on the horizontal axis and glass on the vertical. We could have learned the same information about the plot if we had switched those variables. In many data sets, however, the variables are often related in such a way that one variable appears to have an impact on the other. In the last lesson, we examined countries that are the top consumers of bottled water per person. If we compared this to the amount of plastics that these countries are disposing in landfills, it is natural to think that a higher rate of drinking bottled water could lead to a response in the amount of plastic waster created in that country. In this case we would refer to the bottled water consumed as the **explanatory variable** (also referred to in science and math as the **independent variable**). The explanatory variable should be placed on the horizontal axis. The amount of plastic waste is called the **response variable** (also referred to in science and math as the **dependent variable**), which be placed on the vertical axis. There are most likely other variables involved, like the total population, recycling rate, and consumption of other plastics, so we are not implying that the bottled water consumption is the sole cause of change in plastic waste, and without actual data it is difficult to even comment on the strength of the relationship, but it makes sense to look at the general relationship in these terms. It is very natural to think of this as a cause and effect relationship, though you will learn in a later chapter that it is very dangerous to assume such a relationship without performing a properly controlled statistical experiment.

Line Plots

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

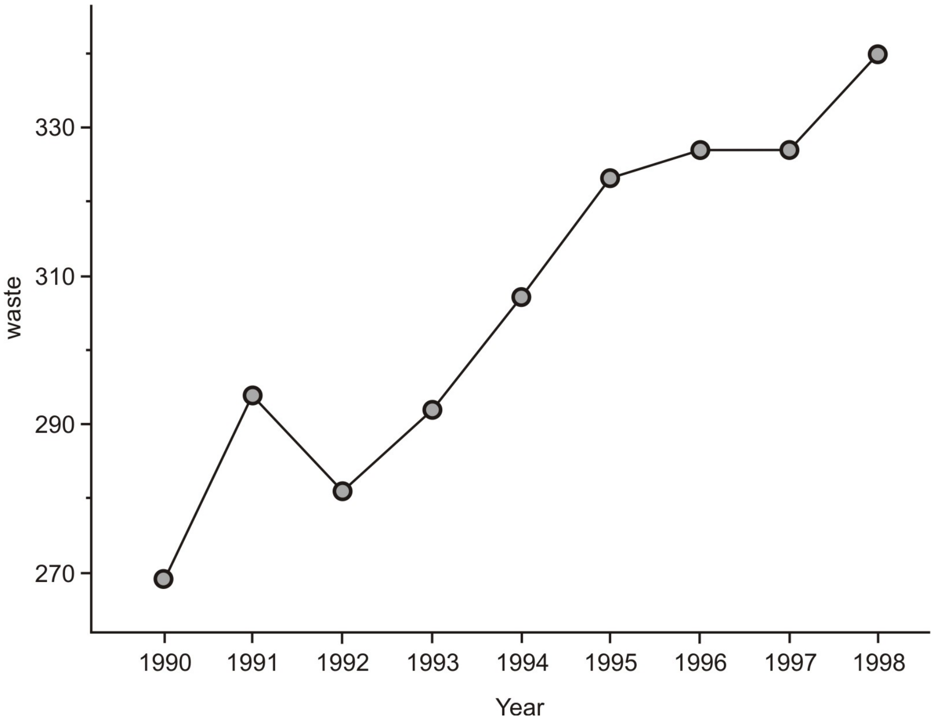
Table 2.21:

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.zerowasteamerica.org/MunicipalWasteManagementReport1998.htm

In this example, the time in years is the explanatory variable and the amount of municipal waste is the response variable. It is not the passage of time that *causes* our waste to increase. Other factors such as population growth, economic conditions, and societal habits and attitudes contribute as causes. But 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 is changing over a period of time. Here is the line plot for the US Municipal Waste data:

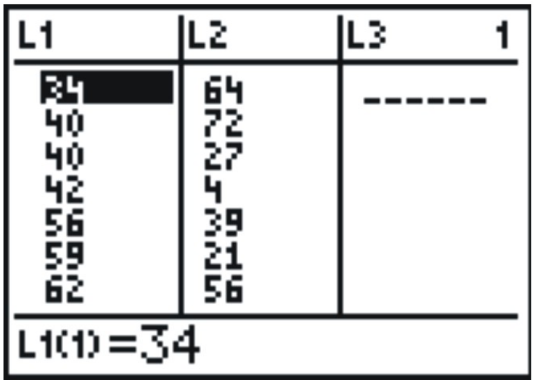


It is easy to see general trends from this type of plot. For example, we can spot the year in which the most dramatic increase occurred by looking at the steepest line (1990). We can also spot the years in which the waste output decreased and/or remained about the same (1991 and 1996). It would be interesting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

Scatterplots and Line Plots on the Graphing Calculator

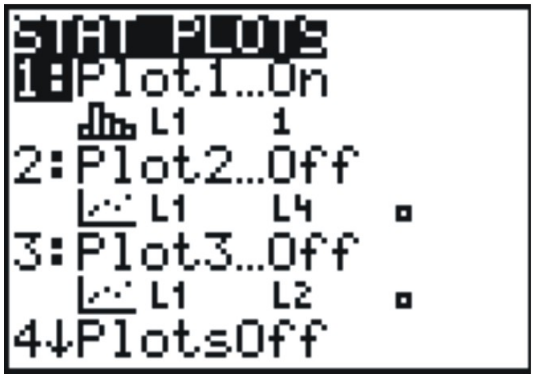
Scatterplots

Enter the data from the scatterplot example of recycling rates. Place the paper rates in L_1 and the glass rates in L_2



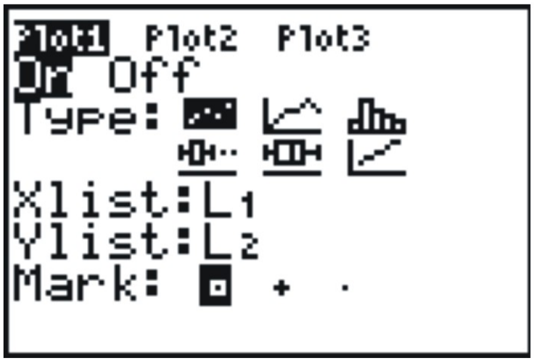
```
L1      |L2      |L3     1
 34     | 64     | ------
 40     | 72     |
 40     | 27     |
 42     | 4      |
 56     | 39     |
 59     | 21     |
 62     | 56     |
L1(1)=34
```

Next, press **2ND** [**STAT-PLOT**] to enter the STAT-PLOTS menu and choose the first plot.



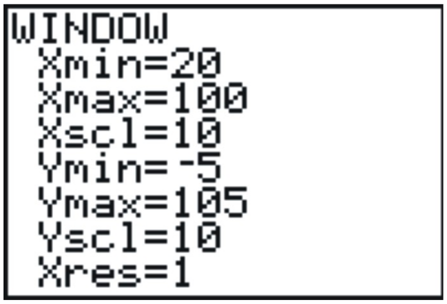
```
STAT PLOTS
1:Plot1...On
     L1    1
2:Plot2...Off
     L1    L4    □
3:Plot3...Off
     L1    L2    □
4↓PlotsOff
```

Change the settings to match the following screenshot:

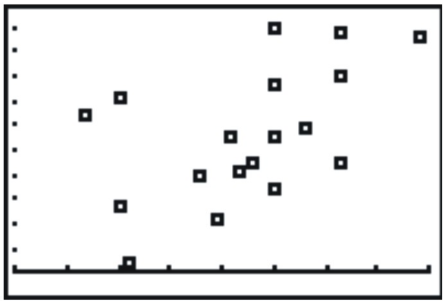


```
Plot1  Plot2  Plot3
On  Off
Type:
Xlist:L1
Ylist:L2
Mark:  □  +  ·
```

This selects a scatterplot with the explanatory variable in L_1 and the response variable in L_2. In order to see the points better, you should choose either the square or the plus sign for the mark. Finally, set an appropriate Window to match the data. In this case, we looked at our lowest and highest percentages in each variable, and added a bit of room to create a pleasant window. Press [**GRAPH**] to see the result, Which is shown below.

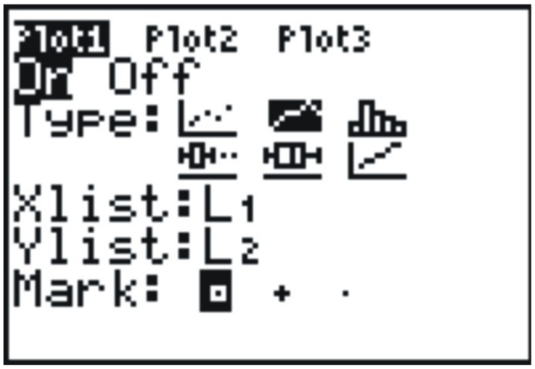


```
WINDOW
 Xmin=20
 Xmax=100
 Xscl=10
 Ymin=-5
 Ymax=105
 Yscl=10
 Xres=1
```



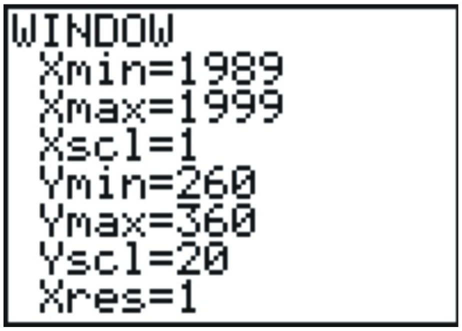
Line Plots

Your graphing calculator will also draw a line plot and the process is almost identical to that for creating a scatterplot. Enter the data from the US Municipal waste example into your lists. The only change that you need to make is to choose a line plot in the Plot1 menu.

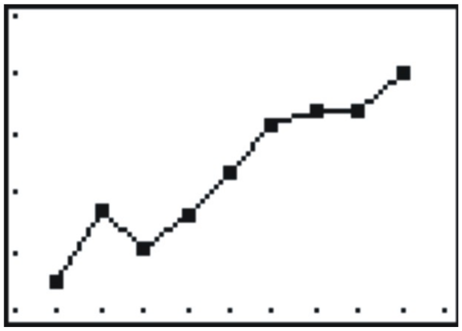


```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: ▫ + ·
```

Set an appropriate window, and graph the resulting plot.



```
WINDOW
 Xmin=1989
 Xmax=1999
 Xscl=1
 Ymin=260
 Ymax=360
 Yscl=20
 Xres=1
```



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 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, or **weak** if the points appear more randomly scattered. If the values of the response variable generally increase as the values of the explanatory variable also increase, the data has a **positive association.** If the response variable generally decreases as the explanatory variable increases, the data has a **negative association.** In a **line graph,** there is significance to the change between consecutive points so those points are connected. Line graphs are used often when the explanatory variable is time.

Points to Consider

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

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 2.22:

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.puducherry.gov.in/envisnew/INDUSTRIAL%20SOLID%20WASTE.htm

(a) Create a bar graph for this data.

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

Table 2.23:

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 2.24:

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

Table 2.24: (continued)

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

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

(a) Create a dot plot for this data.

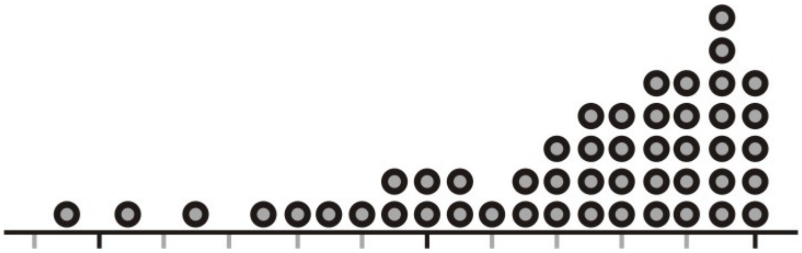
(b) Discuss the shape, center, and spread of this distribution.

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

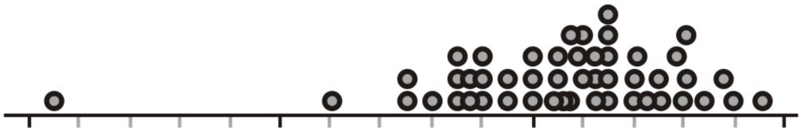
(d) Use your stem and leaf plot to find the median percentage for this data.

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

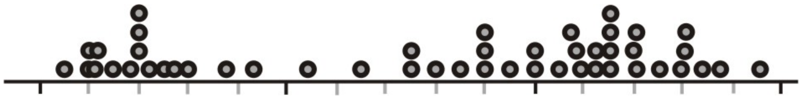
(a)



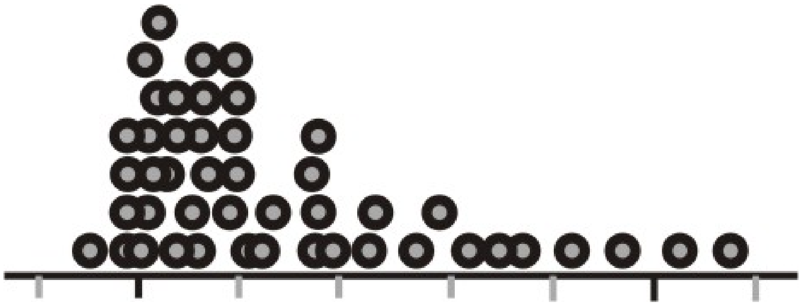
(b)



(c)

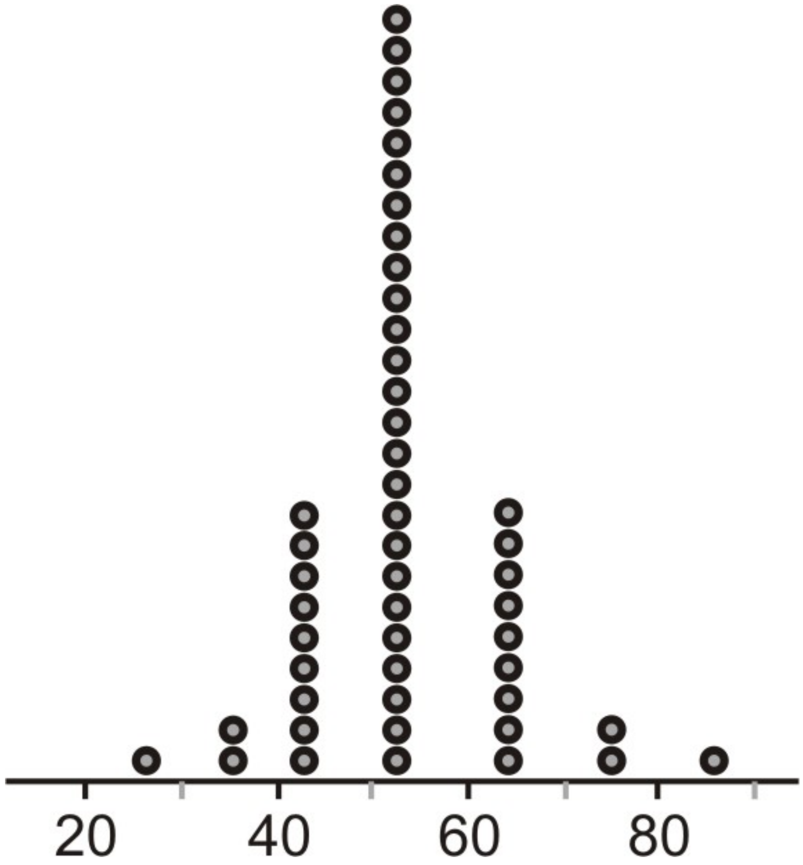


(d)

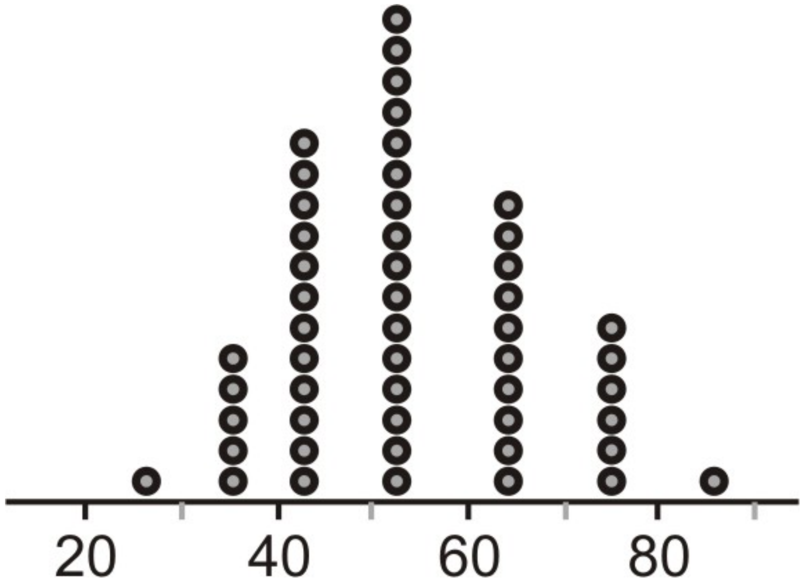


Questions 4 and 5 refer to the following dot plots:

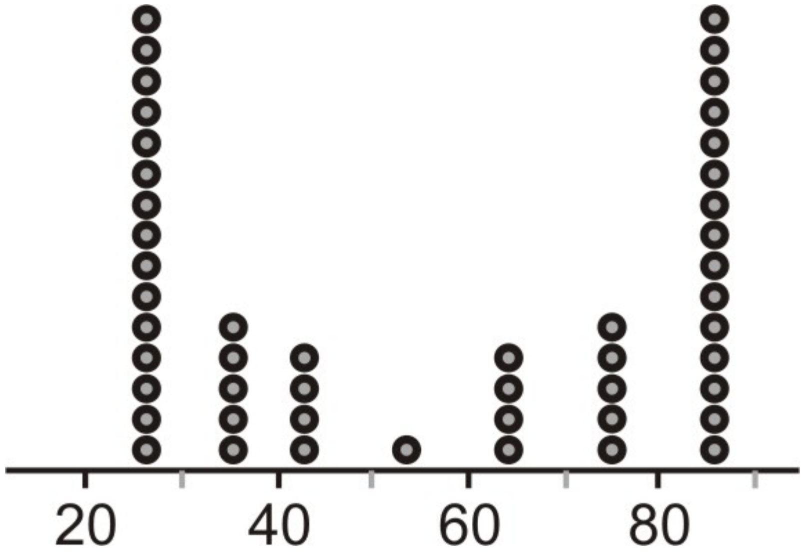
(a)



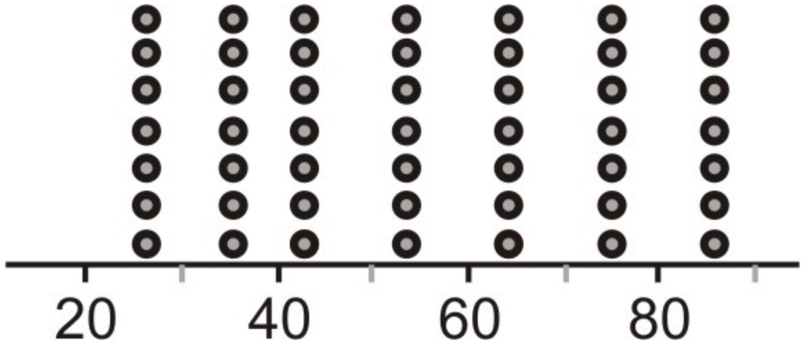
(b)



(c)



(d)



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 is the data including both variables.

Table 2.25:

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

Table 2.25: (continued)

State	Percentage	Total Amount of Municipal Waste in Thousands of Tons
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
South Carolina	34	8361
South Dakota	42	510
Tennessee	40	9496
Utah	19	3760

Table 2.25: (continued)

State	Percentage	Total Amount of Municipal Waste in Thousands of Tons
Vermont	30	600
Virginia	35	9000
Washington	48	6527
West Virginia	20	2000
Wisconsin	36	3622
Wyoming	5	530

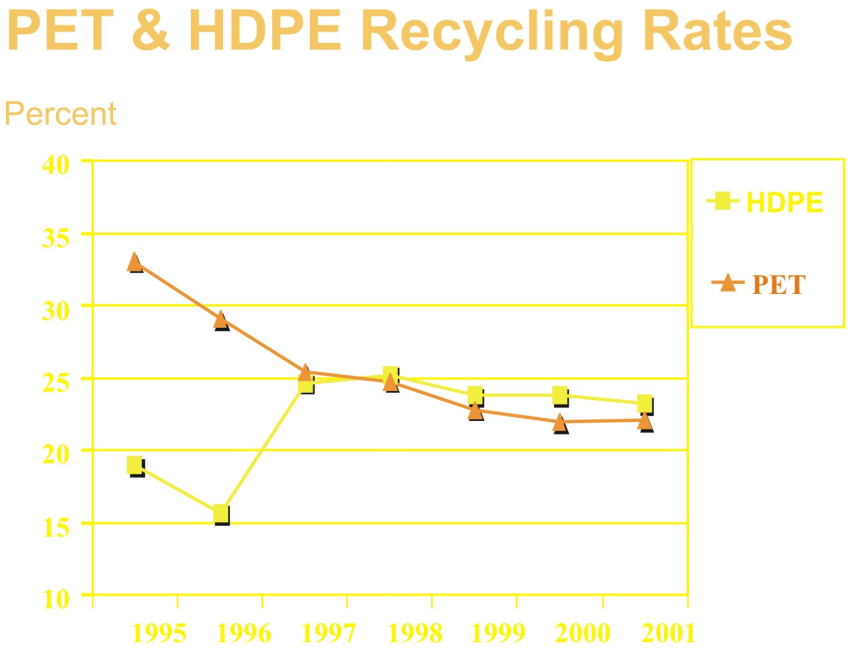
(a) Identify the variables in this example and identify which one is the explanatory and which one is the response variable.

(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.

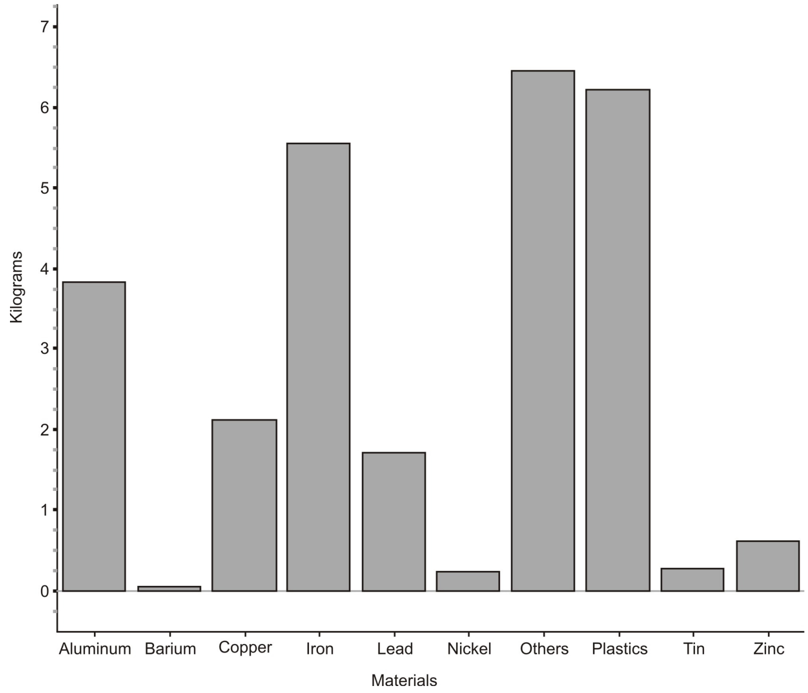


Source: National Association for PET Container Resources, American Plastics Council

(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?

Review Answers

1. (a)



(b) Divide each weight by 27.

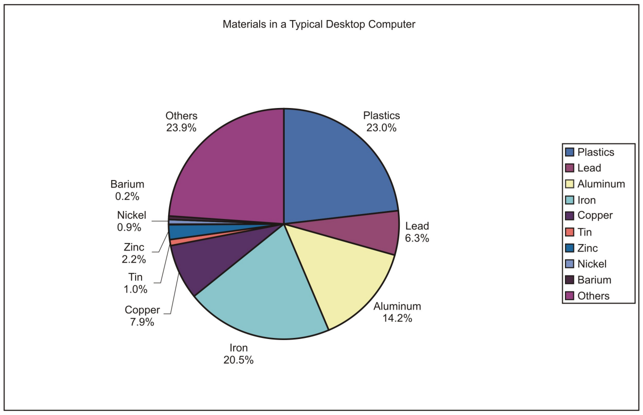
Table 2.26:

Material	Kilograms	Approximate Percentage of Total Weight
Plastics	6.21	23
Lead	1.71	6.3
Aluminum	3.83	14.2
Iron	5.54	20.5
Copper	2.12	7.8
Tin	0.27	1
Zinc	0.60	2.2
Nickel	0.23	0.8
Barium	0.05	0.2
Other elements and chemicals	6.44	23.9

Table 2.26: (continued)

Material	Kilograms	Approximate Percentage of Total Weight

The data from the table has been adjusted a bit from the original source to simplify the problem and to account for rounding errors. a.



2. (a)

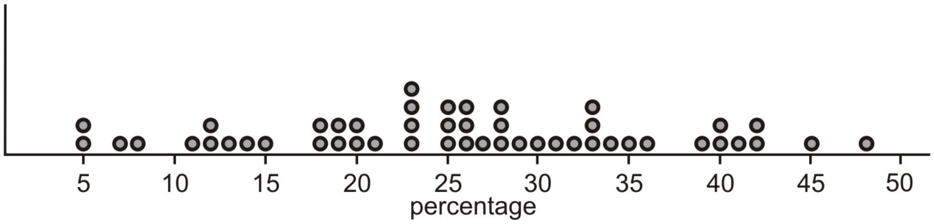


Figure: Percentage of Municipal Waste Recycled by US States, 1998

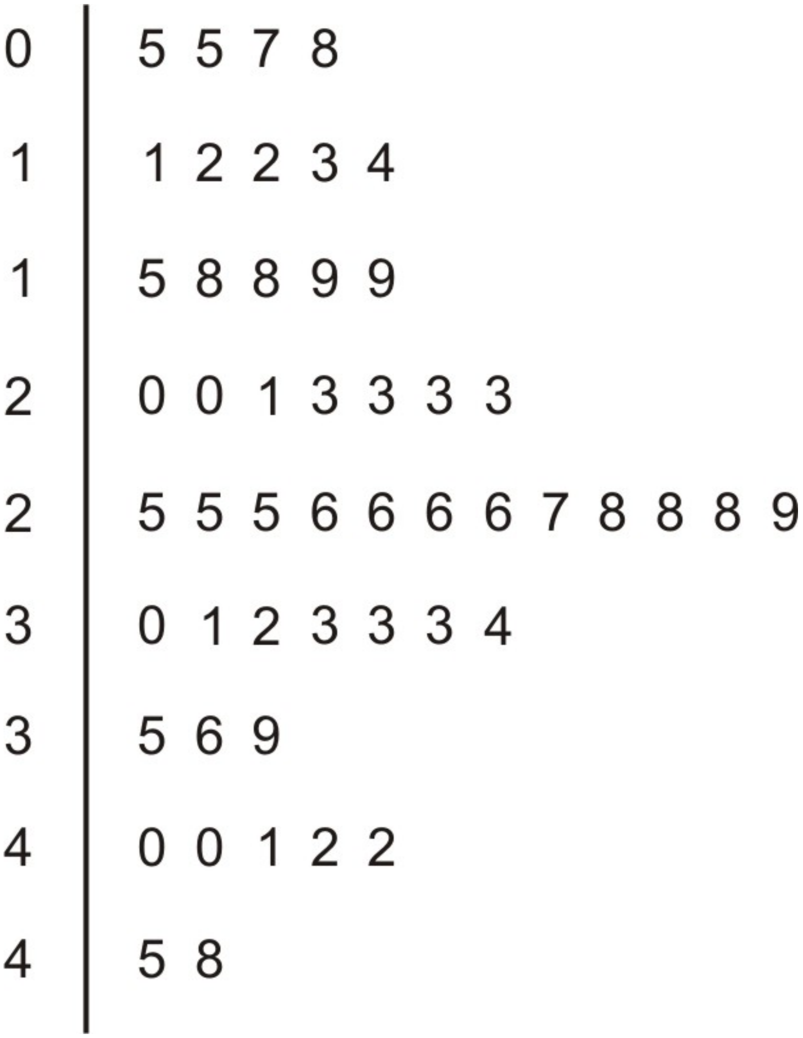
(b) This data is fairly symmetric, is centered around 25 or 26 percent, and has a spread of 43 percentage points (the range).

(c)

0	5 5 7 8
1	1 2 2 3 4 5 8 8 9 9
2	0 0 1 3 3 3 3 5 5 5 6 6 6 6 7 8 8 8 9
3	0 1 2 3 3 3 4 5 6 9
4	0 0 1 2 2 5 8

1 | 3 means 13 percent

In this example, we chose the stems to represent every 10 percentage points, which compresses the data and forfeits some of the information about the shape of the distribution. Instead, we can split the stems in half $(10 - 14$, and $15 - 19)$. This plot is much more informative about the true shape of the data.



0	5 5 7 8
1	1 2 2 3 4
1	5 8 8 9 9
2	0 0 1 3 3 3 3
2	5 5 5 6 6 6 6 7 8 8 8 9
3	0 1 2 3 3 3 4
3	5 6 9
4	0 0 1 2 2
4	5 8

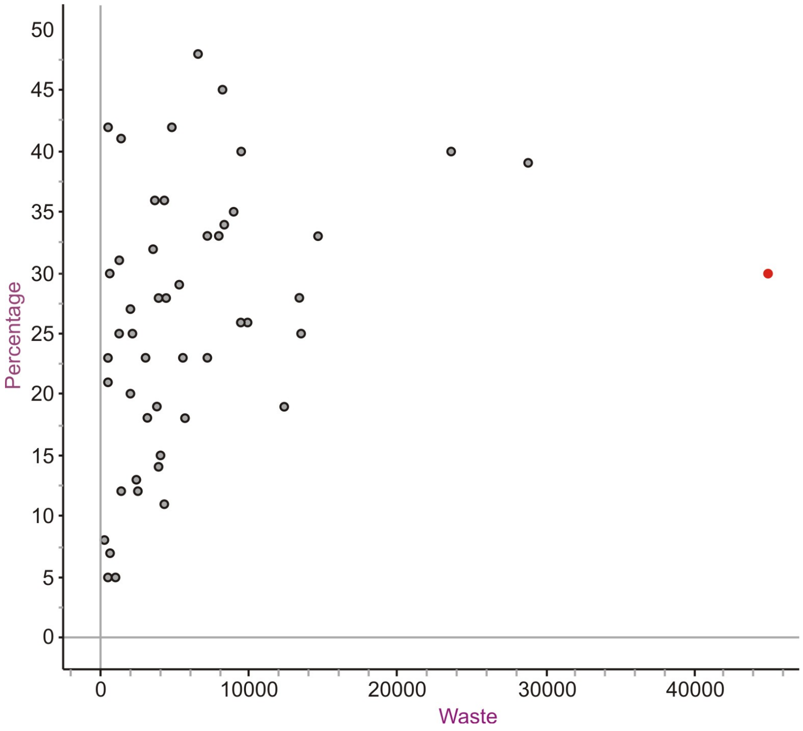
(d) There are 49 data points, so the 25^{th} value, counted from either end, is the median. In this case, that is 26.

3. (a) This data set is mound-shaped and skewed left.
   (b) This data set has one obvious outlier. The remainder of the data is mound-shaped and fairly symmetric.
   (c) This data appears to be bimodal.
   (d) This data set appears mound-shaped and skewed right.
4. (a) mound-shaped and symmetric
   (b) mound-shaped and symmetric
   (c) bimodal
   (d) uniform

5. All four distributions have almost the same center, which appears to be around 52.
6. The distribution is symmetric, so the center should be very close, if not equal to the mean. Standard deviation is a measure of the typical distance away from the mean. Most of the data points in this distribution are concentrated very close to the center.
7. c. The bimodal shape means that most of the points are located at the extreme values making their distance from the mean greater.
8. (a) total amount of municipal waste: explanatory variable percentage of total waste recycled: response variable In this situation, we would most likely be interested in showing if states that have more or less total waste would have different recycling performance, i.e. does municipal waste *explain* a response in the recycling rate.

   (b) $13,386,000$ tons. The data is indicated to be recorded in thousands of tons.

   (c)



   (d) There does appear to be at least one obvious outlier. California creates by far the most waste, most likely due to its large population. If we ignore California, there appears to be only a weak positive association between the amount of waste and the percentage recycled. If we removed two other points that appear atypical, New York and Florida, there is almost no association between the two. States with low average waste creation have recycling rates varying from the lowest, to the highest rates. If we remove the three potential outliers and rescale the axes, the data cloud is almost a circle, showing virtually no association.

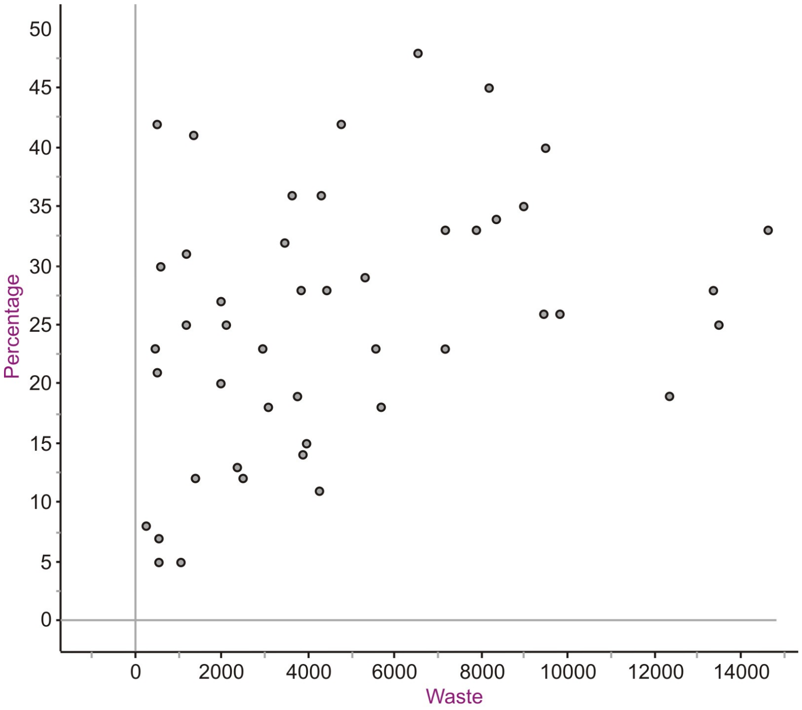


Figure 44

9. (a) HDPE plastic recycling showed a dramatic growth in 1996 and a slight growth the following year, but decline for all other years. PET bottle recycling has declined steadily through the entire time range.
   (b) The recycling rate declined by approximately 20% from 1995 to 2001.
   (c) One contributing cause has been the dramatic growth of immediate use personal size containers like bottled water that are typically consumed away from home and are not as likely to end up in curbside recycling programs.
   (d) The decline was the greatest during 1995 and 1996.

References

National Geographic, January 2008. Volume 213 No.1

- [1]http://www.etoxics.org/site/PageServer?pagename=svtc_global_ewaste_crisis'
- http://www.earth-policy.org/Updates/2006/Update51_data.htm

2.3 Box-and-Whisker Plots

Learning Objectives

- Calculate the values of the $5-$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 we will round out our investigation of different types of visual displays by introducing the box-and-whisker plots. The basic ideas of shape, center, spread, and outliers will be investigated in this context and students will be asked to become proficient in translating and interpreting graphs of univariate data of the various types.

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.

In order to review finding these values from Chapter One, let's turn to another recycling/conservation related issue. The huge population growth in the western United States in recent years, along with a trend toward less annual rainfall in many areas and even drought conditions in others, has put tremendous strain on the water resources available now and the need to protect them in the years to come. Here is a listing of the reservoir capacities of the major water sources for Arizona:

Table 2.27:

Lake/Reservoir	% of Capacity
Salt River System	59
Lake Pleasant	49
Verde River System	33

Table 2.27: (continued)

Lake/Reservoir	% of Capacity
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 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:

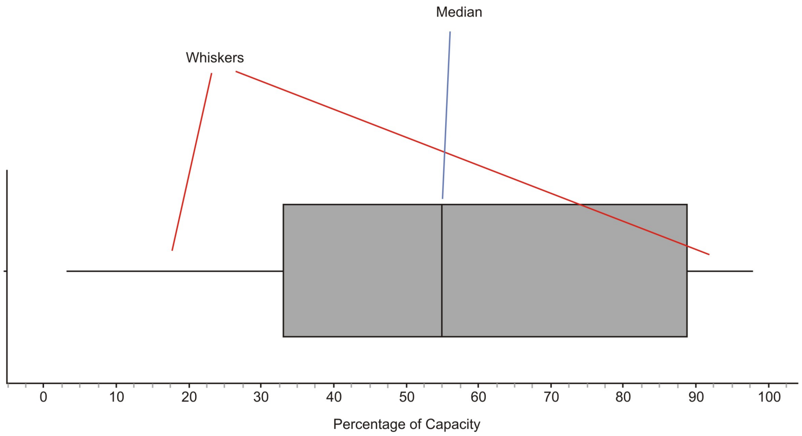
$$3, 9, 33, 49, 51, 59, 85, 89, 95, 98$$

With 10 numbers, the median would be between 51 and 59, or 55. Recall that the lower quartile is the 25^{th} percentile, or where 25% of the data is below that value. In this data set, that number is 33. Similarly, the upper quartile is 89. Therefore the five-number summary is:

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

Box-and-Whisker Plots

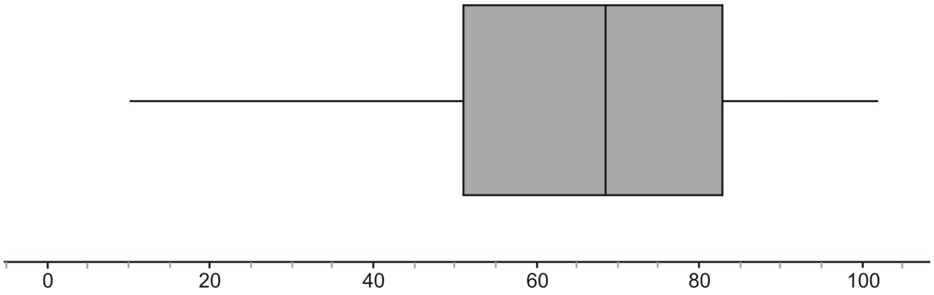
A **box-and-whisker plot** is a very convenient and informative way to represent single-variable data. To create the "box" part of the plot, draw a rectangle that extends from the lower quartile to the upper quartile. Draw a line through the interior of the rectangle at the median. Then we connect the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



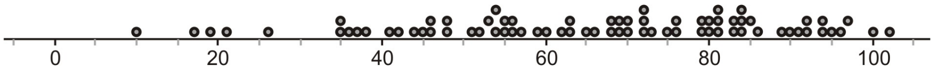
The plot divides the data into quarters. If the number of data points is divisible by 4, then there will be exactly the same number of values in each of the two whiskers, as well as the two sections in the box. In this example, because there are 10 data points, it will only be approximately the same, but approximately 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. Of course, it could be uniform or bimodal also, so you would have to also look at a dot plot or histogram to get a more complete picture of the shape.

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. Many students initially incorrectly interpret this to mean that longer sections contain more data and shorter ones contain less. This is not true and it is important to remember that roughly **the same amount of data is in each section.** What this does tell us is how the data is **spread** in each of those sections. 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:

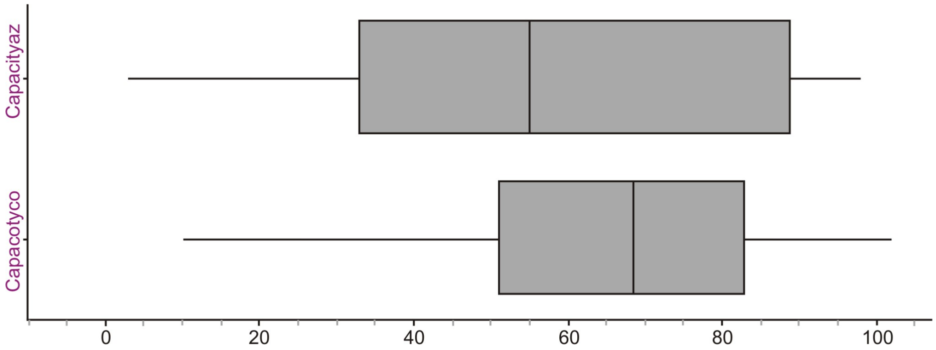


In this case, the third quarter of data (between the median and upper quartile), appears to be a bit more densely concentrated in a smaller area. The data in the lower whisker also appears to be much more widely spread than it is in the other sections. Looking at the dot plot for the same data shows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



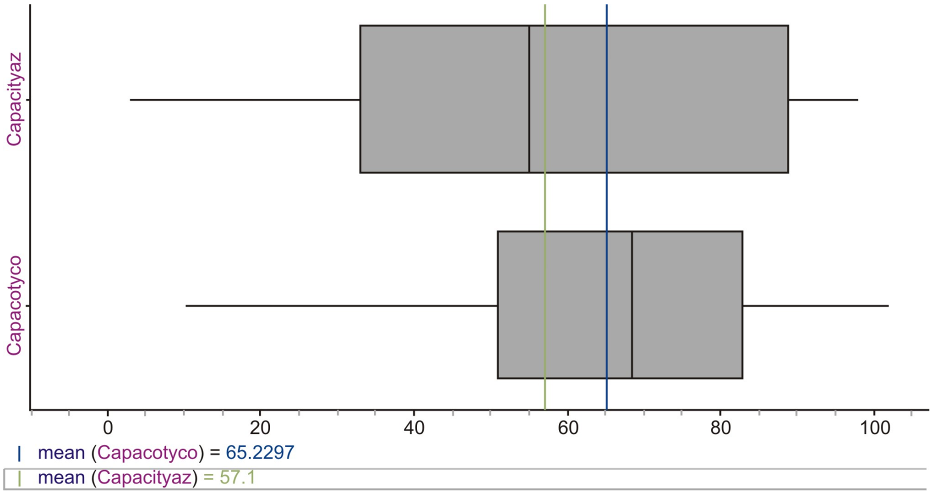
Comparing Multiple Box Plots: Resistance Revisited

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 appears to have a wider spread. The center of the Colorado data (as evidenced by the location of the median) is higher, which would tend to indicate that, in general, Arizona's capacities are lower. In the first chapter we talked about the concept of **resistance**. Recall that the median is a resistant measure of center because it is not affected by outliers, but the mean is not resistant because it will be pulled toward outlying points. This is also true of skewed data. 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.



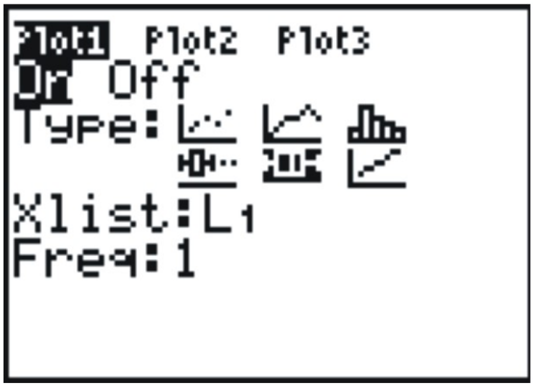
Notice that the long left whisker in the Colorado data causes the mean to be pulled toward the left, making it lower than the median. In the Arizona plot, you can see that the mean is slightly higher than the median due to the slightly elongated right side of the box. If these data sets were perfectly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

Outliers in Box-and-Whisker Plots

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

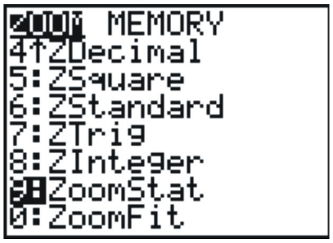
$$80, 83, 77, 95, 85, 74, 34, 68, 90, 82, 75$$

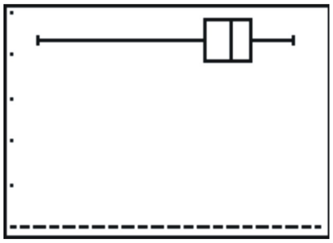
At first glance, the 34 should stand out. It appears as if this point is significantly isolated from the rest of the data, which is the textbook definition of an outlier. Let's use a graphing calculator to investigate this plot. Enter your data into a list as we have done before, and then choose a plot. 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).



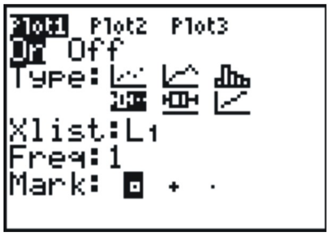
```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
```

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 select number 9 (ZoomStat).

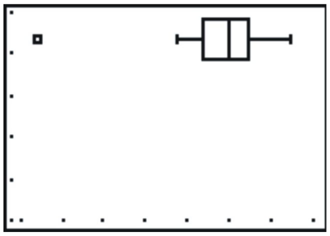




While box plots give us a nice summary of the important features of a distribution, we lose the ability to identify individual points. The left whisker is elongated, but if we did not have the data, we would not know if all the points in that section of the data were spread out, or if it were just the 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 press then graph it.

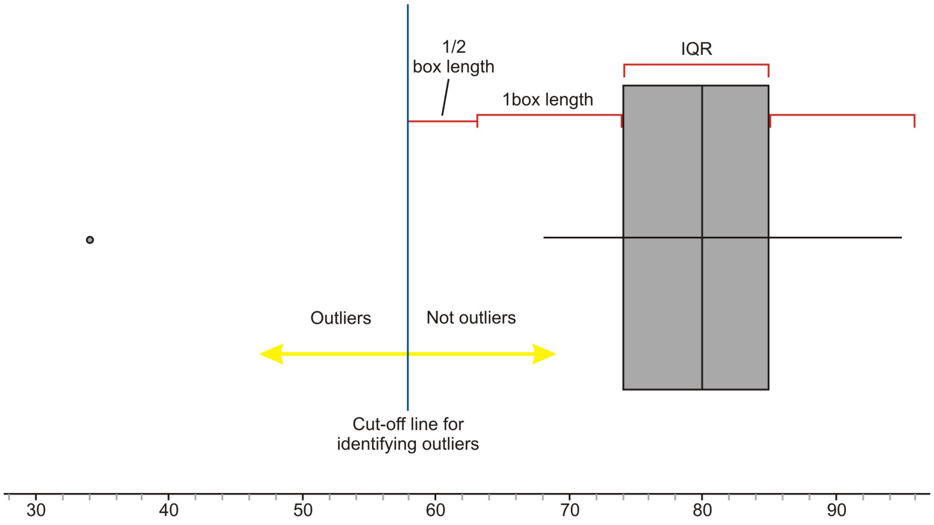


```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: ▪ + ·
```



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 1 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

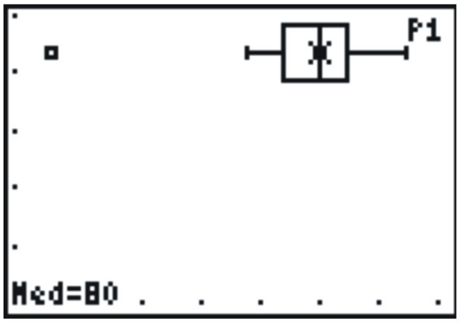
Interquartile range(IQR): $85 - 74 = 11$

$1.5 * \text{IQR} = 16.5$

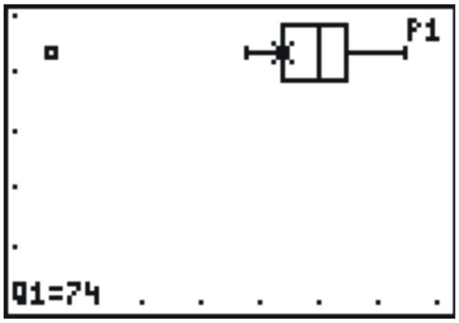
Cut-off for outliers in left whisker: $74 - 16.5 = 57.5$

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.

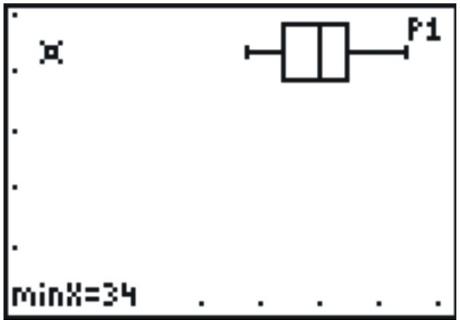
If you press [**ZOOM**], and use the left or right arrows, the calculator will trace the values of the five-number summary, as well as the last point on the left whisker.



Med=80



Q1=74



minX=34

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 2.28:

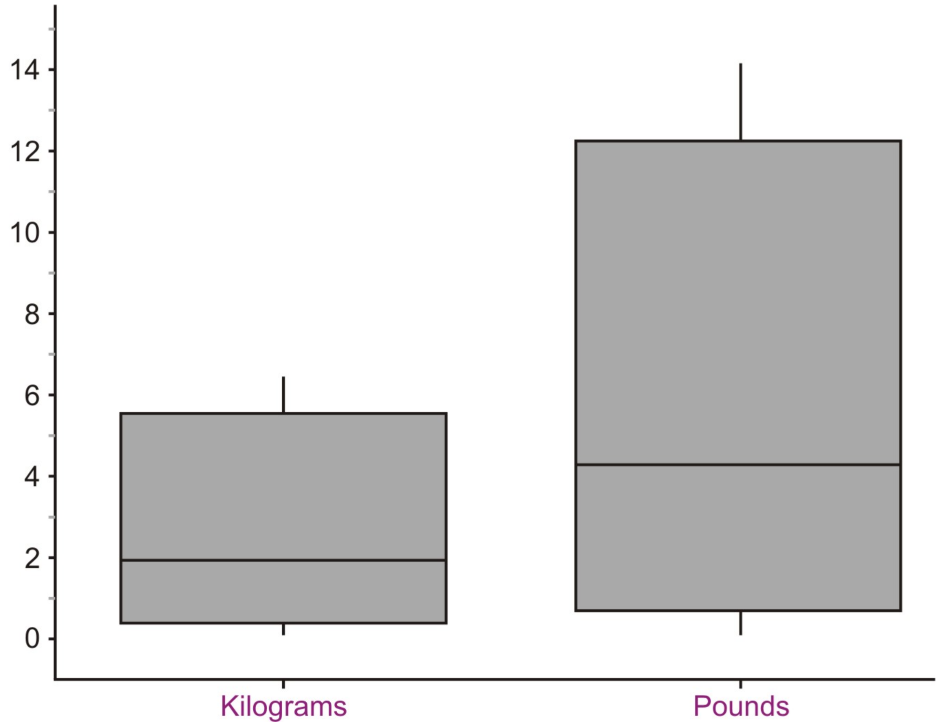
Material	Kilograms
Plastics	6.21
Lead	1.71
Aluminum	3.83
Iron	5.54
Copper	2.12
Tin	0.27
Zinc	0.60
Nickel	0.23
Barium	0.05
Other elements and chemicals	6.44

Here is a similar set of data given in pounds.

Table 2.29:

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

The source of this data set was in India, so like much of the rest of the world, the data was given in metric units, or kilograms. If we want to convert these weights to pounds, what would be different about this distribution? To convert from kilograms to pounds, we multiply the number of kilograms times 2.2. Think about how, if at all, the shape, center, and spread would change. If you multiple all values by a factor of 2.2, then the calculation of the mean would also be multiplied by 2.2, so the center of the distribution should be increased by the same factor. Similarly, calculations of the range, interquartile range, and standard deviation will also be increased by the same factor. So the center and the measures of spread will increase proportionally. This should result in the graph maintaining the same shape, but being stretched out or elongated. Here are the side-by-side box plots for both distributions showing the effects of changing units.



Lesson Summary

The **five-number** summary is 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 (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. 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 units** of a distribution, the center and spread will be affected, but the shape will stay the same.

Points to Consider

1. 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?
2. Which plots are most useful to interpret the ideas of shape, center, and spread?
3. What effects do other transformations of the data have on the shape, center, and spread?

Review Questions

1. Here is 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 $1.5 * \text{IQR}$ 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 is 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 $1.5 * \text{IQR}$ 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 approximately 4 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 2.30:

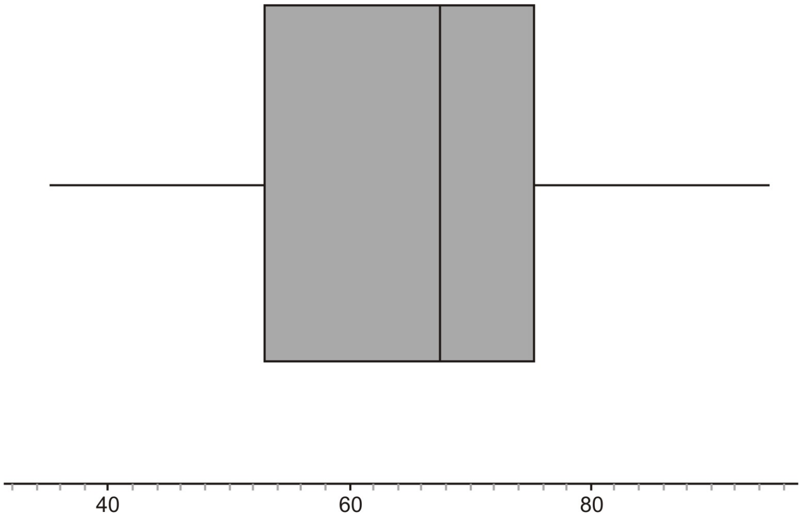
State	Average Price of a Gallon of Gasoline (US$)	Average Price of a Liter of Gasoline (US$)
Alaska	3.458	
Washington	3.528	
Idaho	3.26	
Montana	3.22	
North Dakota	3.282	
Minnesota	3.12	
Michigan	3.352	
New York	3.393	
Vermont	3.252	
New Hampshire	3.152	
Maine	3.309	

Average Prices of a Gallon of Gasoline on March 16, 2008

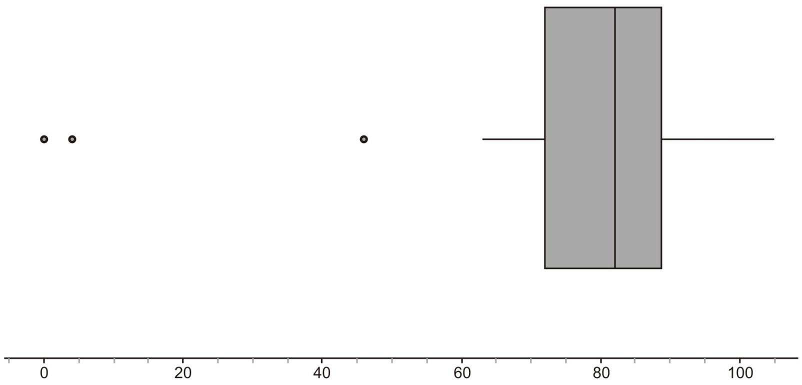
Figure: Average prices of a gallon of gasoline on March 16, 2008. *Source:* AAA, http://www.fuelgaugereport.com/sbsavg.asp

Review Answers

1. (a) $35, 53, 67.5, 75.5, 95$
   (b) $\text{IQR} = 22.5$.
   $1.5 * \text{IQR} = 33.75$
   Upper bound for outliers $> 33.75 + 75.5 = 109.25$. There is no data above 95 Lower bound for outliers $< 53 - 33.75 = 19.25$. There is no data below 35, so there are no outliers.
   (c)

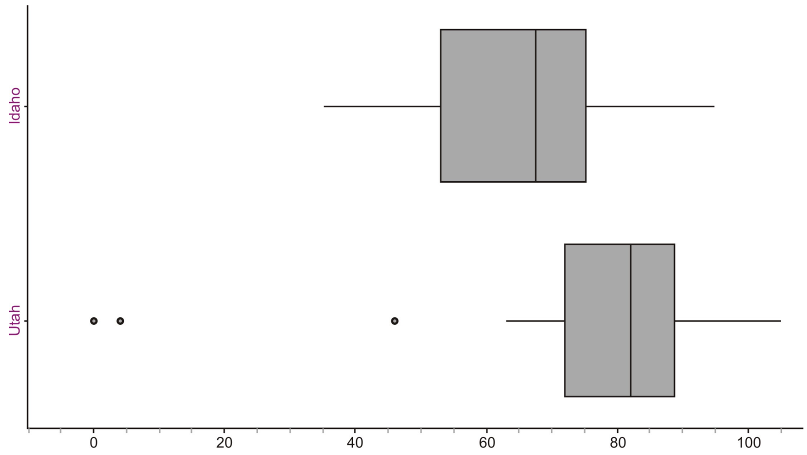


   (d) The distribution of Idaho reservoir capacities is roughly symmetric and is centered somewhere in the middle 60 percents. The capacities range from 35 up to 95 and the middle 50% of the data is between 53 and 75.5%.
   (e) The data between the median and the upper quartile is slightly more compressed which causes the median to be slightly larger than the median. The mean is approximately 65.7.
2. (a) $\{0, 72, 82, 89, 105\}$
   (b) $\text{IQR} = 17$.
   $1.5 * \text{IQR} = 25.5$
   Upper bound for outliers $> 25.5 + 89 = 114.5$. There is no data above 105 Lower bound for outliers $< 72 - 25.5 = 49.5.63$ is the last value that is above this point, so $0, 4$, and 46 are all outliers 35.
   (c)



(d) The distribution of Utah reservoir capacities has three outliers. If those points were removed, the distribution is roughly symmetric and is centered somewhere in the low 80 percents. The capacities range from 0 up to 105 including the outliers and the middle 50% of the data is between 72 and 89%.

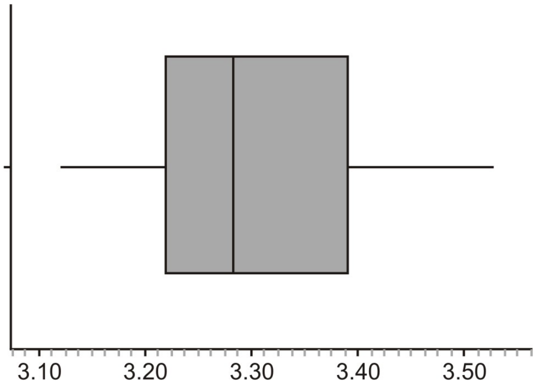
(e) There are three extreme outliers, the mean is not resistant to the pull of outliers and there will be significantly lower than the median. The mean is approximately 75.4.



3. If we disregard the three outliers, the distribution of water capacities in Utah is higher than that of Idaho at every point in the five number summary. From this we might conclude it is centered higher and that the reservoir system is overall at safer levels in Utah, than in Idaho. Again eliminating the outliers, both distributions are roughly symmetric and their spreads are also fairly similar. The middle 50% of the data for Utah is just slightly more closely grouped.
4. e. If the mean is greater than the median, then it has been pulled to the right either by an outlier, or by a skewed right shape. The median will not be affected by either of those things.
5. (a) $3.12, 3.22, 3.282, 3.393, 3.528$
   (b) $\text{IQR} = .173$
   $1.5 * \text{IQR} = .2595$
   Upper bound for outliers $> .2595 + 3.393 = 3.6525$. There is no data above 3.528
   Lower bound for outliers $< 3.22 - .2595 = 2.9605$. There is no data below 3.12, so

there are no outliers.

(c)

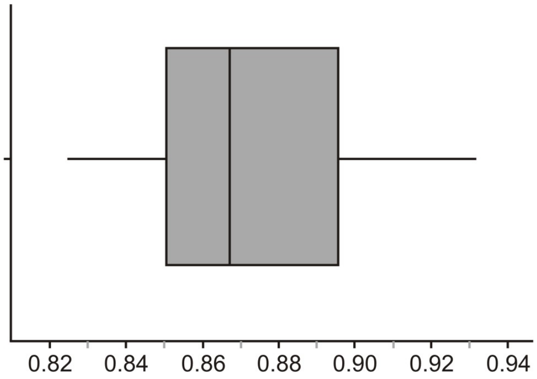


(d) By dividing the data by 3.7854, we will obtain the average cost per liter. The mean, median will be decreased, being divided by 3.7854. The same is true for the measures of spread (range, IQR, and standard deviation), which will result in the data being compressed into a smaller area if we were to graph both distributions on the same scale. The shape of the distributions will remain the same.

(e)

Table 2.31:

State	Average Price of a Gallon of Gasoline (US$)	Average Price of a Liter of Gasoline
Alaska	3.458	0.914
Washington	3.528	0.932
Idaho	3.26	0.861
Montana	3.22	0.851
North Dakota	3.282	0.867
Minnesota	3.12	0.824
Michigan	3.352	0.886
New York	3.393	0.896
Vermont	3.252	0.859
New Hampshire	3.152	0.833
Maine	3.309	0.874



References

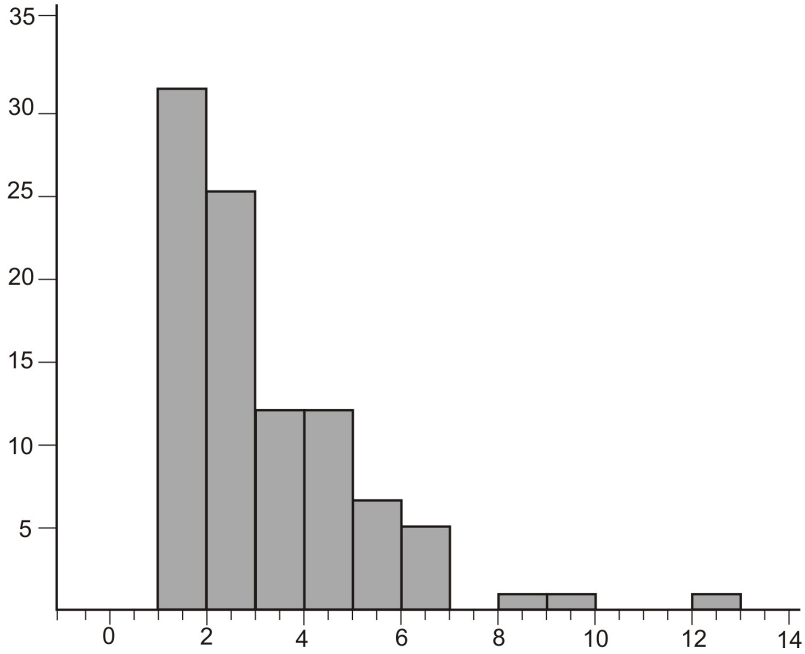
[1] Kunzig, Robert. Drying of the West. National Geographic, February 2008, Vol. 213, No. 2, Page 94.

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

2.4 Chapter Review

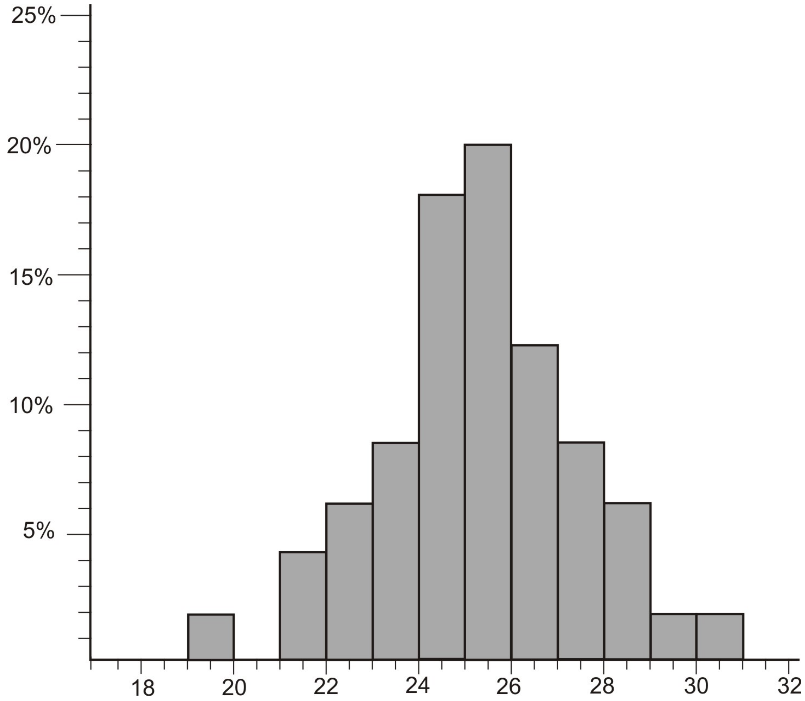
Part One: Questions

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



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

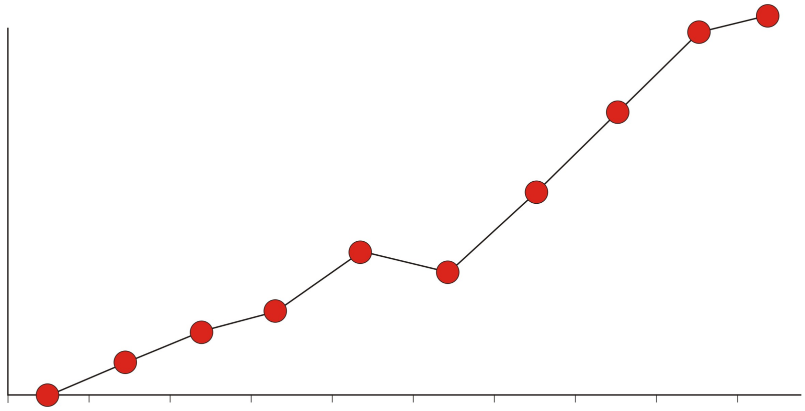
2. Sean was given the following relative frequency histogram to read.



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

(a) 16
(b) 24
(c) 32
(d) 68

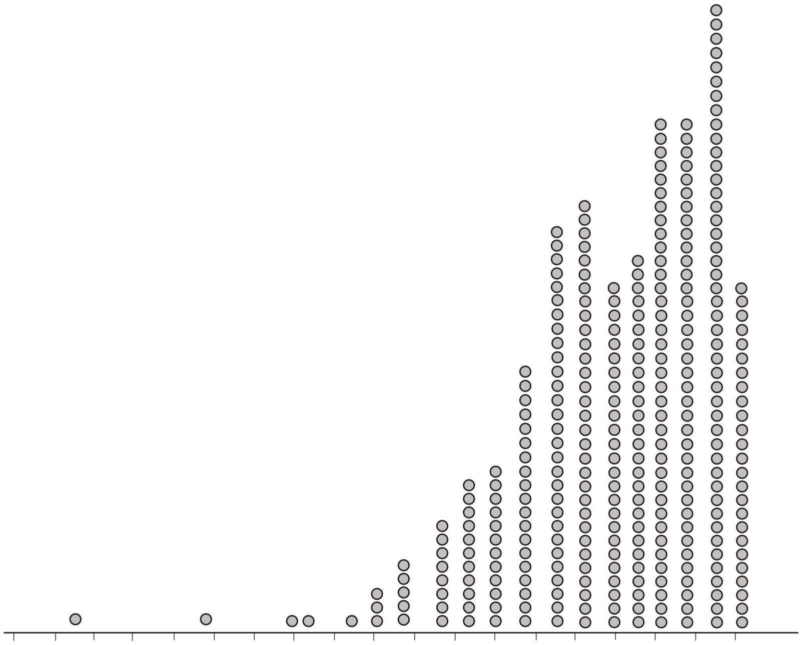
3. Tianna was given a graph for a homework question in her statistics class, but she forgot to label the graph or the axes and couldn't remember if it was a frequency polygon, or an ogive plot. Here is her graph:



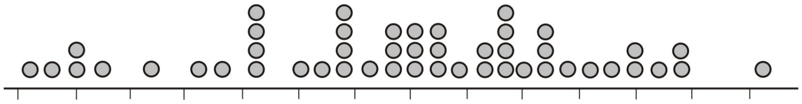
Identify which of the two graphs she has and briefly explain why.

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

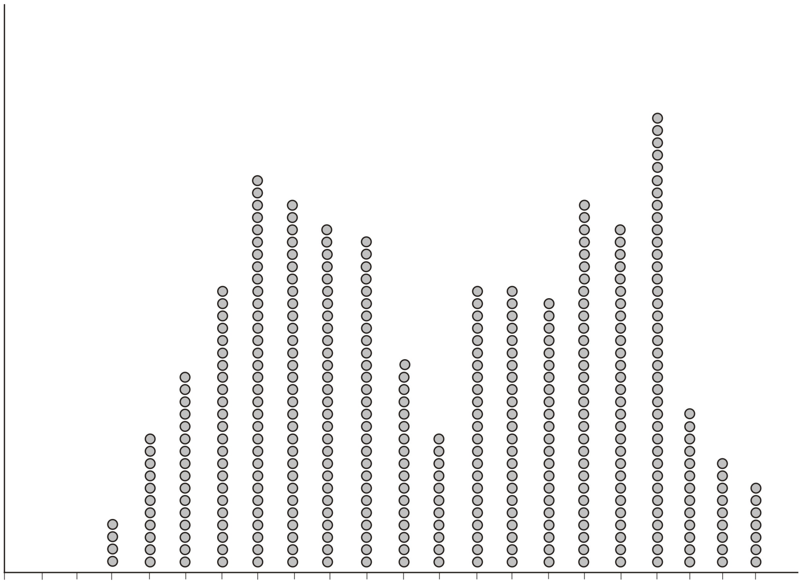
4.



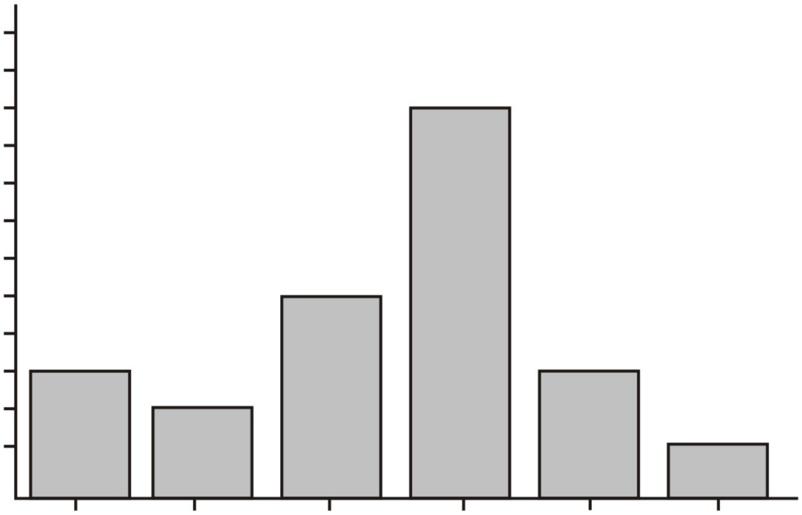
5.



6.



7.



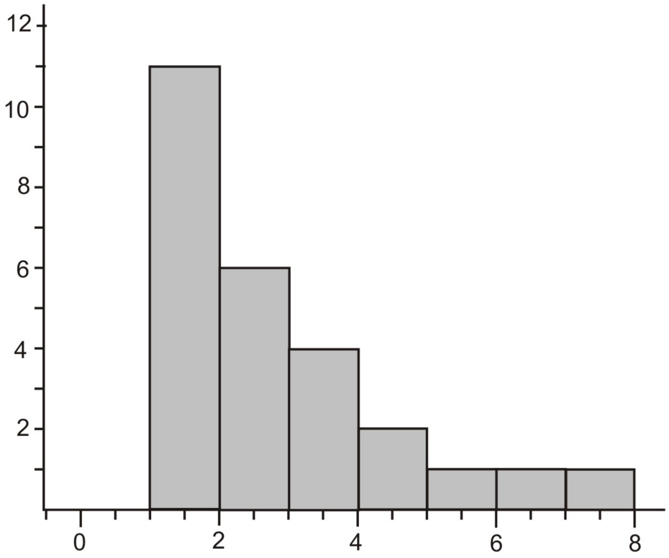
(a) Endy collected and graphed the heights of all the 12^{th} grade students in his high school.

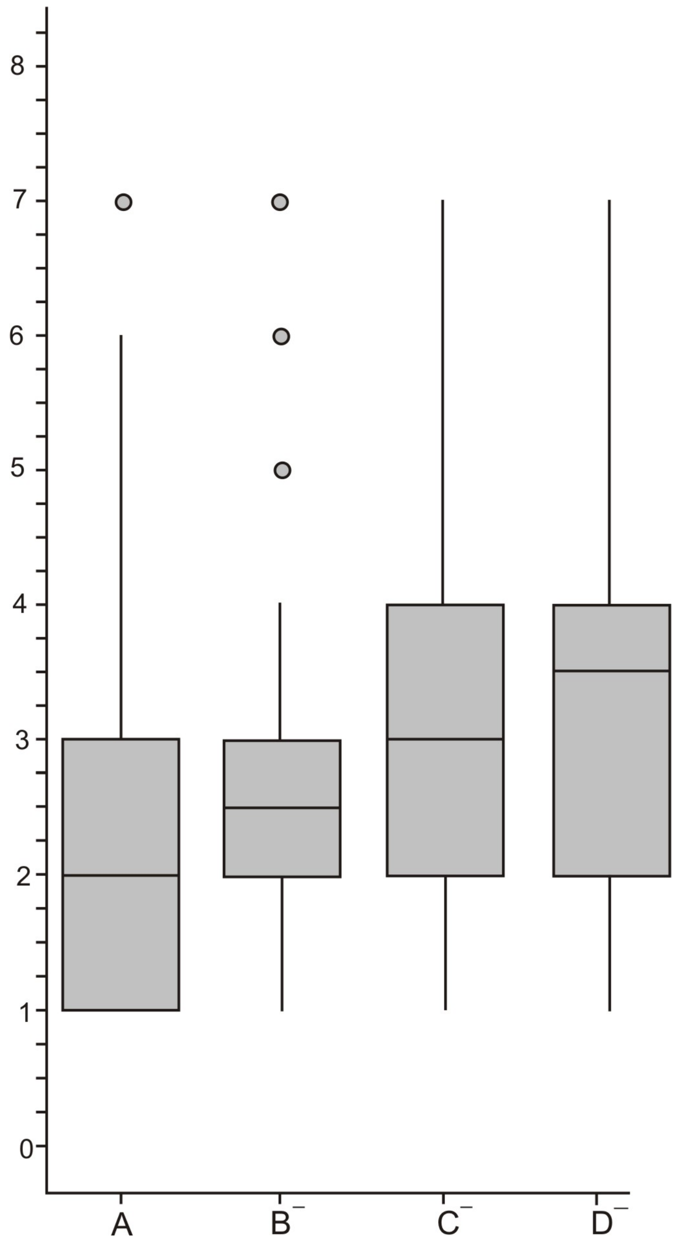
(b) Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or bank change. She created a plot of the dates of the pennies.

(c) Thamar asked her friends what their favorite movie was this year and graphed the results.

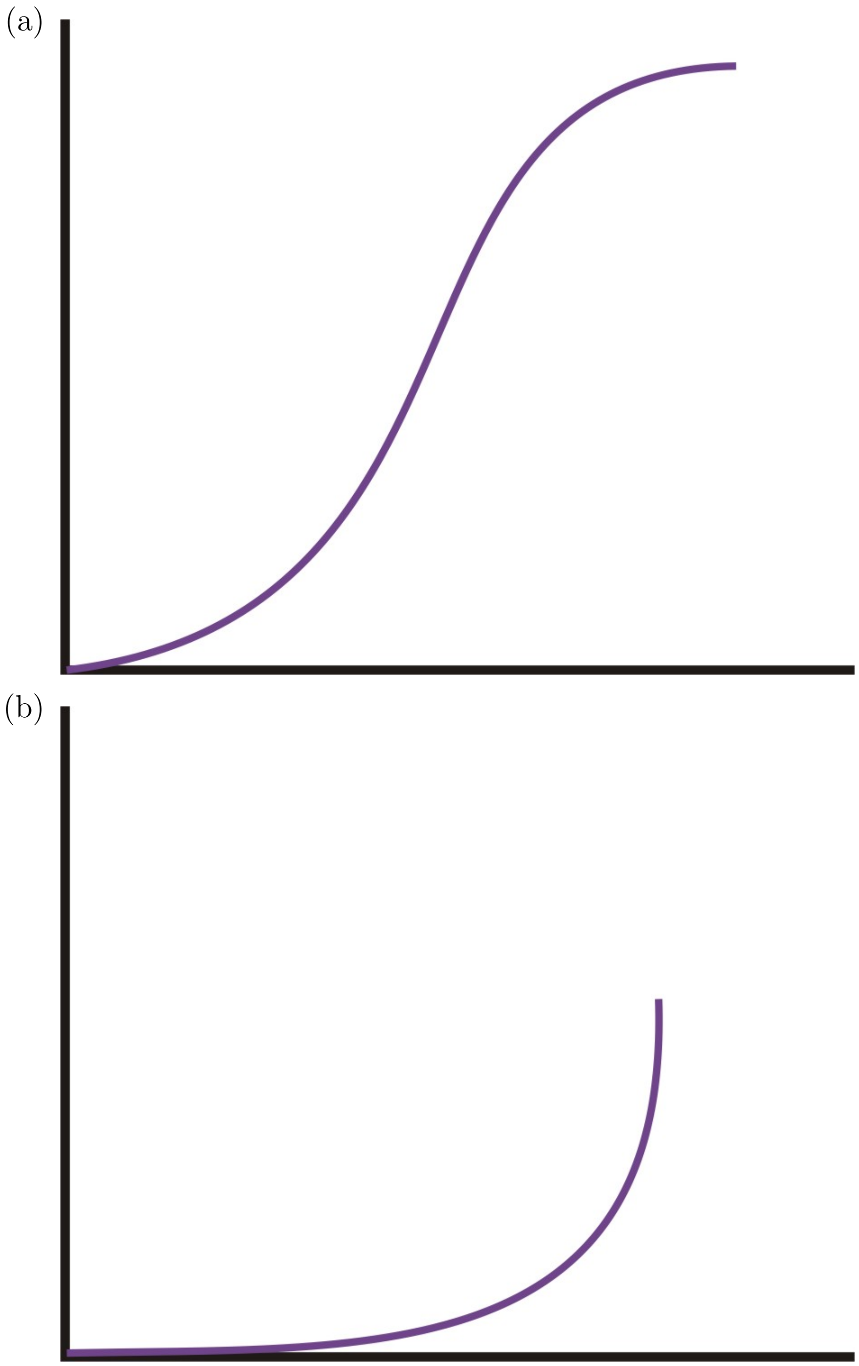
(d) Jeno bought a large box of doughnut holes at the local pastry shop, weighed each of them and then plotted their weights to the nearest tenth of a gram.

8. Which of the following box plots matches the histogram?

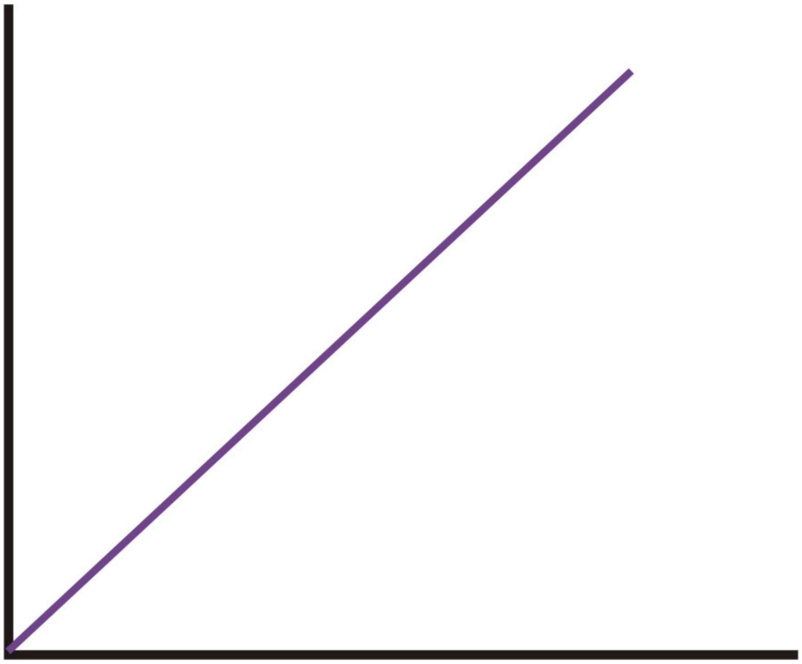

12
10
8
6
4
2
0
2
4
6
8



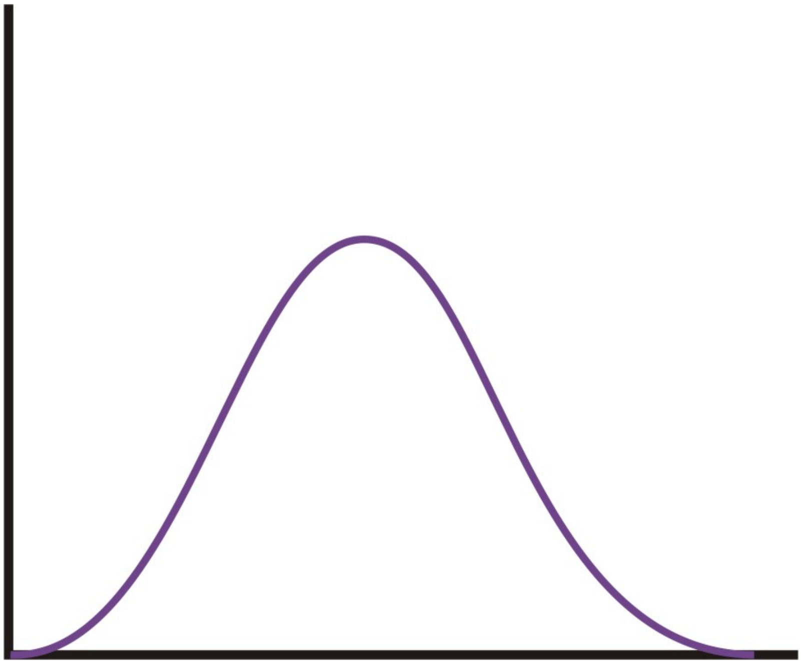
9. If a data set is roughly symmetric with no skewing or outliers, which of the following would be an appropriate sketch of the shape of the corresponding ogive plot?


(a)
(b)

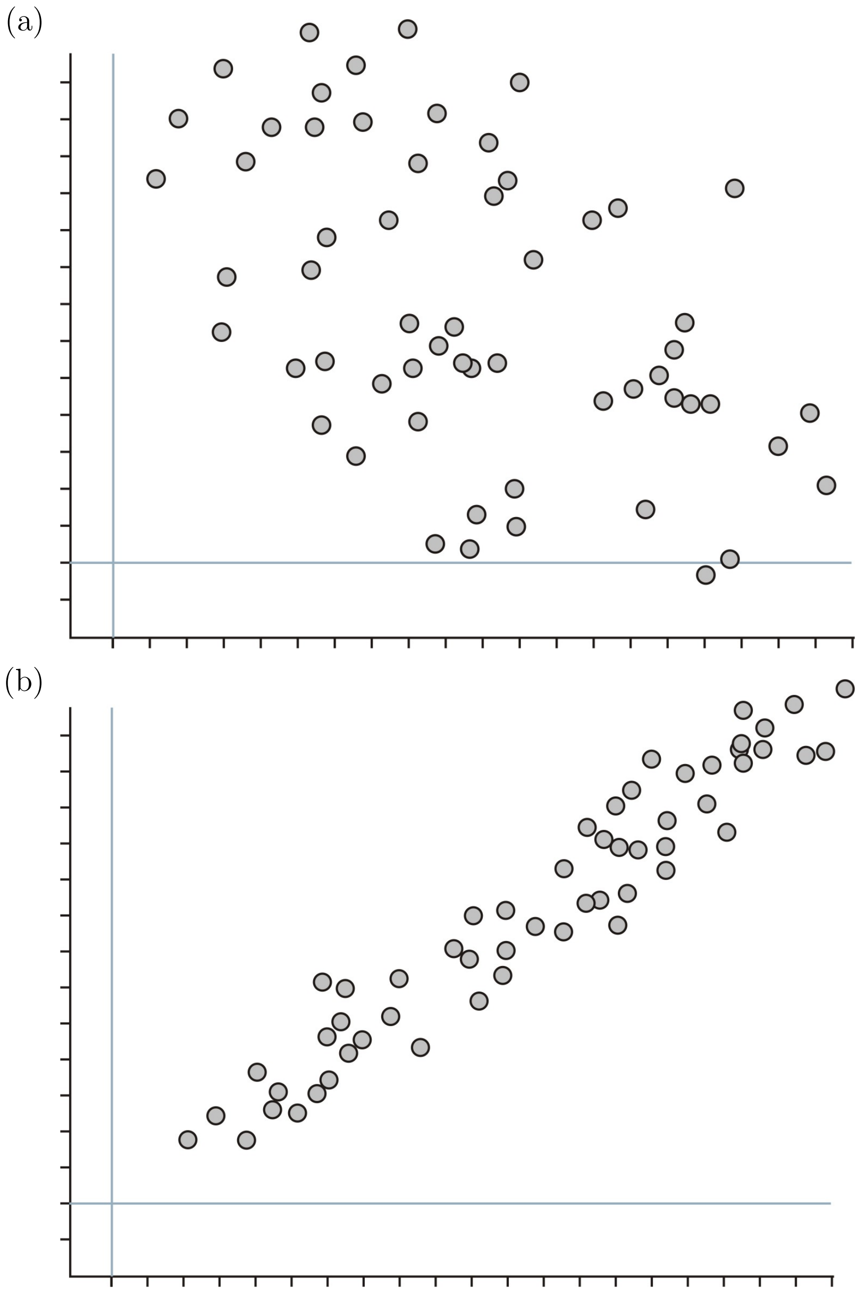
(c)



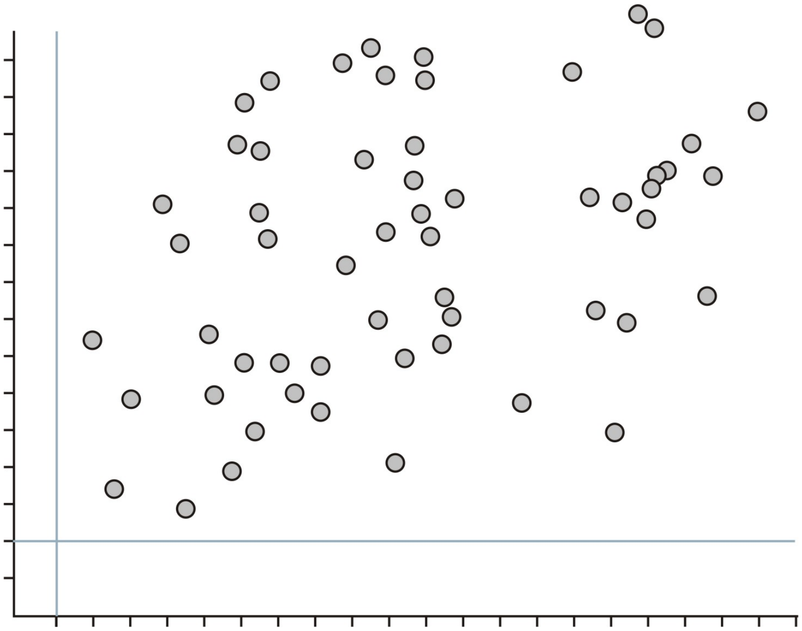
(d)



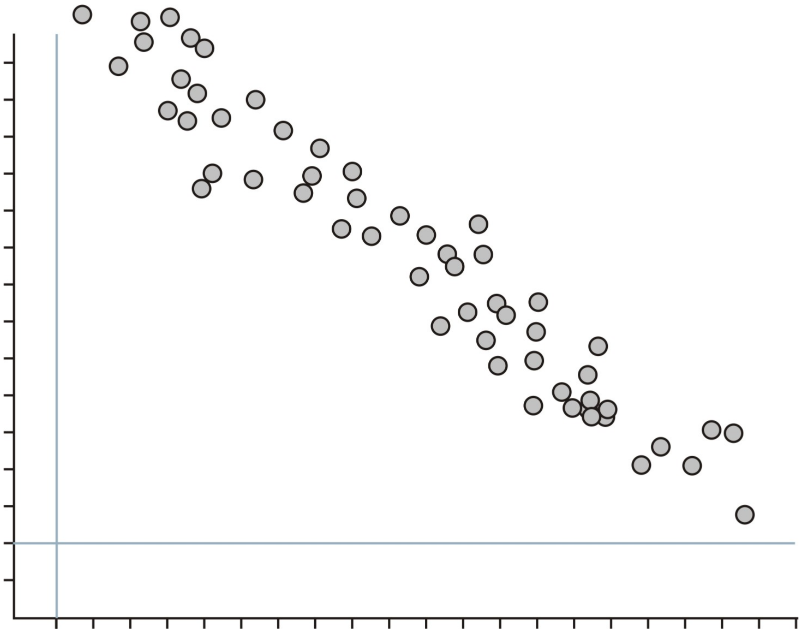
10. Which of the following scatterplots shows a strong, negative association?


(a)
(b)

(c)



(d)



Part One: Answers

1. b
2. c
3. It must be a frequency polygon. At one point in the graph, there is a decreasing line. An ogive plot represents the cumulative data up to that point, so it can never decrease.
4. b

5. d
6. a
7. c
8. a
9. a
10. d

Part Two: Open-Ended Questions

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

Table 2.32:

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**
Emirates Office Tower	**Dubai**	**1163**
Tuntex Sky Tower	**Kaohsiung**	**1140**

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

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

Table 2.33:

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 90^{th} percentile for this data (i.e. the height that 90% of the data is less than)

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

Table 2.34:

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

Table 2.34: (continued)

Year*	Adults	Jacks
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/newsreleases/Sacto_adult_and_jack_escapement_thru%202007.pdf

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

In case you are not up on your salmon facts there are two terms in this chart that may be unfamiliar. Fish escapement refers to the number of fish who "escape" the hazards of the open ocean and return to their freshwater streams and rivers to spawn. A "Jack" salmon is a fish that returns to spawn before reaching full adulthood.

(a) Create one line graph that shows both the adult and jack populations for those years. The data from 1971 to 1995 represents 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 2.35:

Island	Approximate Area (sq.km)
Baltra	8
Darwin	1.1
Española	60
Fernandina	642
Floreana	173
Genovesa	14
Isabela	4640
Marchena	130

Table 2.35: (continued)

Island	Approximate Area (sq.km)
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.

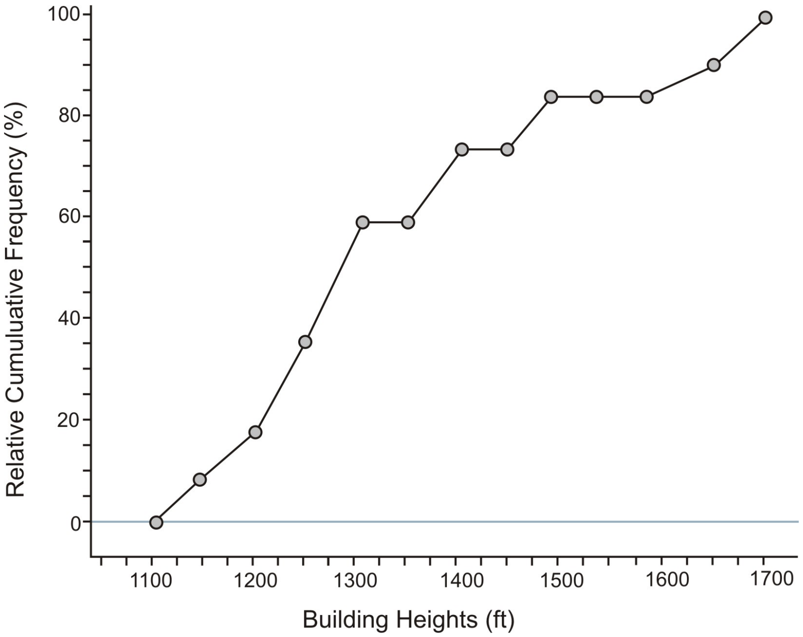
4. Investigation: The National Weather Service maintains a vast array of data on a variety of topics. Go to: http://lwf.ncdc.noaa.gov/oa/climate/online/ccd/snowfall.html. You will find records for the mean snowfall for various cities across the US.
   (a) Create a back-to-back stem-and-leaf plot for all the cities located in each of two geographic regions. (Use the simplistic breakdown found at the following page 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.

Part Two: Open-Ended Answers

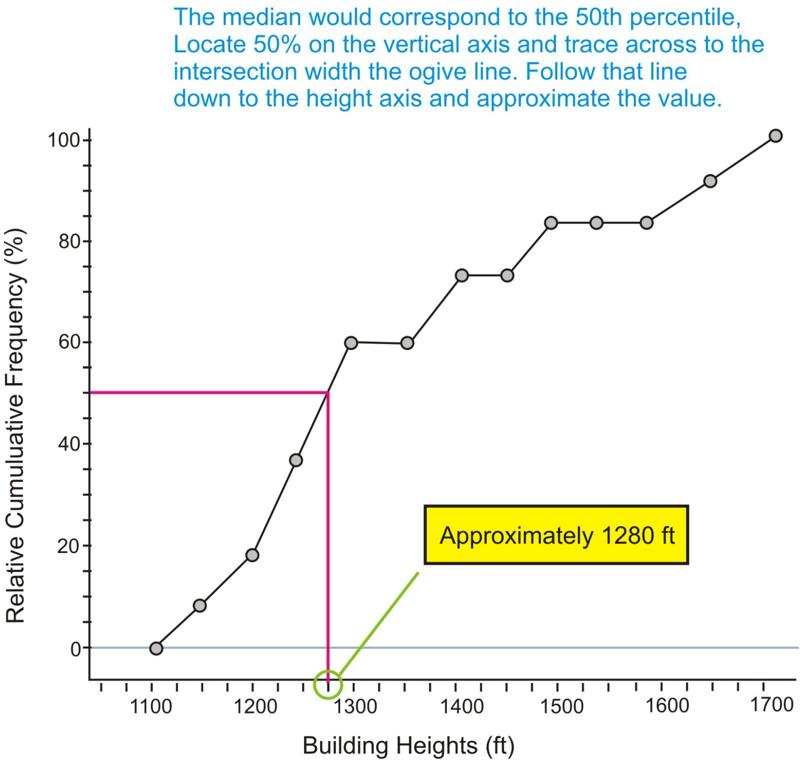
1. (a)

Table 2.36:

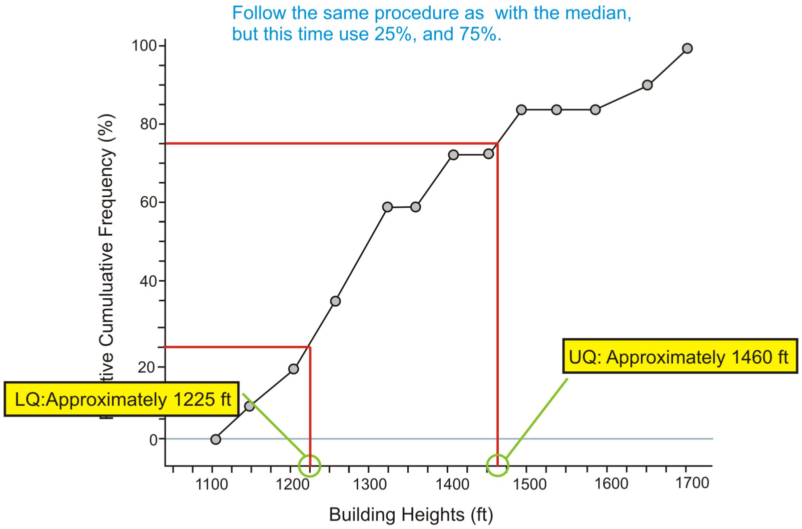
Class	Frequency	Relative Frequency(%)	Cumulative Frequency	Relative Cumulative Frequency(%)
[1100-1150)	1	7.1	1	0
[1150-1200)	1	7.1	2	7.1
[1200-1250)	3	21.4	5	14.3
[1250-1300)	3	21.4	8	35.7
[1300-1350)	0	0	8	57.1
[1350-1400)	2	14.3	10	57.1
[1400-1450)	0	0	10	71.4
[1450-1500)	2	14.3	12	85.7
[1500-1550)	0	0	12	85.7
[1550-1600)	0	0	12	85.7
[1600-1650)	1	7.1	13	92.9
[1650-1700)	1	7.1	14	100



(b)



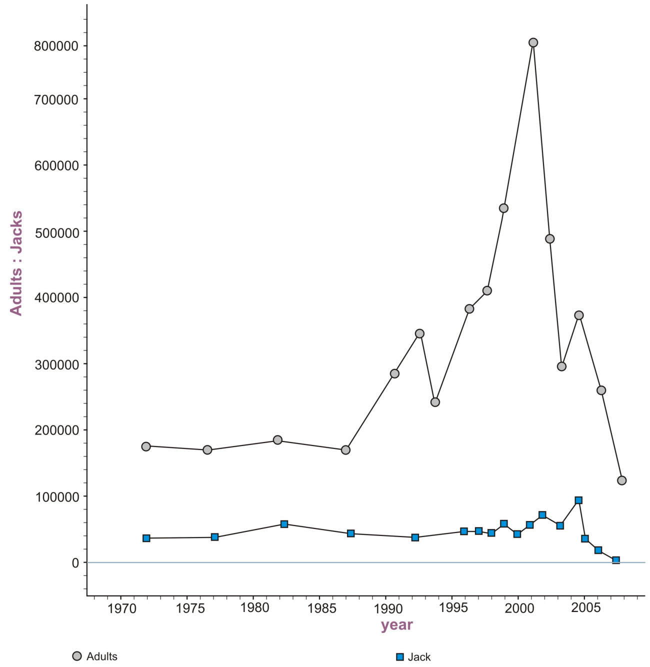
(c)



(d) approximately 1625 ft

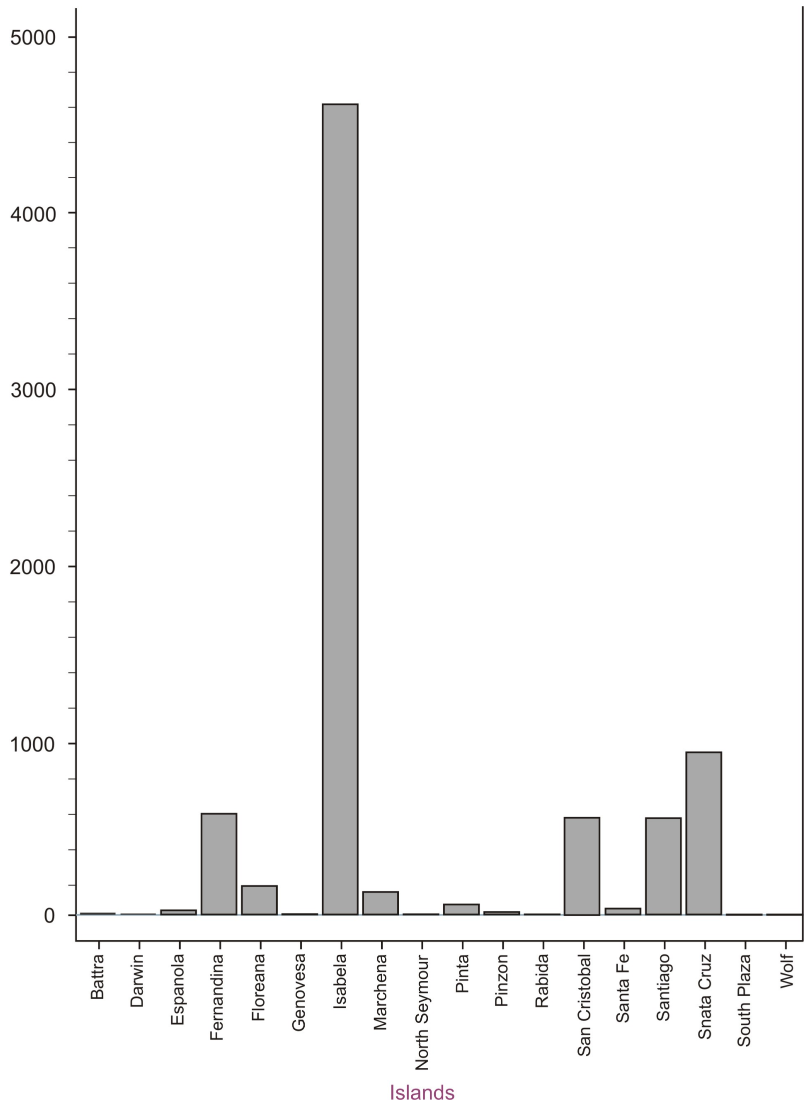
2. (a) There isn't necessarily a *wrong* way or right way to create this graph and to interpret the different time intervals, but a year should be the same distance apart for the entire graph so that the rate of change of the lines means the same thing across the entire plot. In this case, we plotted the average as a point in the middle of the five-year interval. It is possible that a student could devise a better representation, as long as the relationship in the data is clearly and correctly

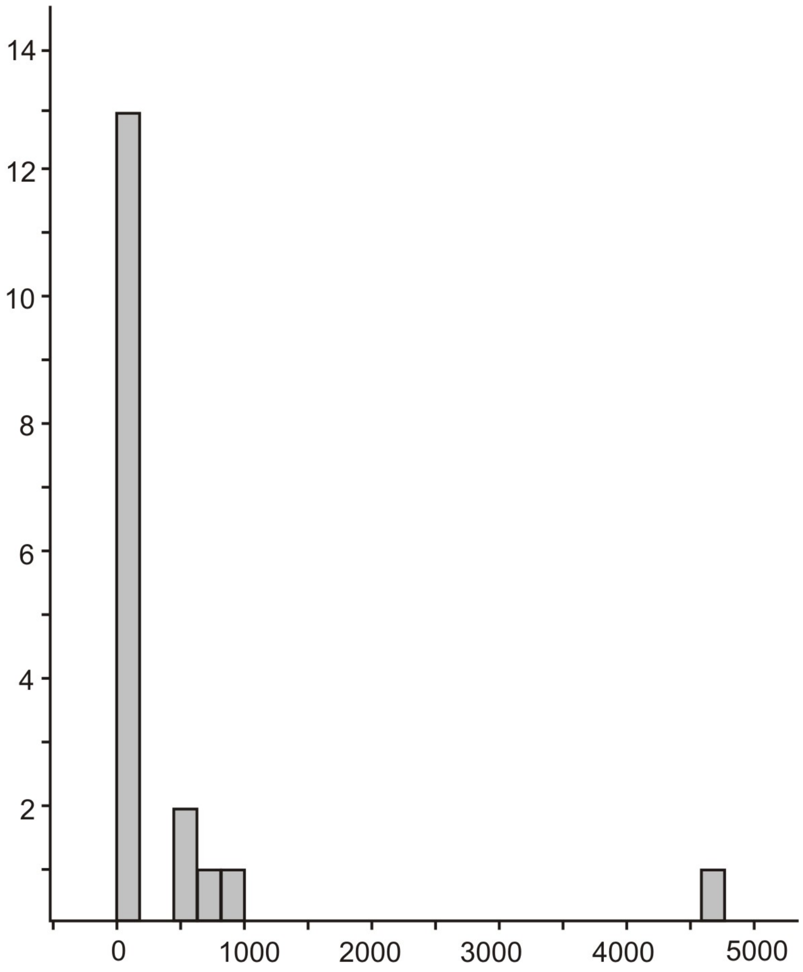
represented.

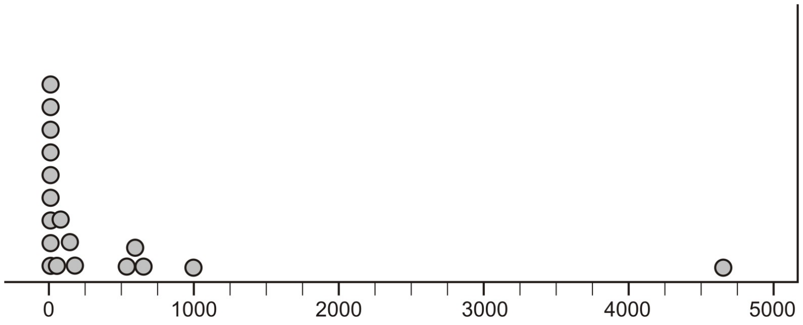


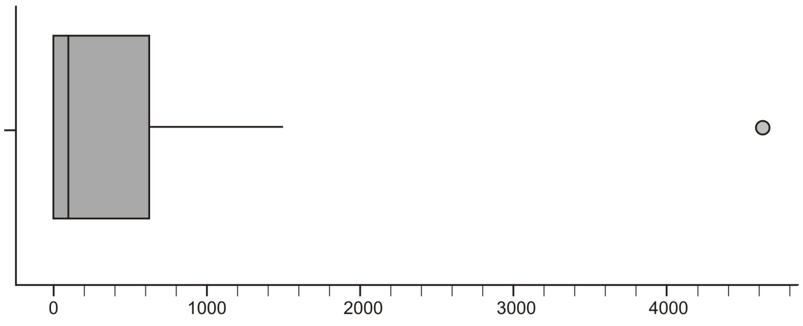
(b) Answers will vary, but comments should focus on features of the plot that are placed *in the context* of the actual situation. For example, the plot of adult salmon increases dramatically after 1995 to a peak in 2002. This could be due to many factors, one of which was the inclusion of the Chinook salmon under the endangered species act. The plot for the Jack salmon stays relatively horizontal, indicating that the Jack population remained relative constant until the most recent downturn. Other comments could be made and interested students might be encouraged to research things such as climate conditions or changes in the management of the salmon populations that may have led to the increases or decreases.

3. (a) The various plots are shown below:


5000
4000
3000
2000
1000
0
Battra
Darwin
Espanola
Fernandina
Floreana
Genovesa
Isabela
Marchena
North Seymour
Pinta
Pinzon
Rabida
San Cristobal
Santa Fe
Santiago
Snata Cruz
South Plaza
Wolf
Islands

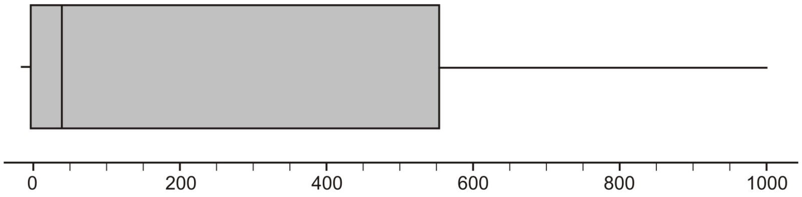

14
12
10
8
6
4
2
0
1000
2000
3000
4000
5000


0
1000
2000
3000
4000
5000

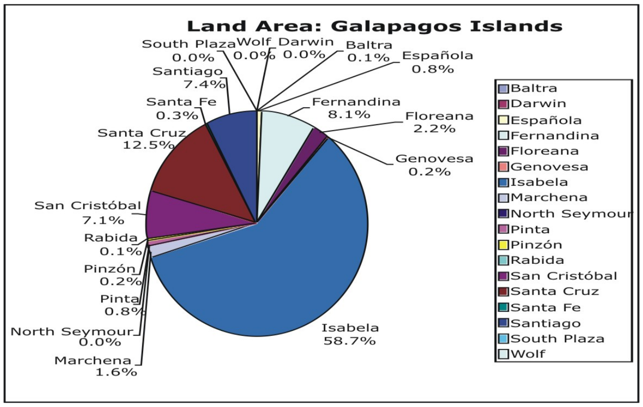

0
1000
2000
3000
4000

The only plot that does not seem to be a good fit is a stem-and-leaf plot. There is an extremely wide spread with the outlier, and creating meaningful stems would be difficult.

(b) The plot is spread very widely, extending from a group of islands with almost no significant area, to the largest island, Isabela, which is so large at 4600 mi2 that it is an extreme outlier. Even without the outlier, there is still a significant variation in the remaining islands. Ignoring Isabela, the distribution is still significantly skewed right. You can see this in all three graphs and it shows that most of the islands in the archipelago are smaller. The box plot does not appear to have a left whisker, but it is in fact, so small in relation to the scale of the graph, that it is indistinguishable. Here is a box-and-whisker plot without the outlier that has been rescaled.



The center would most appropriately be measured by the median because the extreme skewing and outliers will raise the mean substantially. The median island size is approximately 42 square kilometers.



4. .

Image Sources

(1) .

Chapter 3

An Introduction to 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 past records indicate that in the past five years, the company's profit has 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 decisions concerning a successful investment in the software company and the decision of not betting $100 for the next outcome of a die are both based on probabilities of certain sample results. Namely, the software company's profit has been declining for the past five years and the outcome of rolling a 2 ten times in a row is quite strange. From these sample results, we might conclude that we are not going to invest our money in the software company or continue betting on this die. In this chapter you will learn mathematical ideas and tools that can help you understand such situations.

3.1 Events, Sample Spaces, and Probability

Learning Objectives

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

An **event** is something that occurs or happens. Flipping a coin is an event. Walking in the park and passing by a bench is an event. Anything that could possibly happen is an event.

Every event has one or more possible **outcomes**. Tossing a coin is an event but getting a tail is the outcome of the event. Walking in the park is an event and finding your friend sitting on a bench is an outcome of the event.

In statistics, the process of taking a measurement or making an observation is called an **experiment.** For example, tossing a coin and recording the up face in a table of data is an experiment because a measurement is taken.

Experiment

The process of taking a measurement or making an observation.

Keep in mind that the definition of an experiment in statistics is broader than the one used in science. A scientific experiment involves scientific instrumentations such as thermometers, microscopes, telescopes and tubes. A statistical experiment may involve all these items but it mainly involves recording data and measurements. For example, we may conduct an experiment to learn which brand of coffee a customer may prefer among three brands, recording a voter's opinion on a particular political issue, or measuring the amount of carbon monoxide present in a certain environment. Any kind of observation, measuring, and recording that you may conduct can be considered a statistical experiment.

Suppose a coin is tossed once. There are two possible outcomes, either a head (H) or a tail (T). Notice that if the experiment is conducted only once, you will observe only one of the two possible outcomes. These individual outcomes for an experiment are each called **simple events**. Here is another example: a die has six possible outcomes: $1, 2, 3, 4, 5, 6$. 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.

Simple Event

The simplest outcome of an experiment.

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. In other words, the simple events 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 an experiment.

Sample Space

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:

Experiment: We want to investigate 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:

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

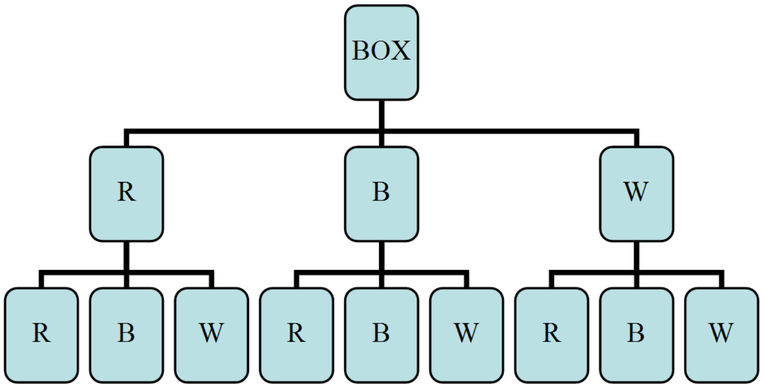
Similarly, the sample space of tossing a coin is either head (H) or tail (T) so we write $S = \{H, T\}$.

Example:

Experiment: 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?

Solution:

It is probably best if we draw a diagram to illustrate all the possible drawings.



As you can see from the 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 diagram above, we can see that the sample space is:

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

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

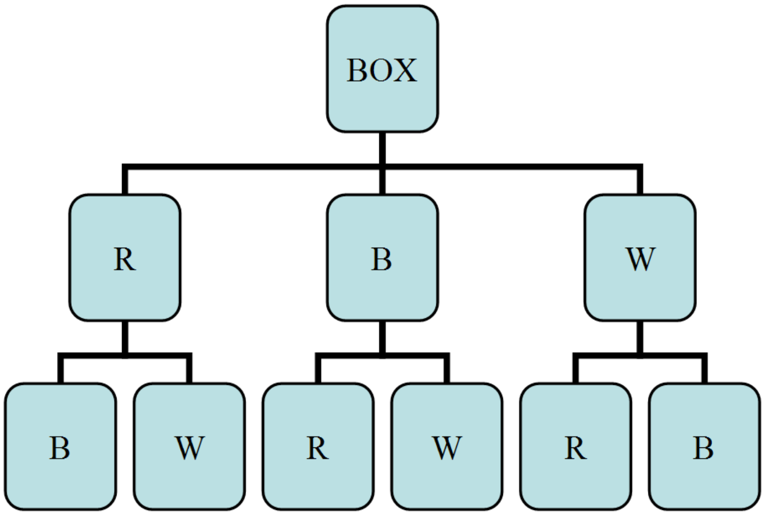
Where the first column represents the first drawing and the first row represents the second drawing.

Example:

Experiment: Consider the same experiment as in the example before last but 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 diagram below illustrates this case:



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

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

Now let us return to the concept of probability and relate it to the concepts that we have just studied. You may be familiar with the meaning of probability and may have used the term as a synonym with informal words like "chance" and "odds." For the time being, we will begin our treatment of probability using these informal concepts and then later, we will solidify these meanings into formal mathematical definitions.

As you probably know from your previous math courses, if you toss a fair coin, the chance of getting a tail (T) is the same as the chance of getting a head (H). Thus we say that the probability of observing a head is 50% (or 0.5) and the probability of observing a tail is also 50%. We also say sometimes "the odds are $50 - 50$."

The probability, P, of an outcome, A, always falls somewhere between two extremes: 0 (or 0%), which means the outcome is an impossible event and 1 (or 100%) represents an outcome that is guaranteed to happen. These two extremes are generally not seen in real life situations. 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 between 0 (impossible) and 1 (always).

In addition, the probabilities of possible outcomes of an event must all add up to 1. This 1 represents a certainty that one of the outcomes must happen. For example, tossing a coin will produce either a head or a tail. Each of these two outcomes has a probability of 50%, or 1/2. However, the total probabilities of the coin to land head or tail is $1/2 + 1/2 = 1$.

Property 2

$$\sum_{\text{all outcomes}} P(A) = 1$$

The sum of the probabilities of all possible outcomes must add up to 1.

Notice that tossing a coin or throwing a dice results in outcomes that are all equally probable, that is, each outcome has the same probability as the other outcome in the same sample space. Getting a head or a tail from tossing a coin produces equal probability for each outcome, 50%. Throwing a die also has 6 possible outcomes but they all have the same probability, 1/6. We refer to this kind of probability as the *classical probability*. It is the simplest kind of probability. Later in this lesson, we will deal with situations where each outcome in a given sample space has different probability.

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 that the outcome A happens is $P(A)$. We use the following formula to calculate the probability of an outcome to occur:

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

The following examples show you how to use this formula.

Example:

When tossing two coins, what is the probability of getting head-head (HH)? Is the probability classical?

Solution:

Since there are 4 elements (outcomes) in the set of sample space: $\{HH, HT, TH, TT\}$, its size then is 4. Further, there is only 1 HH outcome to occur. Using the formula above,

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

Notice that each of these 4 outcomes is equally probable, namely, the probability of each is 1/4. Thus it is a classical probability. Notice also that the total probabilities of all possible outcomes add up to one: $1/4 + 1/4 + 1/4 + 1/4 = 1$

Example:

What is the probability of throwing a dice and getting either $2, 3$, or 4?

Solution:

The sample space for a fair dice has a total of 6 possible outcomes. However, the total number of outcomes for our case is 3 hence,

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

So, there is a probability of 50% that we will get $2, 3$, or 4.

Example:

Consider an experiment of tossing two coins. Assume the coins are not balanced. The design of the coins is to produce the following probabilities shown in the table:

Table 3.1:

Sample Space	Probability
HH	4/9
HT	2/9
TH	2/9
TT	1/9

Figure: Probability table for flipping two weighted coins.

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

Solution:

Notice that the simple events HT and TH 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

1. An **event** is something that occurs or happens with one or more **outcomes**.
2. An **experiment** is the process of taking a measurement or making an observation.
3. A **simple event** is the simplest outcome of an experiment.
4. The **sample space** is the set of all possible outcomes of an experiment, typically denoted by S.

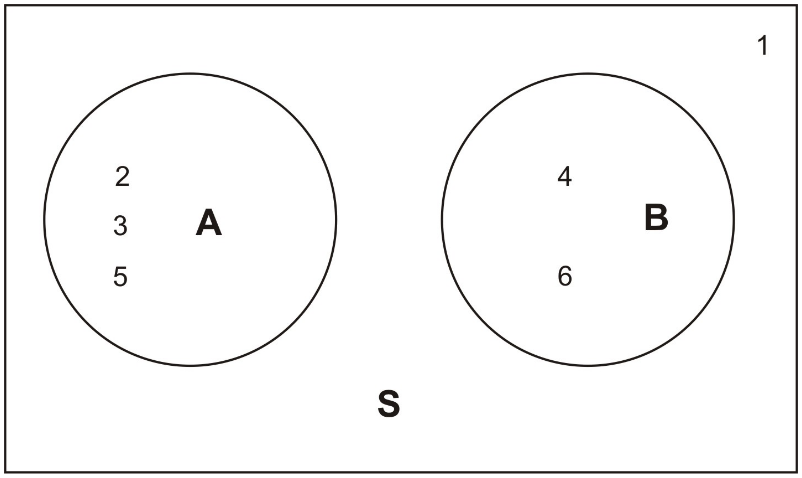
Review Questions

1. Consider an experiment composed of throwing a die followed by throwing a coin.
   (a) List the simple events and assign a probability for each simple event.
   (b) What are the probabilities of observing the following events?

$$\begin{aligned} &\text{A: }\{\text{2 on the die, H on the coin}\} \\ &\text{B: }\{\text{Even number on the die, T on the coin}\} \\ &\text{C: }\{\text{Even number on the die}\} \\ &\text{D: }\{\text{T on the coin}\} \end{aligned}$$

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:

$$\begin{aligned} P(1) = P(2) = P(4) = 2/9 \\ P(3) = P(5) = P(6) = 1/9 \end{aligned}$$



(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.
   (a) Refer to the blue marbles as $B1$ and $B2$ and the red ones as $R1$, $R2$, and $R3$. List the outcomes in the sample space.
   (b) Determine the probability of observing each of the following events:

A: {2 blue marbles are drawn}
B: {1 red and 1 blue are drawn}
C: {2 red marbles are drawn}

Review Answers

1. (a) $\{1T, 1H, 2T, 2H, 3T, 3H, 4T, 4H, 5T, 5H, 6T, 6H\}$
   (b) A: 1/12 B: 1/4 C: 1/2 D: 1/2
2. (a) 4/9
   (b) 1/3
3. (a) $\{B1B2, B1R1, B1R2, B1R3,$
   $B2B1, B2R1, B2R2, B2R3,$
   $R1B1, R1B2, R1R2, R1R3,$
   $R2B1, R2B2, R2R1, R2R3,$
   $R3B1, R3B2, R3R1, R3R2\}$
   (b) A: 1/10 B: 3/5 C: 3/10

3.2 Compound Events

Learning Objectives

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

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.

Definition The **union** of two events A and B occurs if either event A or event B or both occur on a single performance of an experiment. We denote the union of the two events by the symbol $A \cup B$. You can say this symbol with either "A union B" or "A or B".

The word "**or**" is used with **union**:
$A \cup B$ means everything that is in set A **OR** in set B **OR** in both sets.

Definition The **intersection** of two events A and B occurs if both event A and event B occur on a single performance of an experiment. We denote the intersection of two events by the symbol $A \cap B$. The most common way to say this symbol is "A and B".

The word "**and**" is used with **intersection**:
$A \cap B$ means everything that is in set A **AND** in set B.

Example:

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

$$\begin{aligned}&\text{A: \{observe an even number\}}\\&\text{B: \{observe a number less than or equal to 3\}}\end{aligned}$$

1. Describe $A \cup B$ for this experiment.
2. Describe $B \cap B$ for this experiment.
3. Calculate $P(A \cup B)$ and $P(A \cap B)$, assuming the die is fair.

Solution:

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

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

$$\begin{aligned} A \cup B &= \{2,4,6\} \cup \{1,2,3\} \\ &= \{1,2,3,4,6\} \end{aligned}$$

2. The intersection of A and B is the event that occurs if we observe *both* an even number and a number that is equal to or less than 3 on a single toss of a die.

$$\begin{aligned} A \cap B &= \{2,4,6\} \cap \{1,2,3\} \\ &= \{2\} \end{aligned}$$

In other words, the intersection of A and B is the simple event to observe a 2.

3. Remember the probability of an event is the sum of the probabilities of the simple events,

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

Similarly,

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

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

Example:

Refer to the above example and define the new events

C: {observe a number that is greater than 5}
D: {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 $A \cap B \cap C$

Solution:

1. Event C corresponds to finding the simple event $S(C) = C = \{6\}$. So

$$\begin{aligned} A \cup B \cup C &= \{2, 4, 6\} \cup \{1, 2, 3\} \cup \{6\} \\ &= \{1, 2, 3, 4, 6, \} \end{aligned}$$

2. Event D corresponds to finding the simple event $S(D) = D = \{5\}$. So

$$\begin{aligned} A \cap D &= \{2, 4, 6\} \cap \{5\} \\ &= \phi \end{aligned}$$

Where ϕ is the empty set. This says that there are no elements in the set $A \cap D$. This means that you will not observe any events that combine sets A and D.

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

$$\begin{aligned} (A \cap B) \cap C &= (\{2, 4, 6\} \cap \{1, 2, 3\}) \cap \{6\} \\ &(\{2\} \cap \{6\}) \\ &= \phi \end{aligned}$$

Again, we get the empty set.

Lesson Summary

1. The **union** of two events A and B, $A \cup B$, occurs if either event A or event B or both occur on a single performance of an experiment. A union is an **"or" relationship**.
2. The **intersection** of two events A and B, $A \cap B$, occurs only if both event A and event B occur on a single performance of an experiment. An intersection is an **"and" relationship**.
3. 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.
- Using the complement of an event to calculate the probability of an event.
- Understanding the complementary rule.

Definition The **complement** A' of an event A consists of all the simple events (outcomes) that are *not* in the event A.

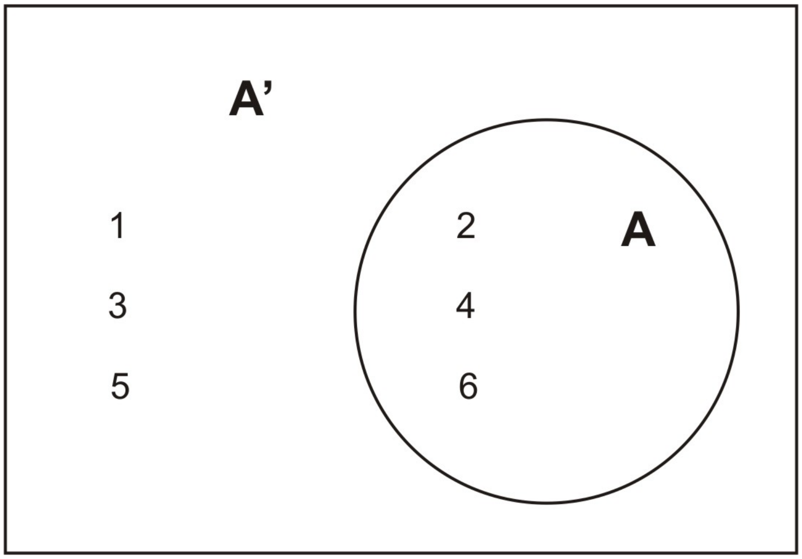
Let us refer back to the experiment of throwing one die. As you know, the sample space of a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. If we define the event A as

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

Then, $A = \{1, 3, 5\}$, which includes all the simple events of the set S that are odd. Thus, the **complement** of A is the set of simple events that will not occur in A. So A' will include all the elements that are not odd in the sample space of the set S:

$$A' = \{2, 4, 6\} \, .$$

The Venn diagram is shown below.



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

The Complementary Rule

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 below, it is sometimes easier to calculate the probability of the complement of an event rather than the event itself. Then the probability of the event, $P(A)$, is calculated using the relationship:

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

Example:

If 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?

Solution:

First, ask the question, what is the probability of the simple event? It is

$$P(A) = \{\text{you will get the flu}\} = 0.43$$

The complement is

$$P(A') = \{\text{you will not get the flu}\} = 1 - P(A) = 1 - 0.43 = 0.57$$

Example:

Two coins are tossed simultaneously. Here is an event:

$$\text{A: }\{\text{observing at least one head }\}$$

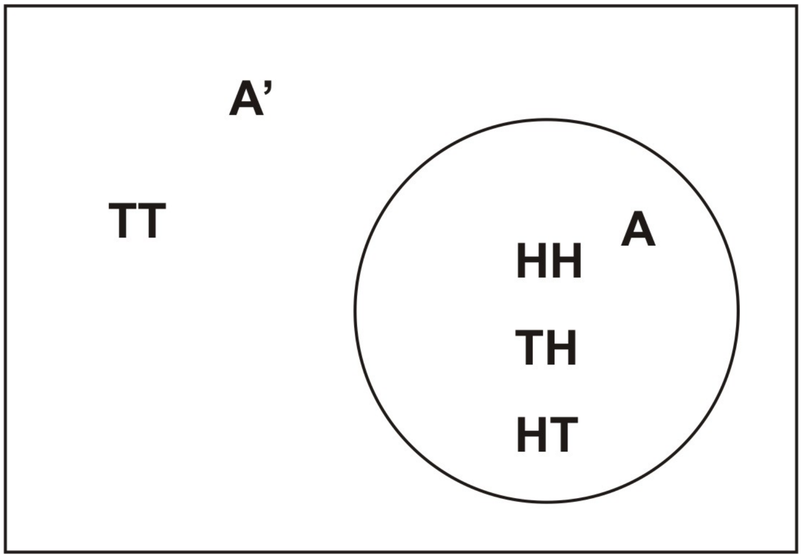
What is the complement of A and how would you calculate the probability of A by using the complementary relationship?

Solution:

Since the event A is observing all simple events $A = \{HH, HT, TH\}$, then the complement of A is defined as the event that occurs when A does not occur, namely, all the events that do not have heads, namely,

$$A' = \{\text{observe no heads}\} = \{TT\}$$

We can draw a simple Venn diagram that shows A and A' in the toss of two coins.



The second part of the problem is to calculate the probability of A using the complementary relationship. Recall that $P(A) = 1 - P(A')$. So by calculating $P(A')$, we can easily calculate $P(A)$ by subtracting it from 1.

$$P(A') = P(TT) = 1/4$$

and

$$P(A) = 1 - P(A') = 1 - 1/4 = 3/4.$$

Obviously, we could have gotten the same result if we had calculated the probability of the event of A occurring directly. The next example, however, will show you that sometimes it is easier to calculate the complementary relationship to find the answer that we are seeking.

Example:

Here is a new kind of problem. Consider the experiment of tossing a coin ten times. What is the probability that we will observe at least one head?

Solution:

Before we begin, we can write the event as

$$\text{A} = \{\text{observe at least one head in ten tosses}\}$$

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 look like this: $HTTHTHHTTH$, another $THTHHHTHTH$, etc. Is there a way to calculate the number of simple events for this experiment? The answer is yes but we will learn how to do this later in the chapter. For the time being, let us just accept that there are $2^{10} = 1024$ simple events in this experiment.

To calculate the probability, each time we toss the coin, the chance is the same for heads and tails to occur. We can therefore say that each simple event, among 1024 events, is equally likely to occur. So

$$P(\text{any simple event among } 1024) = \frac{1}{1024}$$

We are being asked to calculate the probability that we will observe at least one head. You may find it difficult to calculate since the heads will most likely occur very frequently during 10 consecutive tosses. However, if we calculate the complement of A, i.e., the probability that *no heads* will be observed, our answer may become a little easier. The complement A' is easy, it contains only one simple event:

$$A' = \{TTTTTTTTTT\}$$

Since this is the only event that no heads appear and since all simple events are equally likely, then

$$P(A') = \frac{1}{1024}$$

Now, because $P(A) = 1 - P(A')$, then

$$P(A) = 1 - P(A') = 1 - \frac{1}{1024} \approx 0.999 = 99.9\%$$

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

Lesson Summary

1. The **complement** A' of an event A consists of all the simple events (outcomes) that are *not* in the event A.
2. The **Complementary Rule** states that the sum of the probabilities of an event and its complement must equal 1, or for an event A, $P(A) + P(A') = 1$.

Review Questions

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

   A: {At least one head is observed}

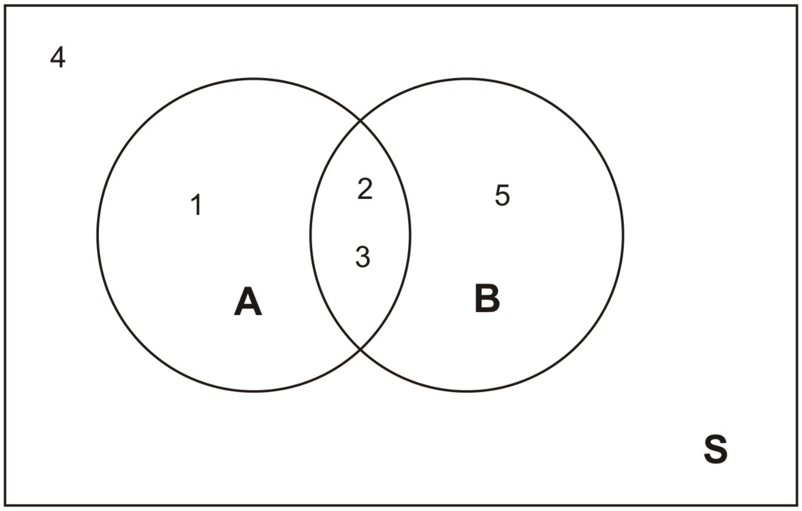
   B: {The number of heads observed is odd}

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

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:

   $$P(1) = 1/10, P(2) = 2/10, P(3) = 3/10, P(4) = 1/10, P(5) = 3/10.$$

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



Review Answers

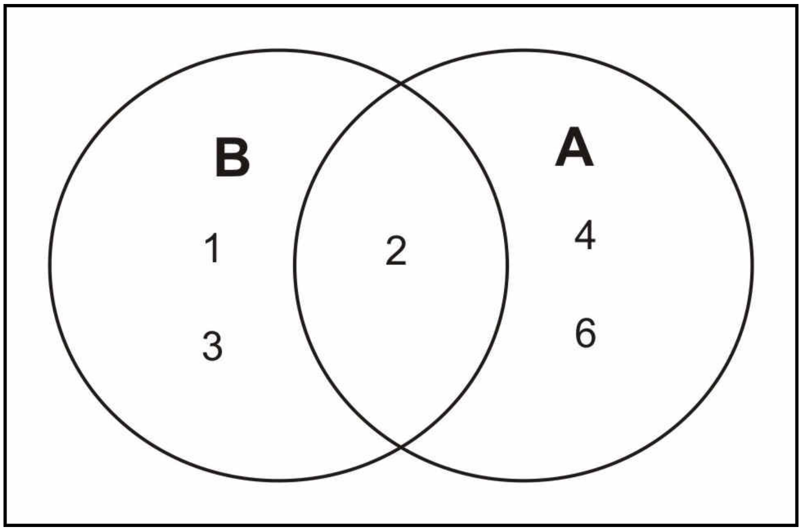
1. (a) all: $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
   (b) A : $\{HHH, HHT, HTH, THH, HTT, THT, TTH\}$
   (c) B : $\{HHH, HTT, THT, TTH\}$
   (d) $A \cup B$ same as A, A' : $\{TTT\}$, $A \cap B$ same as B
   (e) $P(A) = P(A \cup B) = 7/8, P(B) = P(A \cap B) = 1/2, P(A') = 1/8$
2. $4/10, 2/10, 3/10, 5/10, 9/10, 7/10, 5/10, 1/10$.

3.4 Conditional Probability

Learning Objectives

- Calculate the conditional probability that event A occurs, given that event B occurs.

Sometimes, we wish to change the probability of an event when we are bound to certain conditions. For example, we know that the probability of observing an even number on a throw of a die is 0.5 (simple event A). However, suppose that we throw the die and the result is a number that is 3 or less (simple 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 the event B, we have reduced our sample space from 6 simple events to 3 simple events. In other words, with the introduction of a particular condition (the 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 B is the number 2. We conclude that the probability that A occurs, given that B occurs is 1 : 3, or 1/3. We denote it by the symbol

$P(A|B)$, which reads "the probability of A, given B". So for the die toss experiment, we write

$$P(A|B) = \frac{1}{3}.$$

Conditional Probability of Two Events

Definition Conditional Probability If A and B are two events, then the probability of the event A to occur, *given* that event B occurs is called a **conditional probability**. We denote it by the symbol $P(A|B)$, which reads "the probability of A given B."

However, we want to show a systematic way of calculating conditional probabilities. Take the ratio of the probability of the part of A that falls within the reduced sample space B (i.e., the intersection of the two sample spaces A and B) to the total probability of the reduced sample space.

To calculate the conditional probability that event A occurs, given that event B occurs, 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 follows:

$$\begin{aligned}&\text{A} = \{\text{observe an even number}\}\\&\text{B} = \{\text{observe a number less than or equal to 3}\}\end{aligned}$$

We use the formula,

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

and get,

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

Example:

A medical research center is conducting experiments to examine the relationship between cigarette smoking and cancer in a particular city in the US. Let A represent an individual that smokes and let C represent an individual that develops cancer. So 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, 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).

How can these simple events be studied, along with their associated probabilities, to examine the relationship between smoking and cancer?

Solution:

We have

A: {individual smokes}
C: {individual develops cancer}
A': {individual does not smoke}
C': {individual does not develop cancer}

A very powerful way of determining 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 $P(C|A')$:

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

Before we enter our data into the formula, we need to calculate the value of the denominator. $P(A)$ is the probability of the individuals who smoke 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. Thus

$$\begin{aligned} P(A) &= P(AC) + P(AC') \\ &= 0.10 + 0.30 \\ &= 0.40 \\ &= 40\% \end{aligned}$$

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%. Continuing on with our calculations,

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

Similarly, we calculate the conditional probability of a nonsmoker that develops cancer:

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

Where $P(A') = P(A'C) + P(A'C') = 0.05 + 0.55 = 0.6 = 60\%$. It is also equivalent to using the complementary relation $P(A') = 1 - P(A) = 1 - 0.40 = 0.60$.

So what is our conclusion from these calculations? We can clearly see that there exists a relationship between smoking and cancer: The probability that a smoker develops cancer is 25% and the probability that a nonsmoker develops cancer is only 8%. Taking the ratio between the two probabilities, $25\% \div 8\% = 3.125$, which means a smoker is more than three times more likely to develop cancer than a nonsmoker. Keep in mind, however, that it would not be accurate to say that smoking causes cancer but it does suggest a strong link between smoking and cancer.

Lesson Summary

1. If A and B are two events, then the probability of the event A to occur, given that event B occurs is called a **conditional probability**. We denote it by the symbol $P(A|B)$, which reads "the probability of A given B."
2. Conditional probability can be found with the equation $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

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.

i. List the possible outcomes in the sample space.

ii. Two events are defined as follows:

$$\begin{aligned} &\text{A: \{At least one head appears\}} \\ &\text{B: \{Only one head appears\}} \end{aligned}$$

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, two red, and two blue. Two marbles are randomly selected without replacement and their colors are recorded.

i. List the possible outcomes in the sample space.

ii. Let the following events be defined:

$$\begin{aligned} &\text{A: \{Both marbles have the same color\}} \\ &\text{B: \{Both marbles are red\}} \\ &\text{C: \{At least one marble is red or white\}} \end{aligned}$$

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

Review Answers

1. $0.21, 0.5$
2. $3/4, 1/2, 1/2, 1, 2/3$
3. $1/3, 0, 1/14, 1/7, 1$

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 how to represent them in a Venn diagram.
- Understand the condition of independence.

When the probabilities of certain events are known, we can use those probabilities to calculate the probabilities of their respective unions and intersections. We use two rules: the additive and the multiplicative rules to find those probabilities. The examples that follow will illustrate how we can do so.

Example:

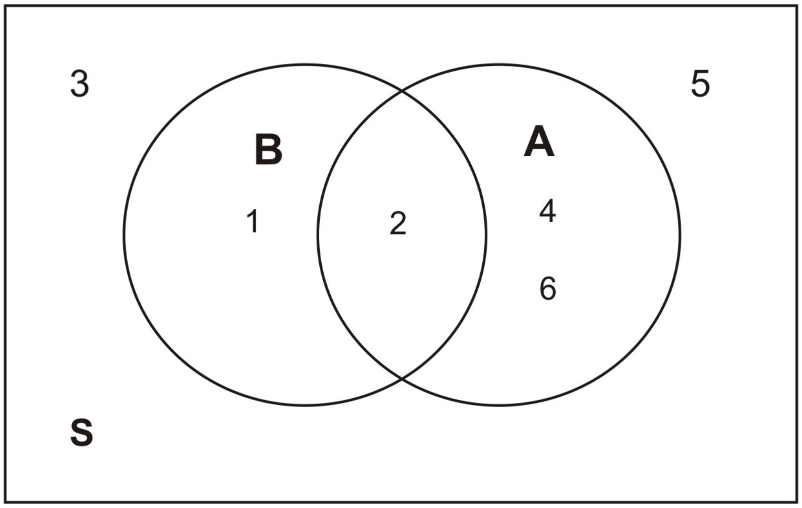
Suppose we have a loaded (unfair) die. We toss it several times and record the outcomes. If we define the following events:

$$\begin{aligned}&\text{A: }\{\text{observe an even number}\}\\&\text{B: }\{\text{observe a number less than 3}\}\end{aligned}$$

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

Solution:

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



Therefore,

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

Since

$$\begin{aligned} P(A) &= P(2) + P(4) + P(6) = 0.4 \\ P(B) &= P(1) + P(2) = 0.3 \\ P(A \cap B) &= P(2) = 0.1 \end{aligned}$$

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

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

But since

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

Substituting, yields

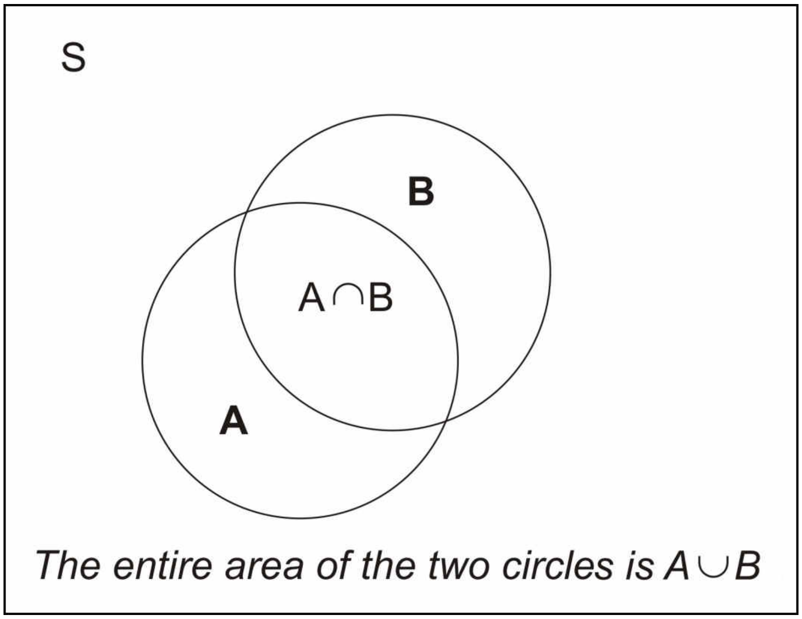
$$P(A) + P(B) = P(A \cup B) + P(2)$$

However, $P(2) = P(A \cap B)$, thus

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

Or,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.4 + 0.3 - 0.1 = 0.6 \end{aligned}$$



What we have demonstrated is that the probability of the union of two events, A and B, can be obtained by adding the individual probabilities of A and B and subtracting $P(A \cap B)$. The Venn diagram above illustrates this union. Formula (1) above is called the **Additive Rule of Probability.**

Additive Rule of Probability

The union of two events, A and B, can be obtained by adding the individual probabilities of A and B and subtracting $P(A \cap B)$. The Venn diagram above illustrates this union.

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

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?

Solution:

Our event is

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

The event E consists of 22 cards; namely, 13 spade cards and 9 face cards that are not spade. Be careful, if we say that we have 12 face cards, we would be over counting the face-spade cards!

To find $P(E)$ we use the additive rules of probability. First, let

$$\begin{aligned} \text{C} &= \{\text{card selected is a spade}\} \\ \text{D} &= \{\text{card selected is a face card}\} \end{aligned}$$

Note that $P(E) = P(C \cup D)$. Remember, event C consists of 13 cards and event D consists of 12 face cards. Event $P(C \cap D)$ consists of the 3 face-spade cards: The king, jack and, queen of spades cards. Using the additive rule of probability formula,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= 0.250 + .231 - .058 \\ &= 0.423 \\ &= 42.3\% \end{aligned}$$

I hope that you have learned, through this example, the reason why we subtract $P(C \cap D)$. It is because we do not want to count the face-spade cards twice.

Example:

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

Solution:

Let

$$\begin{aligned} \text{A} &= \{\text{person selected is male}\} \\ \text{B} &= \{\text{person selected is under 18}\} \end{aligned}$$

From the percents given,

$$P(A) = 0.842 \qquad P(B) = 0.183 \qquad P(A \cap B) = 0.141$$

The probability of a person selected is male or under 18 $P(A \cup B)$:

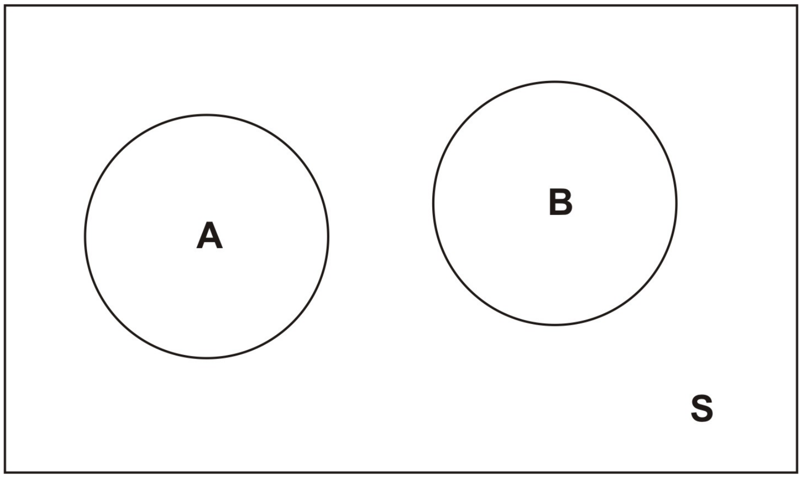
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.842 + 0.183 - 0.141 \\ &= 0.884 \\ &= 88.4\% \end{aligned}$$

This means that 88.4% of the people arrested in the mid 1990's are either males or under 18.

It happens sometimes that $A \cap B$ contains no simple events, i.e., $A \cap B = \{\phi\}$, the empty set. In this case, we say that the events A and B are **mutually exclusive.**

Definition If $A \cap B$ contains no simple events, then A and B are **mutually exclusive**.

The figure below is the 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.



This figure shows that the events A and B have no simple events in common, that is, events A and B can not occur simultaneously, and therefore, $P(A \cap B) = 0$.

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, that is

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

Notice that since the two events are mutually exclusive, there is no over-counting.

Example:

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

Solution:

Let

A: {observe only one head}
B: {observe two heads}

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

Recall from previous section that the conditional probability rule is used to compute the probability of an event, given that another event had already occurred. The formula is

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

Solving for $P(A \cap B)$, we get

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

This result is the **Multiplicative Rule of Probability.**

Multiplicative Rule of Probability

If A and B are two events, then

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

This says that the probability that both A and B occur equals to the probability that B occurs times the conditional probability that A occurs, given that B occurs.

Keep in mind that the conditional probability and the multiplicative rule of probability are simply variations of the same thing.

Example:

In a certain city in the US some time ago, 30.7% of all employed female workers were white-collar workers. If 10.3% of all 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?

Solution:

We first define the following events

$$\begin{aligned} \text{F} &= \{\text{randomly selected worker who is female}\} \\ \text{W} &= \{\text{randomly selected white-collar worker}\} \end{aligned}$$

We are seeking 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

$$\begin{aligned} P(F) &= 10.3\% = 0.103 \\ P(W|F) &= 30.7\% = 0.307 \end{aligned}$$

Now using the multiplicative rule of probability we get,

$$P(F \cap W) = P(F)P(W|F) = (0.103)(0.30) = 0.0316 = 3.16\%$$

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

Example:

A college class has 42 students of which 17 are males and 25 are females. 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 a female and the second is male?

Solution:

Here we may define two events

$$\begin{gathered}\text{F1} = \{\text{first student selected is female}\}\\ \text{M2} = \{\text{second student selected is male}\}\end{gathered}$$

In this problem, we have a conditional probability situation. We want to determine the probability that the first student 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.

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

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

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

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 a female and the second one is a male.

Example:

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

$$\begin{aligned} \text{A} &= \{\text{first toss is head}\} \\ \text{B} &= \{\text{second toss is also head}\} \end{aligned}$$

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

Solution:

You would probably say no. Let's see if this is so. The sample space of this experiment is

$$S = \{HH, HT, TH, TT\}$$

Each of these simple events has a probability of $1/4 = 25\%$. Looking back at the problem, we have events A and B.

Since the first toss is a head, we have

$$P(A) = P(HH) + P(HT) = 1/4 + 1/4 = 1/2$$

And since the second toss is also a head,

$$P(B) = P(HH) + P(TH)1/4 + 1/4 = 1/2$$

Now, what is the conditional probability? Here it is,

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

What does this tell us? It tells us that $P(B) = 1/2$ and $P(B|A) = 1/2$ also. Which means knowing that the first toss resulted in a head does not affect the probability of the second toss. In other words,

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

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

Condition of Independence

If event B is independent of event A, then the occurrence of A does not affect the probability of the occurrence of event B. So we write,

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

Example:

The table below gives the number of physicists (in thousands) in the US cross classified by specialties ($P1, P2, P3, P4$) 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 the event $B1$ independent of $P3$?

Table 3.3:

	Industry ($B1$)	Academia ($B2$)	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). This data is hypothetical.

Solution:

The problem may appear a little difficult at first but it is actually much simpler, especially, if we make use of the condition of independence. All we need to do is to calculate $P(B1|P3)$ and $P(B1)$. If those two probabilities are equal, then the two events $B1$ and $P3$ are indeed independent. Otherwise, they are dependent. From the table we find,

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

And

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

Thus, $P(B1|P3) \neq P(B1)$ and so the event $B1$ is dependent on the event $P3$. This lack of independence results from the fact that the percentage of nuclear physicists who are working in the industry (3.5%) is not the same as the percentage of all physicists who are in the industry (10.9%).

Lesson Summary

1. 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 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
2. If $A \cap B$ contains no simple events, then A and B are **mutually exclusive**. Mathematically, this means $P(A \cup B) = P(A) + P(B)$.
3. The **Multiplicative Rule of Probability** states $P(A \cap B) = P(B)P(A|B)$.
4. If event B is **independent** of event A, then the occurrence of A does not affect the probability of the occurrence of event B. Mathematically, $P(B) = P(B|A)$.

Review Questions

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

$$\begin{aligned} &\text{A: } \{\text{Sum of the numbers is odd}\} \\ &\text{B: } \{\text{Sum of the numbers is } 9, 11, \text{ or } 12\} \end{aligned}$$

(a) Are events A and B independent? Why?
(b) Are events A and B mutually exclusive? Why?

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?

Review Answers

1. (a) No; $P(A) \neq P(A|B)$
   (b) No; $P(A \cap B) \neq 0$
2. 0.891

3.6 Basic Counting Rules

Learning Objectives

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

Inferential Statistics is a method of statistics that consists of drawing conclusions about a population based on information obtained from a subset or sample of the population. The main reason a sample of the population is only taken rather than the entire population (a census) is because it is less costly and it can be done more quickly than a census. In addition, because of the inability to actually reach everyone in a census, a sample can actually be more accurate than a census.

Once a statistician decides that a sampling is appropriate, the next step is to decide how to select the sample. That is, what procedure should we use to select the sample from the population? 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. It would not be also reasonable to estimate the average income of the entire state of California by sampling the average income of the wealthy residents of Beverly Hills. Therefore, the goal of sampling is to obtain a representative sample. For now, we will only study one powerful way of taking a sample from a population. It is called **random sampling**.

Random Sampling

A random sampling is a procedure in which each sample of a given size is equally likely to be the one selected. Any sample that is obtained by random sampling is called a **random sample.**

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 random sample.

Example:

Suppose you randomly select 4 cards from an ordinary deck of 52 cards and all the cards selected are kings. Would you conclude that the deck is still an ordinary deck or do you conclude that the deck is not an ordinary one and probably contains more than 4 kings?

Solution:

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 drawing 4 kings is not unusual, it is actually certain. However, if the deck was shuffled well, getting 4 kings is highly improbable. The point of this example is that if you want to select a random sample of 4 cards to draw an inference about a population, the 52 cards deck, it is important that you know *how* the sample was selected from the deck.

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?

Solution:

First we must require that each ticket has an equal chance of winning. That is, each ticket must have a probability of 1/100 of being selected. One fair way of doing that is mixing all the tickets in a container and *blindly* picking 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, say a million tickets, and we were asked to select 5 winning tickets. There are several standard procedures for obtaining random samples using a computer or a calculator.

Sometimes experiments have so many simple events that it is impractical to list them. However, in some experiments we can develop a counting rule, with the use of tree diagrams that can aid us to do that. The following examples show how that is done.

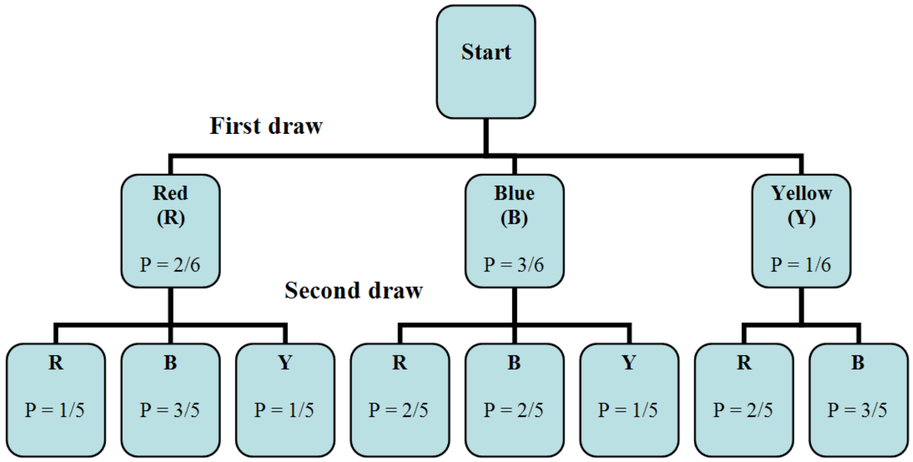
Example:

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 set it aside.

Then we will draw another one, record its color. With the aid of a tree diagram, calculate the probability of each outcome of the experiment.

Solution:

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



The tree diagram shows us the two stages of drawing two balls without replacing them back into the box. In the first stage, we pick a ball blindly. Since there are 2 red, 3 blue, and 1 yellow, then the probability of getting a red is 2/6. The probability of getting a blue is 3/6 and the probability of getting a yellow is 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 of getting the second ball, we look back at the tree diagram and observe the followings.

There are eight possible outcomes for the experiment:

RR: red on the 1^{st} and red on the 2^{nd}
RB: red on the 1^{st} and blue on the 2^{st}

And so on. Here are the rest,

$$RY, BR, BB, BY, YR, YB.$$

Next, we want to calculate the probabilities of each outcome.

$$
\begin{aligned}
P(R\ 1^{st} \text{ and } R\ 2^{st}) &= P(RR) = 2/6 \cdot 1/5 = 2/30\\
P(R\ 1^{st} \text{ and } B\ 2^{st}) &= P(RB) = 2/6 \cdot 3/5 = 6/30\\
P(RY) &= 2/6 \cdot 1/5 = 2/30\\
P(BR) &= 3/6 \cdot 2/5 = 6/30\\
P(YB) &= 3/6 \cdot 2/5 = 6/30\\
P(YB) &= 3/6 \cdot 1/5 = 3/30\\
P(YB) &= 1/6 \cdot 2/5 = 2/30\\
P(YB) &= 1/6 \cdot 3/5 = 3/30
\end{aligned}
$$

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

The method used to solve the example above can be generalized to any number of stages. This method is called the **Multiplicative Rule of Counting.**

The Multiplicative Rule of Counting

(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 the series of events A followed by B.

Another way of stating it:

(II) You have k sets of elements, n_1 in the first set, n_2 in the second set,..., and n_k in the kth set. Suppose 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:

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

Solution:

Let's summarize what we have.

Entrees:3
Appetizer: 5
Dessert: 4

We use the multiplicative rule above to calculate the number of different dinner meals that can be selected. We simply multiply all the number of choices per item together:

$$(3)(5)(4) = 60$$

There are 60 different dinners that can be ordered by the customers.

Example:

Here is a classic example. In how many different ways can you seat 8 people at a dinner table?

Solution:

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 and its own symbol. So we say "eight factorial," and we write 8!.

Factorial Notation

$$n! = n(n-1)(n-2)\ldots 1$$

Example:

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

Solution:

This is a more difficult problem than the examples above and we will use the second version of the Multiplicative Rule of Counting. We need to analyze it in the following way:

The executive positions can be denoted by $k = 3$ sets of elements that correspond to

$n_1 =$ The number of candidates that are available to fill the first position
$n_2 =$ The number of candidates remaining to fill the second position

$n_3 =$ The number of candidates remaining to fill the third position

Hence,

$$\begin{aligned} n_1 &= 30 \\ n_2 &= 29 \\ n_3 &= 28 \end{aligned}$$

The number of different ways to fill the three positions is

$$n_1 n_2 n_3 = (30)(29)(28) = 24,360 \text{ possible positions.}$$

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

Counting Rule for Permutations

The number of ways to arrange **in order** n different objects within r positions is

$$P_r^n = \frac{n!}{(n-r)!}$$

Example:

Let's go back to the previous example but this time we want to compute the number of ordered seating arrangements we have for 8 people for only 5 seats.

Solution:

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,

$$\begin{aligned} P_r^n &= \frac{n!}{(n-r)!} = \frac{8!}{(8-5)!} \\ &= \frac{8!}{3!} \\ &= \frac{40,320}{6} \\ &= 6720 \end{aligned}$$

Another way of solving this problem is to use the Multiplicative Rule of Counting,

Since there are only 5 seats available for 8 people, then for the first seat, there are eight people. For the second seat, there are seven remaining people, since one person has already been seated. For the third seat, there are 6 people, since two people are already seated. For the fifth seat, there are 4 people. After that we run out of seats. Thus

$$(8)(7)(6)(5)(4) = 6720.$$

Of course, the permutation rule is more powerful since it has the advantage of using the factorial. Most scientific calculators can do factorials permutations, so make sure to know how to do them on your calculator.

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 treasure. How many different slates of three candidates are there, if each candidate must specify which office he or she wishes to run for?

Solution:

Each slate is a list of one person for each of three positions, the provost, the general director and the treasure. If, for example, Mr. Smith, Mr. Hale, and Ms. Osborn wish to be on a slate together, there are several different slates possible, depending on which one will run for provost, general director and treasurer. So we are not just asking for the number of different groups of three names on a slate but we are also asking for a specific *order*, since it makes a difference which name is listed in which position.

So,

$$\begin{aligned} n &= 13 \\ r &= 3 \end{aligned}$$

Using the permutation formula,

$$P_r^n = \frac{n!}{(n-r)!} = \frac{13!}{(13-3)!} = 1716$$

There are 1716 different slates of officers.

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 rule will not work here since order is not important. We have a new formula that will compute different combinations.

Counting Rule for Combinations

The number of combinations of n objects taken r at a time is

$$C_r^n = \frac{n!}{r!(n-r)!}$$

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:

Back to our example above. How many different groups of three are there, taken out of 13 people?

Solution:

As explained in the previous paragraph, we are interested in combinations rather than permutations of 13 people taken 3 at a time. We use the combination formula

$$\begin{aligned} C_r^n &= \frac{n!}{r!(n-r)!} \\ C_3^{13} &= \frac{13!}{3!(13-3)!} = \frac{13!}{3!10!} = 286 \end{aligned}$$

There are 286 different groups of 3 to go to the convention.

In the above computation you can see that the difference between the formulas for nCr and nPr is in the factor $r!$ in the denominator of the fraction. Since $r!$ is the number of different orders of r things, and combinations ignore order, then 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 five books and read them in whichever order that pleases you. How many different combinations of 5 books are available from a list of 10?

Solution:

Since considerations of order in which the books are selected are not important, we compute the number of combinations of 10 books taken 5 at a time. We use the combination formula

$$C_r^n = \frac{n!}{r!(n-r)!}$$

$$C_5^{10} = \frac{10!}{5!(10-5)!} = 256$$

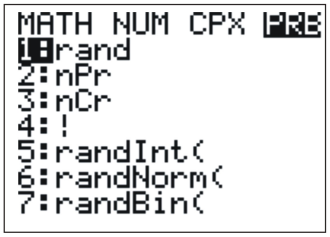
There are 252 different groups of 5 books that can be selected from a list of 10 books.

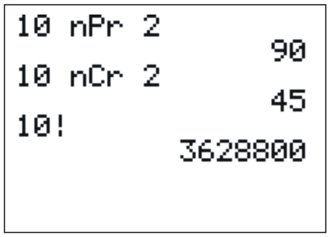
Technology Note

The TI-83/84 calculators and the EXCEL spreadsheet have commands for factorials, permutations, and combinations.

Using the *TI-83/84* Calculators

Press [**MATH**] and then choose PRB (Probability). You will see the following choices, among others: nPr, nCr, and **!** The screens show the menu and the proper uses of these commands.





```
10 nPr 2
              90
10 nCr 2
              45
10!
         3628800
```

Using EXCEL

In Excel the above commands are entered as follows:

- = PERMUT (10,2)
- = COMBIN (10,2)
- = FACT (10)

Lesson Summary

1. **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.

2. A **random sampling** is a procedure in which each sample of a given size is equally likely to be the one selected.
3. The **Multiplicative Rule of Counting** states: 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.
4. The **factorial**, ’ **!** ’, means $n! = n(n-1)(n-2)\dots 1$.
5. The number of **permutations** (***ordered*** arrangements) of n different objects within r positions is $P_r^n = \frac{n!}{(nr)!}$
6. 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) Look for a pattern in the results of a) through c) and try to figure out the number of outcomes for tossing a coin n times.
2. Flying into Los Angeles from Washington DC, you can choose one of three airlines and can 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?

Review Answers

1. (a) 4
   (b) 8
   (c) 32
   (d) 2^n
2. 6
3. $2,598,960$
4. $2,600,000$

Image Sources

Chapter 4

Discrete Probability Distribution

Introduction

In *An Introduction to Probability* we illustrated how probability can be used to make an inference about a population from a set of data that is observed from an experiment. Most of these experiments were simple events that were described in words and denoted by capital letters. However, in real life, most of our observations are in the form of numerical data. These data are observed values of what we call random variables. In this chapter, we will study random variables and learn how to find probabilities of specific numerical outcomes.

Recall that we defined an experiment as a process in which a measurement is obtained. For example, counting the number of cars in a parking lot, measuring the average daily rainfall in inches, counting the number of defective tires in a production line, or measuring the weight in kilograms of an African elephant cub. All these are called **quantitative variables.**

If we let x represent a quantitative variable that can be measured or observed in an experiment, then we will be interested in finding the numerical value of this quantitative variable. For example, $x =$ the weight in kg of an African elephant cub. If, however, the quantitative variable x takes a random outcome, we refer to it as a **random variable**.

Definition A **random variable** represents the numerical value of a simple event of an experiment.

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?

Solution:

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 (2 yes votes) to the second, 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 the 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 none (0) to a maximum number that the restaurant can handle.

There are two types of random variables, *discrete and continuous.* In this chapter, we will only describe and discuss discrete random variables and the aspects that make them important for the study of statistics.

4.1 Two Types of Random Variables

Learning Objectives

- Learn to distinguish between the two types of random variables: continuous and discrete.

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 $232/3$ or 12.23 students. The cost of a loaf of bread is also discrete, say $3.17, where we are counting dollars and cents, but not fractions of a cent.

However, if we are measuring the tire pressure in an automobile, we are dealing with a continuous 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 variables. Discrete random variables come usually from counting, say, the number of chickens in a coop, or the number of passing scores on an exam or the number of voters who showed up to the polls.

One way of distinguishing discrete and continuous variables is between any two values of a continuous variable, there are an infinite number of other valid values. This is not the case for discrete variables; between any two discrete values, there are an integer number $(0, 1, 2, \ldots)$ of valid values. For a discrete variable, these are considered **countable** values since you could count a whole number of them.

Discrete Random Variables and Continuous Random Variables

Random variables that assume a *countable number of values* are called discrete.

Random variables that can take any of the *countless number of values* are called continuous.

Example:

Here is a list of discrete random variables:

1. The number of cars sold by a car dealer in one month: $x = 0, 1, 2, 3, \ldots$
2. The number of students who were protesting the tuition increase last semester: $x = 0, 1, 2, 3, \ldots$. Notice that x could become very large.
3. The number of applicants who have applied for the vacant position at a company: $x = 0, 1, 2, 3, \ldots$

4. The number of typographical errors in a rough draft of a book: $x = 0, 1, 2, 3, \ldots$

Example:

Here is a list of continuous random variables:

1. The length of time it took the truck driver to go from New York city to Miami: $x > 0$, where x is the time.
2. The depth of oil drilling to find oil: $0 < x < c$, where c is the maximum depth possible.
3. The weight of a truck in a truck weighing station: $0 < x < c$, where c is the maximum weight possible.
4. The amount of water loaded in a $12-$ ounce bottle in a bottle filling operation: $0 < x < 12$.

Lesson Summary

1. A **random variable** represents the numerical value of a simple event of an experiment.
2. Random variables that assume a **countable** number of values are called **discrete**.
3. Random variables that can take any of the **countless** number of values are called **continuous**.

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.

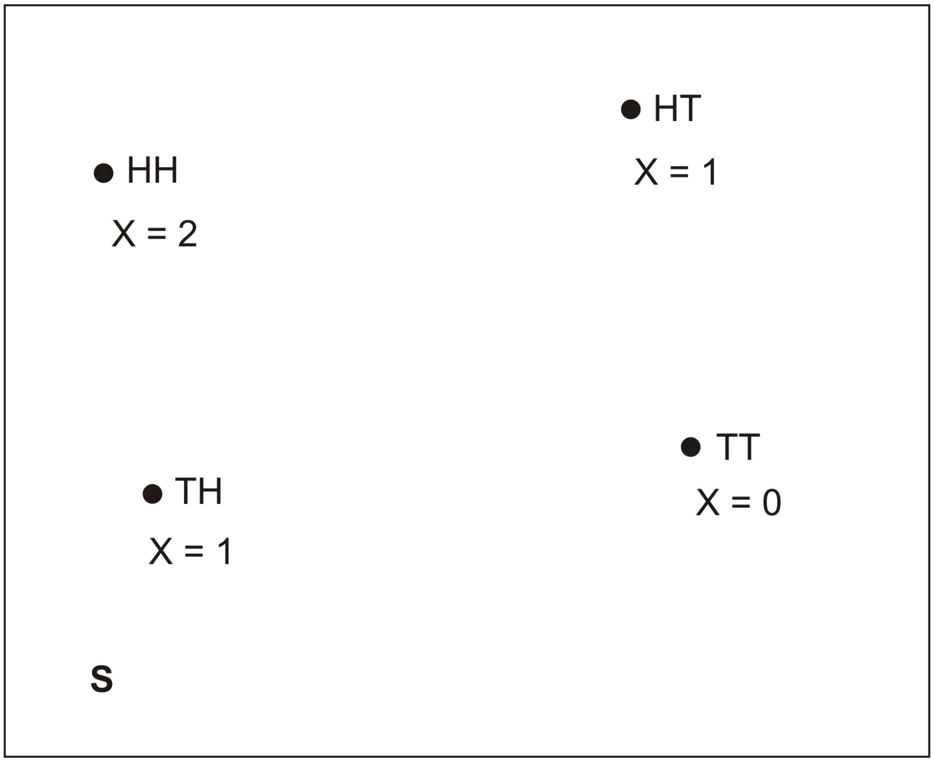
Now, we want to specify the possible values that a discrete random variable can assume. The example below illustrates how.

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.

Solution:

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 probability $1/4$. Since x is the number of heads observed, then $x = 0, 1, 2$. The Venn diagram below shows the two-coin experiment.



From the Venn diagram, we can identify the probabilities of the simple events associated with each value of x:

$$\begin{aligned}
P(x=0) &= P(TT) = 1/4\\
P(x=1) &= P(HT) + P(TH) = 1/4 + 1/4 = 1/2\\
P(x=2) &= P(HH) = 1/4
\end{aligned}$$

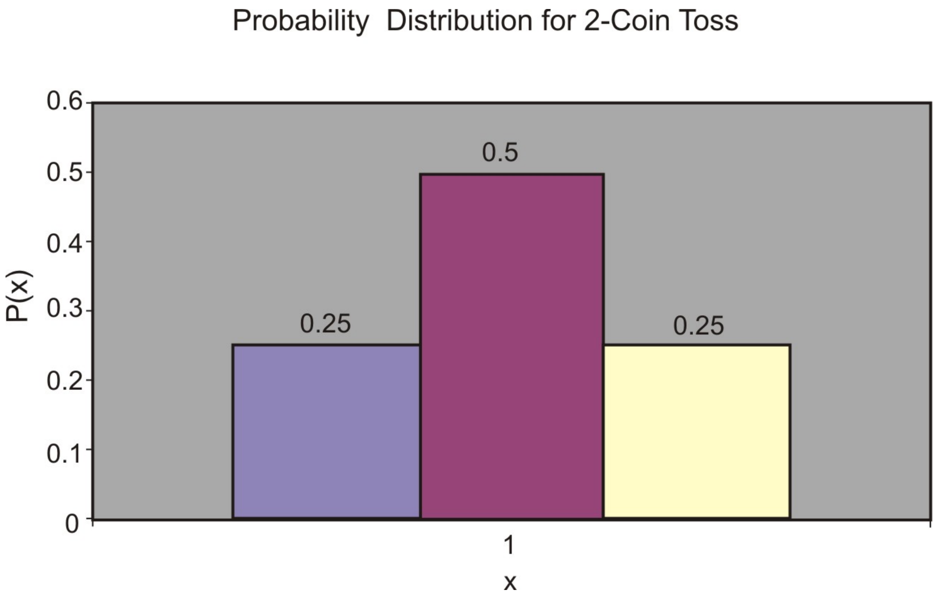
Thus, we have just had a complete description of the values of the random variables and have calculated the associated probabilities that are distributed over these values. We refer to it as the **probability distribution.** This probability distribution can be represented in different ways, sometimes in a tabular form and sometimes in a graphical one. Both forms are shown below.

In tabular form,

x	$P(x)$
0	1/4
1	1/2
2	1/4

Figure: The Tabular Form of the Probability Distribution for the Random Variable in the First Example.

In a graphical form,



We can also describe the probability distribution in a mathematical formula, which we will do later in the chapter.

Probability Distribution

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.

Two conditions must be satisfied for all probability distributions:

Two conditions must be satisfied for the probability distribution:

$P(x) \geq 0$, for all values of x

$$\sum_x P(x) = 1$$

The symbol $\sum_x P(X)$ means "add the values of $P(x)$ for all values of x"

Example:

What is the probability distribution of the number of yes votes for three voters (see the first example in the first section Introduction)

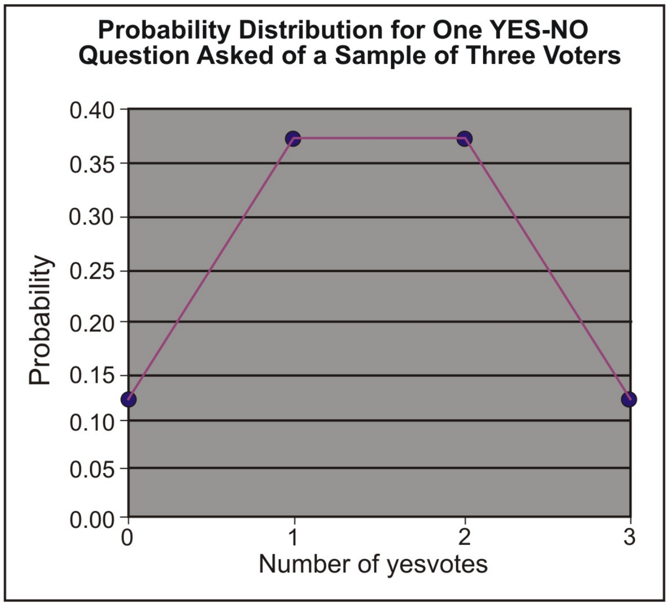
Solution:

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

Table 4.2:

Value of Random Variable	Probability
(number of Yes votes) Probability	
3	$1/8 = 0.125$
2	$3/8 = 0.375$
1	$3/8 = 0.375$
0	$1/8 = 0.125$

Figure: Tabular Representation of the Probability Distribution for the Random Variable in this Example.



The table and the graph are two ways to show the probability distribution. Note that the graph of a probability distribution can be either a line graph or a bar graph.

Lesson Summary

1. 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.

2. All probability distributions must satisfy:

$P(x) > 0$ (for all values of x)

$$\sum_x P(x) = 1$$

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 is greater than zero?
   (d) What is the probability that $x = -2$?

2. A fair die is tossed twice and the up face is recorded. Let x be the sum of the up faces.
   (a) Give the probability distribution of 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? 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?

Review Answers

1. (a) $x = \{-4, 0, 1, 3\}$;
   (b) 1;
   (c) 0.6;
   (d) 0
2. (a)

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

   (b) 15/36;
   (c) 21/36;
   (d) 18/36, 18/36;
   (e) 6/36
3. 7/8

4.3 Mean and Standard Deviation of Discrete Random Variables

Learning Objectives

- Know the definition of the mean, or expected value, of a discrete random variable.
- Know the definition of the standard deviation of a discrete random variable.
- Know the definition of variance of a discrete random variable.
- Find the expected value of a variable.

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:

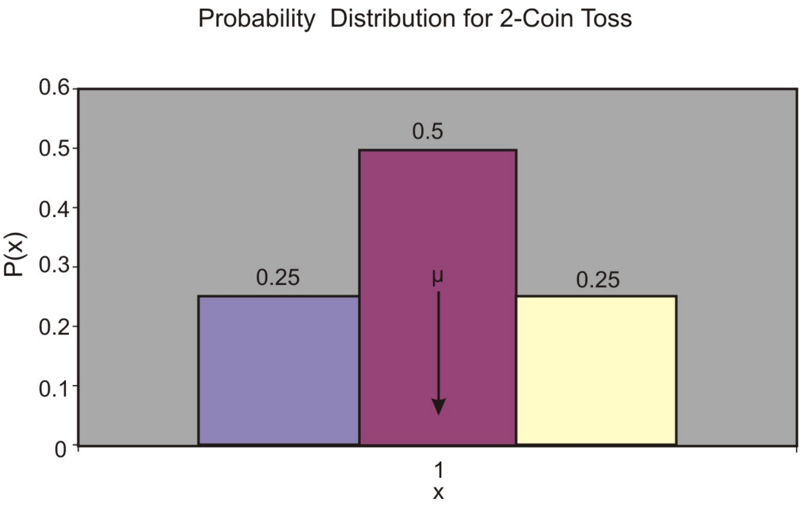
Go back to the $2-$coin experiment in the previous example and calculate the mean μ of the distribution.

Solution:

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 get the population mean, we simply multiply each possible outcome of x by its associated probability and then summing over all possible values of x,

$$\mu = 0(1/4) + 1(1/2) + 2(1/4) = 0 + 1/2 + 1/2 = 1$$



Mean Value or Expected Value

The mean value or expected value of a discrete random variable x is given by

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

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?

Solution:

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 0.1% and the probability that the customer will live is $1 - 0.001 = 99.9\%$. 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 4.3:

Gain,x	Simple events	Probability
\$310	Live	99.9%
$-\$14,690$	Die	0.1%

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 $\$310 - \$15,000 = -\$14,690$ (a loss). Therefore, the expected profit is,

$$\begin{aligned}
\mu &= E(x) = \sum_x xp(x) \\
\mu &= (310)(99.9\%) + (310 - 15,000)(0.1\%) \\
&= (310)(0.999) + (310 - 15,000)(0.001) \\
&= 309.69 - 14.69 = \$295 \\
\mu &= \$295
\end{aligned}$$

This tells us that if the company would sell a very large number of the 1−year \$15,000 policy to too many people, it will make on average a profit of \$295 per sale next year.

Another approach is to calculate the expected payout, not the expected gain

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

Since the company charges \$310 and expects to pay out \$15, the profit for the company is \$295 on every 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 so, we first define the *population variance*, σ^2. It is defined as the average of the squared distance of the values of the random variable x to the mean value μ. The formal definitions of the variance and the standard deviation are shown below.

The Variance

The variance of a discrete random variable is given by the formula

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

The Standard Deviation

The square root of the variance 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. If 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 $p(x)$.

Solution:

a. We use the formula

$$\begin{aligned}
&\mu = E(x) = \sum_x xp(x)\\
&\mu = 0(.002) + 1(.029) + 2(.132) + 3(.309) + 4(.360) + 5(.168)\\
&\mu = 3.50
\end{aligned}$$

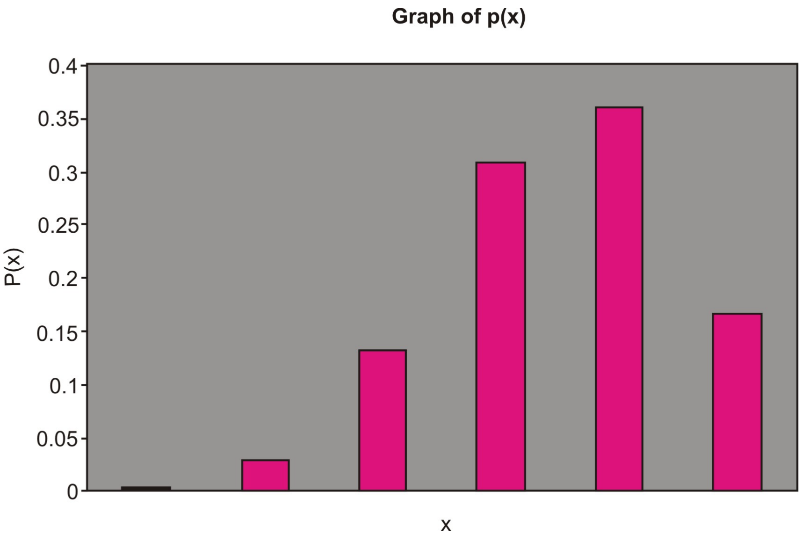
b. We first calculate the variance of x:

$$\begin{aligned}\sigma^2 =& \sum_x (x-\mu)^2 p(x)\\ =&(0-3.5)^2(.002)+(1-3.5)^2(.029)+(2-3.5)^2(.132)\\ &+(3-3.5)^2(.309)+(4-3.5)^2(.36)+(5-3.5)^2(.168)\\ \sigma^2 =&1.05\end{aligned}$$

Now we calculate the standard deviation,

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

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



We can use the mean μ and the standard deviation σ to describe $p(x)$ in the same way we used $\bar{x}$ and s to describe the relative frequency distribution. Notice that $\mu = 3.5$ locates the center of the probability distribution. In other words, if the five cancer patients receive chemotherapy treatment we expect the number x that are cured to be near 3.5. The standard deviation $\sigma = 1.02$ measures the spread of the probability distribution $p(x)$.

Lesson Summary

1. The **mean value** or **expected value** of a discrete random variable x is given by $\mu = E(x) = \sum_x xp(x)$.
2. The **variance** of a discrete random variable is given by $\sigma^2 = \sum_x (x-\mu)^2 p(x)$.
3. The square root of the variance σ^2 is the **standard deviation** of a discrete random variable, $\sigma = \sqrt{\sigma^2}$.

Review Questions

1. Consider the following probability distribution:

x	0	1	2	3	4
$p(x)$	0.1	0.4	0.3	0.1	0.1

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 questioned each inmate to find out how many times the inmate has been convicted. The officer came up with the following table that shows the relative frequencies of x:

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 frequency as approximately the probability, what is the expected value of the number of times of previous convictions of an inmate?

Review Answers

1. (a) 1.7;
(b) 1.21;
(c) 1.1
2. 1.29

4.4 The Binomial Probability Distribution

Learning Objectives

- Know the characteristics of the binomial random variable.
- Know the 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 the Binomial distribution can be applied.
- Use the Binomial distribution to solve statistical problems.

Many experiments result in responses for which there is only two possible outcomes, either, a Yes or a No, Pass or Fail, Good or Defective, Male or Female, etc. A simple example is the toss of a coin, say 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 values of x range from 0 to 6. Many other experiments are equivalent to the toss of a coin, if done n times and we are interested in the probability distribution x of times that one of the two outcomes is observed (from 0 to n). Random variables that have this characteristic are called **binomial random variables.**

For example, let us say that we select 100 students from a large university campus and ask them whether they are in favor of a certain issue that is going 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 do not favor the issue is $1 - p$. Sampling 100 students in this way is equivalent to tossing a coin 100 times.

The experiment that we have been describing is an example of a **binomial experiment.** It can be identified by the following characteristics:

Characteristics of a Binomial Experiment

- The experiment consists of n number of 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 trials are independent of each other.
- The binomial random variable x is the number of successes in the n trials.

Example:

In the following two examples, decide whether x is a binomial random variable.

1. Suppose a university decides to give two scholarships to two students. The pool of applicants is ten students, six males and four females. If all the ten applicants are equally qualified and the university decides to randomly select two. Let x be the number of female students who receive the scholarship.
2. A company decides to conduct a survey on customers to see if their new product, a new brand of shampoo, will sell well. The company chooses 100 randomly selected customers and ask them to state their preference among the new shampoo and two leading shampoos in the market. Let x be the number of the 100 customers who choose the new brand over the other two.

Solution:

1. It is best to review the characteristics of a binomial random variable, in the preceding box. If the first student selected in a female, then the probability that the second student is a female is $3/9$. 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.
2. 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.

Example:

The American Heart Association claims that only 10% of adults over 30 can pass the minimum fitness requirement that is established by them. Suppose that four adults are randomly selected and given the fitness test. Let x be the number of the four who pass the test. What is the probability distribution for x?

Solution:

We first check the characteristics of this experiment. We see that it is a binomial experiment because there are two outcomes (pass or fail the test) and the outcomes are independent because one student's success or failure does not influence any other student's performance. What are the possible values of x? They are $x = 0, 1, 2, 3, 4$. We will take each event and study it.

For $x = 0$ no one passes the fitness test. This is equivalent to

$$\{\text{FFFF}\}$$

F stands for failure on the test. The probability that all will fail the test is:

$$\begin{aligned} p(x = 0) &= p(FFFF) = p(F)p(F)p(F)p(F) \\ &= (.9)(.9)(.9)(.9) = (.9)^4 = 0.6561 \end{aligned}$$

Thus, there is a 65.61% chance that all four will fail the fitness test.

For $x = 1$, only one will pass the test. The list of all possible simple events is:

$$\{\text{SFFF, FSFF, FFSF, FFFS}\}$$

S stands for success on the test. Note that each of these simple events has the same probability, $(.1)(.9)^3$. So the probability of all the simple events is

$$p(x = 1) = 4[(.1)(.9)^3] = 0.2916$$

In other words, there is a chance of 29.16% that only one will pass the test.

For $x = 2$, only two will pass the test. This event has six simple events:

$$\{\text{SSFF, SFSF, SFFS, FSSF, FSFS, FFSS}\}$$

Each of these events has a probability of $(.1)^2(.9)^2$ so

$p(x = 2) = 6[(.1)^2(.9)^2] = 0.0486$

This is a chance of 4.86% that two will pass the test.

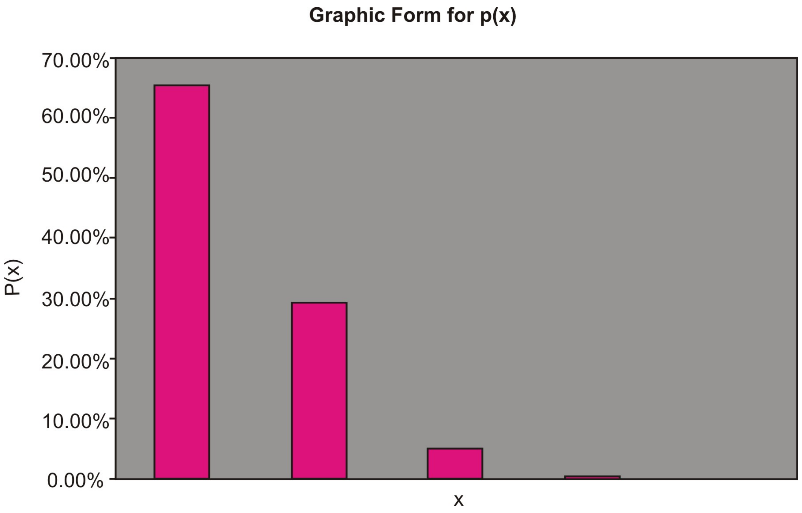
Similarly,

$$\begin{aligned} p(x = 3) &= 4[(.1)^3(.9)] = 0.0036 \\ p(x = 4) &= (.1)^4 = 0.0001 \end{aligned}$$

To summarize, we show all the probability distributions in tabular and graphical forms:

x	$p(x)$
0	65.61%
1	29.16%
2	4.86%
3	0.36%
4	0.01%

Figure: The Probability Distribution in Table Form of the Binomial Experiment in this Example (Four Adults who take a Fitness Test).



Note: $p(4)$ is too small to show on the graph.

In the table above, we displayed the probability distribution of random variable $x-$ the number of people, out of 4, who will pass the fitness test. Although the work done to get the probability distribution is quite involved, it is rather simple compared to most practical situations. Imagine if our sample was not 4 but 100. In this case, there would be over a million possible outcomes. To tabulate such outcomes would be impractical. Fortunately, there is a formula for the binomial distribution that saves us all the numerous calculations.

If an n experiments are performed, then the number of ways to get exactly x successes S is equal to the binomial coefficients. In other words the event of obtaining exactly x successes S in the n trials consists of C_x^n outcomes. From previous chapters, recall that

$$C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the number of simple events that have x successes and $(n-x)$ failures. Here in this chapter, we introduce the above notation $\binom{n}{x}$ that is equivalent to C_x^n of Chapter 3. Both notations are used interchangeably in statistics and it is a good idea to be familiar with both.

The Binomial Probability Distribution

Suppose n experiments are performed, with the probability of successes on any given trial is p. Let x denote the total number of successes in the n trials. Then the probability distribution of the random variable x is given by the formula,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

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:

1. Identify a success.
2. Determine p, the success probability.
3. Determine n, the number of experiments or trials.
4. Use the binomial formula to write the probability distribution of x.

The examples below will help you learn how to use the binomial formula.

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?

Solution:

Showing the four steps listed above.

1. The success is any call that is longer than 3.8 minutes.
2. The probability $p = 25\% = 0.25$.
3. The number of trials n is 3.
4. Thus we can now use the binomial probability formula,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Substituting we have: $p(x) = \binom{3}{x}(.25)^x(1 - .25)^{3-x}$

a. For $x = 2$,

$$\begin{aligned} p(x) &= \binom{3}{2}(.25)^x(1 - .25)^{3-2} \\ &= (3)(.25)^2(1 - 25)^1 \\ &= 0.14 \end{aligned}$$

The probability is 14% that exactly two out of three randomly selected calls will last longer than 3.8 minutes.

b. Here, $x = 0$. We use the binomial probability formula,

$$\begin{aligned} p(x = 0) &= \binom{3}{0}(.25)^0(1 - .25)^{3-0} \\ &= \frac{3!}{0!(3-0)!}(.25)^0(.75)^3 \\ &= 0.422 \end{aligned}$$

The probability is 42.2% that none of the three randomly selected calls will last longer than 3.8 minutes.

Example:

A car dealer knows that from past experience he can make a sale to 20% of the customers that 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?

d. Determine the probability distribution for the number of sales.

Solution:

The success here is making a sale to the customer. The probability that the seller makes a sale to any customer is $p = 20\% = 0.2$. The number of trials is $n = 5$. The binomial probability formula for our case is

$$p(x) = \binom{5}{x}(.25)^x(.8)^{5-x}$$

a. Here we want the probability of exactly 3 sales, $x = 3$:

$$p(x) = \binom{5}{3}(.2)^3(.8)^{5-x} = 0.051$$

This means that the probability that the sales person makes exactly three sales in five attempts is 5.1%.

b. The probability at most one customer means

$$\begin{aligned} p(x \leq 1) &= p(0) + p(1) \\ &= \binom{5}{0}(.2)^0(.8)^{5-0} + \binom{5}{1}(.2)^1(08)^{5-1} \\ &= 0.328 + 0.410 = .738 \end{aligned}$$

c. The probability of at least one sale is

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

We can now apply the binomial probability formula to calculate the five probabilities. However, we can save time by calculating the complement of the probability,

$$\begin{aligned} p(x \geq 1) = 1 - p(x < 1) &= 1 - p(x = 0) \\ 1 - p(0) = 1 - &\binom{5}{0}(.2)^0(.8)^{5-0} \\ &= 1 - 0.328 = 0.672 \end{aligned}$$

This tells us that the salesperson has a chance of 67.2% of making at least one sale in five attempts.

d. Here, we are asked to determine the probability distribution for the number of sales x in five attempts. So we need to compute $p(x)$ for $x = 1, 2, 3, 4,$ and 5. We use the binomial probability formula for each value of x. The table below shows the probabilities.

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.

In many applications of the binomial distribution, it is necessary that we know how to calculate the *mean* and the *standard deviation.* To compute these we use the following formulas shown in the box below.

Mean, Variance, and Standard Deviation for a Binomial Random Variable

Mean: $\mu = np$

Variance: $\sigma^2 = npq = np(1-p)$

Standard Deviation: $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$

Example:

A poll of twenty voters is taken to determine the number in favor of a certain candidate for mayor. Suppose that 60% of all the city's voters favor this candidate.

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

b. Find the probability that $x \le 10$.

c. Find the probability that $x > 12$.

d. Find the probability that $x = 11$.

Solution:

a. Since a 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 is $60\% = 0.6$, the fraction of the total voters who favor the candidate. Therefore, to calculate the mean and the standard deviation,

$$\begin{aligned}\mu &= np = 20(.6) = 12\\ \sigma^2 &= np(1-p) = 20(.6)(.4) = 4.8\end{aligned}$$

The standard deviation

$$\sigma = \sqrt{4.8} = 2.2$$

b. To calculate the probability for $p(x)$,

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

or

$$p(x \le 10) = \sum_{x=0}^{10} p(x) = \sum_{x=0}^{10} \binom{20}{x} (.6)^x (.4)^{20-x}$$

As you can see, this can be very tedious calculations and it is best to resort tables or calculators. If you are using a table, look up the *Cumulative Binomial Probability Table.* To find $p(x \leq 10)$ for $n = 20$ and $p = 0.6$, we first find the column that corresponds to $p = 0.6$ and then the row corresponding for $k = 10$. The value is

$$p(x \leq 10) = 0.245$$

However, please see the box below (Technology Note) to learn more about other options.

c. To find the probability that $p > 12$, the formula says,

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

Using the complementation rule,

$$\begin{aligned} p(x > 12) &= 1 - [p(1) + p(2) + \ldots + p(12)] \\ &= 1 - p(x \leq 12) \\ &= \sum_{x=0}^{12} 1 - p(x) \end{aligned}$$

Consulting tables or calculators (see Box below, Technology Note), $k = 12, p = .6$, we get the result 0.584. Thus

$$P(x > 12) = 1 - 0.584 = 0.416$$

d. To find the probability of exactly 11 voters favor the candidate,

$$p(x = 11) = p(x \leq 11) - p(x \leq 10) = .404 - .245 = .159$$

Technology Note

The TI-83/84 calculators and the EXCEL spreadsheet have commands for the Binomial distribution.

- Press [**DIST**] and scroll down (or up) to **binompdf** (Press [**ENTER**] to place **binompdf** on your home screen.) Type values of μ and x separated by commas and press [**ENTER**].
- Use **binomcdf** (for probability of at most x successes.

Note: it is not necessary to close the parentheses.

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 probability of at least x successes, replace "false" with "true"

Lesson Summary

1. Characteristics of a **Binomial Experiment**

- The experiment consists of n number of 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 trials are independent of each other.
- The binomial random variable x is the number of $S's$ in the n trials.

2. The **binomial probability distribution** is:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

3. For the binomial random variable:

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

Review Questions

1. Suppose x is a binomial random variable with $n = 4, p = 0.2$. Calculate $p(x)$ for the values: $x = 0, 1, 2, 3, 4, 5$. 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.
(b) 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.

Review Answers

1. (a)

x	0	1	2	3	4
$p(x)$	0.4096	0.4096	0.1536	0.0256	0.0016

2. (a)

x	0	1	2	3	4	5
$p(x)$	0.031	0.157	0.312	0.157	0.031	

   (b) $2.5, 1.25$

3. (a) 0.590;
   (b) 0.410

4.5 The Poisson Probability Distribution

Learning Objectives

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

The Poisson 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 is 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 relation to the binomial distribution, if the number of trials n gets larger and larger as the probability of successes p gets smaller and smaller, we obtain the Poisson distribution. The box below shows some of the basic characteristics of the Poisson distribution.

Characteristics of the 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

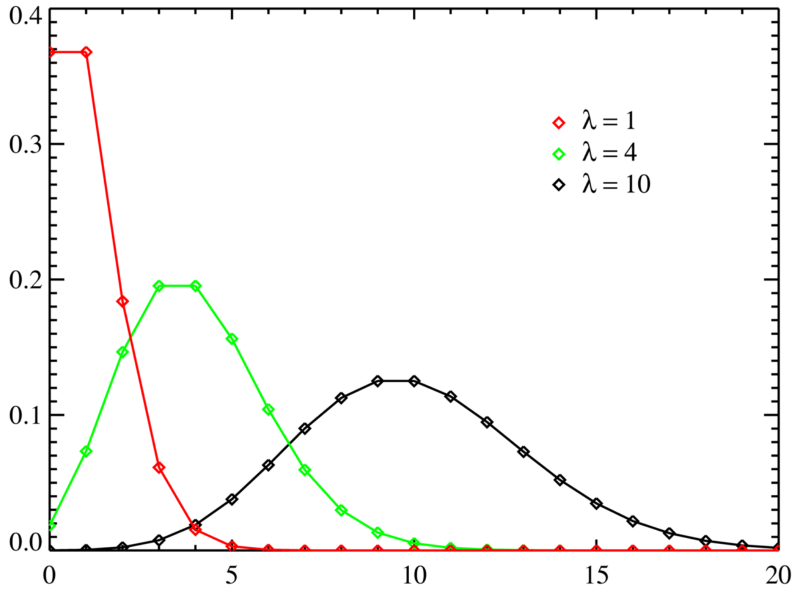
Mean and Variance

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

where

$\lambda =$ The mean number of events during the time, distance, volume or area.

$e =$ the base of the natural logarithm $= 2.718281828\ldots$



Example:

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

- What is the probability distribution for x, the number of fish that you will catch, in 7 hours ?
- What is the probability that you will catch $0, 3$, or 10 fish in seven hours of fishing?
- What is the probability that you will catch 4 or more fish in 7 hours?

Solution:

1. The mean value number is 3 fish in 2 hours or 1.5 fish/1 hour . This means, over seven hours, the mean number events will be $\lambda = 1.5$ fish/hour $*$ 7 hours $= 10.5$ fish. Thus,

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

2. To calculate the probabilities that you will catch $0, 3$, or 10:

$$p(0) = \frac{(10.5)^0 e^{-10.5}}{0!} \approx 0.000027 \approx 0\%$$

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

$$
\begin{aligned}
p(3) &= \frac{(10.5)^3 e^{-10.5}}{3!} \approx 0.0052 \approx 0.5\% \\
p(10) &= \frac{(10.5)^{10} e^{-10.5}}{10!} \approx 0.1212 \approx 12\%
\end{aligned}
$$

The curve in the figure with $\lambda = 10$ is very similar graph of the function for this problem.

3. The probability that you will catch 4 or more fish in 7 hours is,

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

Using the complement rule,

$$
\begin{aligned}
P(x \geq 4) &= 1 - [p(0) + p(1) + p(2) + p(3)] \\
&\approx 1 - 0.000027 - 0.000289 - 0.00152 - 0.0052 \\
&\approx 0.9903
\end{aligned}
$$

Therefore there is about 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.

Solution:

a. The mean and the variance are both equal to λ. Thus,

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

Then the standard deviation is,

$$\sigma = 1.58$$

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

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

c. To find the probability of two or more sightings,

$$p(x \geq 2) = p(2) + p(3) + \ldots = \sum_{x-2}^{\infty} p(x)$$

This is of course an infinite sum and it is impossible to compute. However, we can use the complementation rule,

$$\begin{aligned} p(x \geq 2) &= 1 - p(x \leq 1) \\ &= 1 - [p(0) + p(1)] \end{aligned}$$

—Calculating,—

$$\begin{aligned} &= 1 - \frac{(2.5)^0 e^{-2.5}}{0!} - \frac{(2.5)^1 e^{-2.5}}{1!} \\ &\approx 0.713 \end{aligned}$$

So, according to the Poisson model, the probability that two or more sightings are made in a month is 71.3%

Technology Note

The TI-83/84 calculators and the EXCEL spreadsheet have commands for the Poisson distribution.

Using the *TI-83/84* Calculators

- Press [**DIST**] and scroll down (or up) 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 probability of at most x successes.

Note: it is not necessary to close the parentheses.

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 probability of at least x successes, replace "false" with "true"

Lesson Summary

1. Characteristics of the **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.

2. **Poisson Random Variable**

Mean and Variance

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

where

$\lambda =$ The mean number of events during the time, distance, volume or area.

$e =$ the base of the natural logarithm $= 2.718281828\ldots$

4.6 The Geometric Probability Distribution

Learning Objectives

- Know the definition of the Geometric distribution.

- Identify the characteristics of the Geometric distribution.
- Identify the type of statistical situation to which the Geometric distribution can be applied.
- Use the Geometric distribution to solve statistical problems.

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

The 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. The box below lists the main characteristics of the geometric random variable.

Characteristics of the 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 do we wait until a success is observed? For example, in the world of business, the business owner wants to know the length of time a customer will wait for some type of service. Or, an employer, who is interviewing potential candidates for a vacant position, wants to know how many interviews he/she has to conduct until the perfect candidate for the job is found. Or, 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

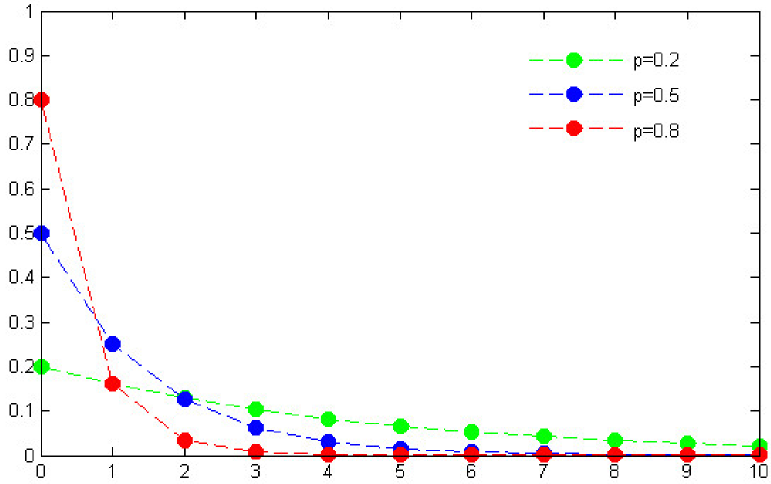
$$\begin{aligned} p(x) &= (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots \\ \mu &= \frac{1}{p} \\ \sigma^2 &= \frac{1-p}{p^2} \end{aligned}$$

where,

$p =$ Probability of an S outcome

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

The figure below plots a few probability distributions of the Geometric distributions. Note how the curve starts high and drops 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 juror being admitted, is 50%.

1. Find the mean and the standard deviation.
2. Find the probability that more than two prospective jurors must be examined before one is admitted to the jury.

Solution:

1. The mean and the standard deviation are,

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

Thus

$$\sigma = \sqrt{2} = 1.41$$

2. To find the probability that more than two prospective jurors will be examined before one is selected,

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

Obviously, this is an infinitely large sum so it is best to use the complementary rule:

$$\begin{aligned} p(x > 2) &= 1 - p(x \le 2) \\ &= 1 - [p(1) + p(2)] \end{aligned}$$

Before we go any further, we need to find $p(1)$ and $p(2)$. Substituting into the formula for $p(x)$:

$$\begin{aligned} p(1) &= (1-p)^{1-1}p = (.5)^0(.5) = 0.5 \\ p(2) &= (1-p)^{2-1}p = (.5)^1(.5) = 0.25 \end{aligned}$$

Then,

$$\begin{aligned} p(x > 2) &= 1 - p(x \le 2) \\ &= 1 - (.5 + .25) = 0.25 \end{aligned}$$

This result says that there is a 25% chance that more than two prospective jurors will be examined before one is admitted to the jury.

Technology Note

The TI-83/84 calculators have commands for the geometric distribution.

- Press **2nd** and scroll down (or up) to **geometpdf** (press **[ENTER]** to place **geometpdf** on your home screen.) Type values of p and x separated by a comma and press **[ENTER]**
- Use **geometcdf**(for probability of at least x successes.

Note: it is not necessary to close the parentheses.

Lesson Summary

1. Characteristics of the **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.

2. Probability distribution, mean, and variance of a **Geometric Random Variable**

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

where,

$p =$ Probability of an S outcome

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

Review Questions

1. A prison reports that the number of escape attempts per month has a Poisson distribution with a mean value of 1.5.
   (a) Calculate the probability that exactly three escapes will be attempted during the next month.
   (b) Calculate the probability that exactly one escape will be attempted during the next month.
2. If 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?
3. An oil company had determined that the probability of finding oil at a particular drilling operation is 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.)

Review Answers

1. (a) 0.1255;
   (b) 0.3347
2. 0.1088
3. 8.19%

Image Sources

Chapter 5

Normal Distribution

5.1 The Standard Normal Probability Distribution

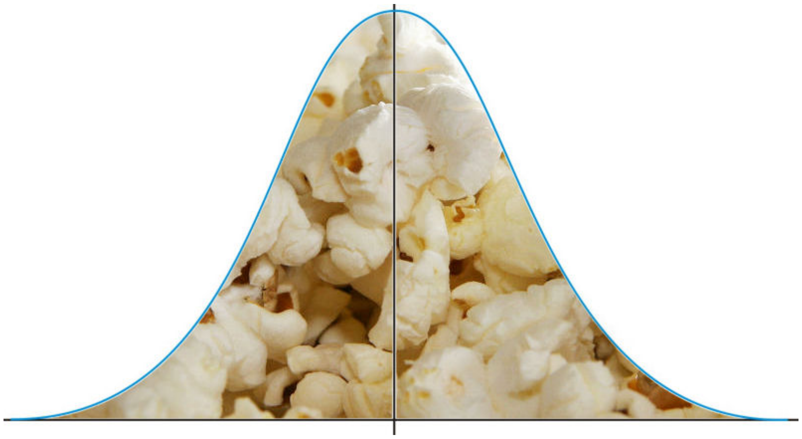
Learning Objectives

- Identify the characteristics of a normal distribution.
- Identify and use the Empirical Rule ($68 - 95 - 99.7$ rule) for normal distributions.
- Calculate a $z-$score and relate it to probability.
- Determine if a data set corresponds to a normal distribution.

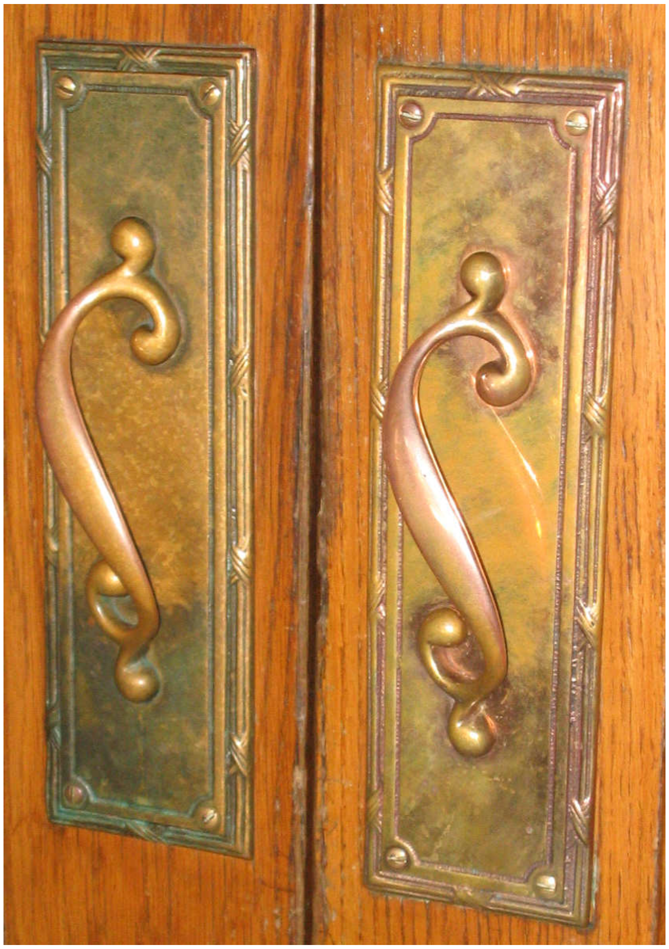
Introduction

Most high schools have a set amount of time in between classes in 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 then the number of students entering decreases again, with one or two students barely making it on time, or perhaps even coming in late! Try the same by watching students enter your school cafeteria at lunchtime. Spend some time in a fast food restaurant or café before, during, and after the lunch hour and you will most likely observe similar behavior.

Have you ever popped popcorn in a microwave? Think about what happens in terms of the rate at which the kernels pop. Better yet, actually do it and listen to what happens! For the first few minutes nothing happens, 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. Try measuring the height, or 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 very students with very low measurements and a couple with very high measurements with the majority of students centered around a particular value.



Sometimes the door handles in office buildings show a wear pattern caused by thousands, maybe millions of times being pulled or pushed to open the door. Often you will see that there is a middle region that shows by far the most amount of wear at the place where people opening the door are the most likely to grab the handle, surrounded by areas on either side showing less wear. On average, people are more likely to have grabbed the handle in the same spot and less likely to use the extremes on either side.

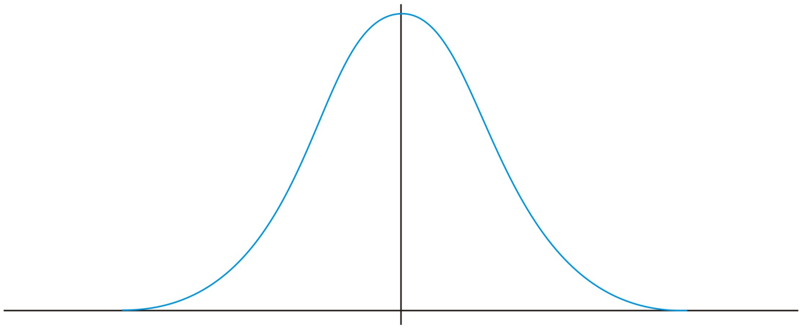


All of these examples show a typical pattern that seems to be a part of many real life phenomena. In statistics, because this pattern is so pervasive, it seems to fit to call it “normal”, or more formally the **normal distribution**. The normal distribution is an extremely important concept because it occurs so often in the data we collect from the natural world, as well as 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

If you think of graphing 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, or bell curve.

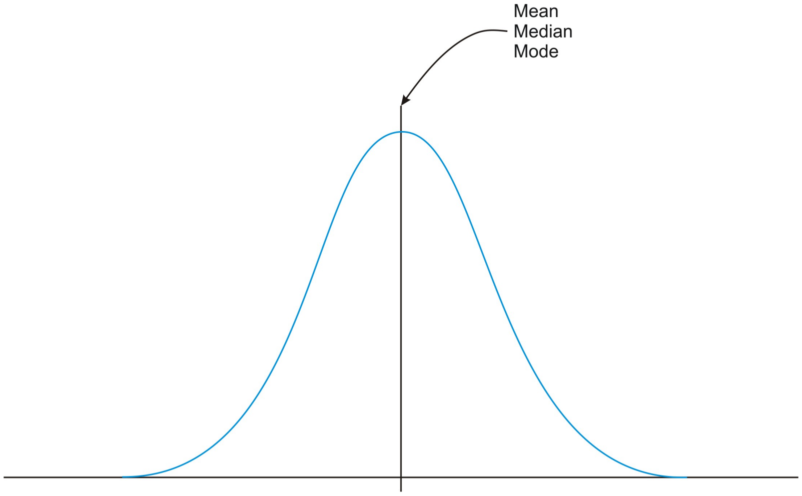


The Normal Distribution

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. In practical data collection, the distribution will never be exactly symmetric, so just like situations involving probability, a true normal distribution results from an infinite collection of data, or from the probabilities of a continuous random variable.

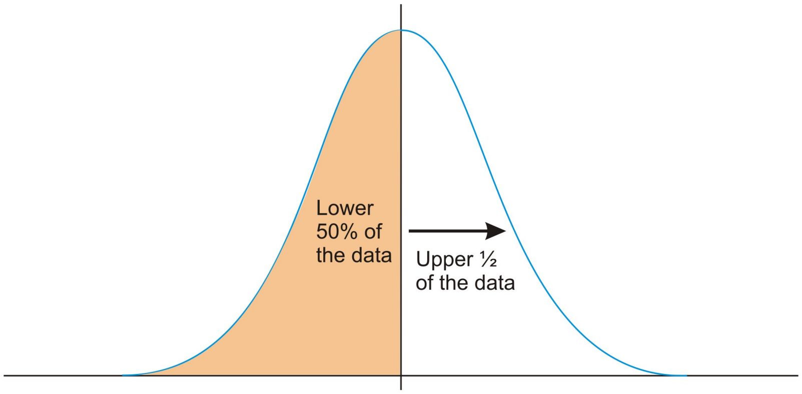
Center

Due to this exact symmetry the center of the 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, **mean, median, and mode** are equal.



Normal Distribution Center

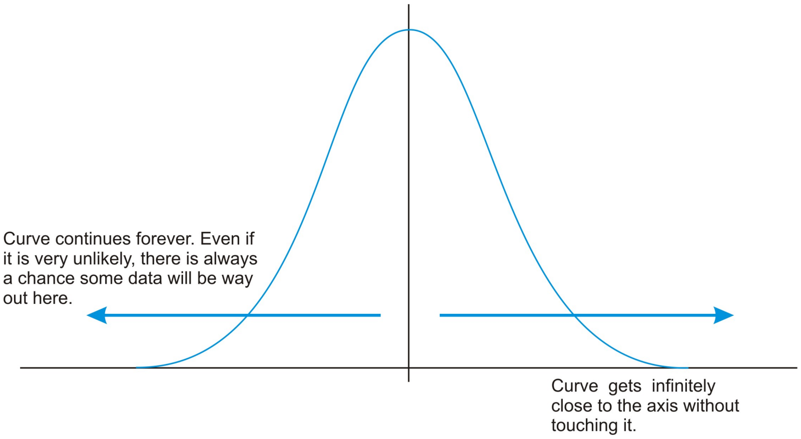
It is also important to realize that this center peak divides the data into two equal parts.



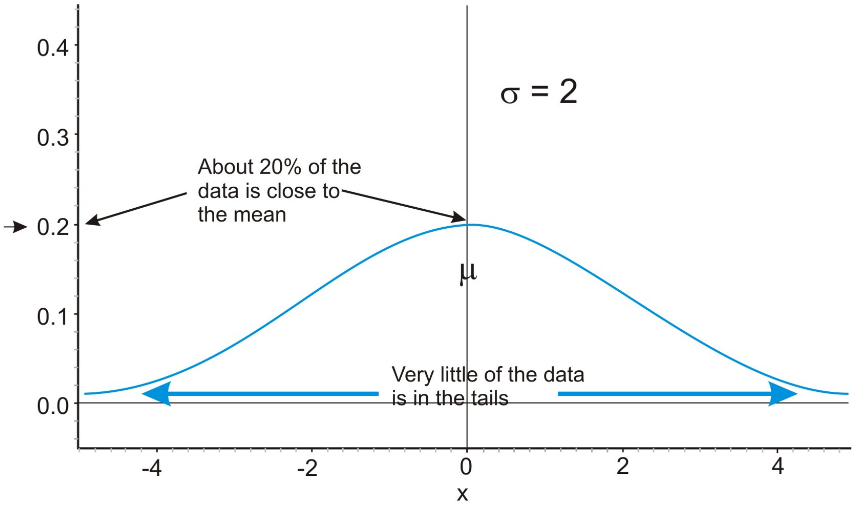
Normal Distribution divided into equal halves by the center

Spread

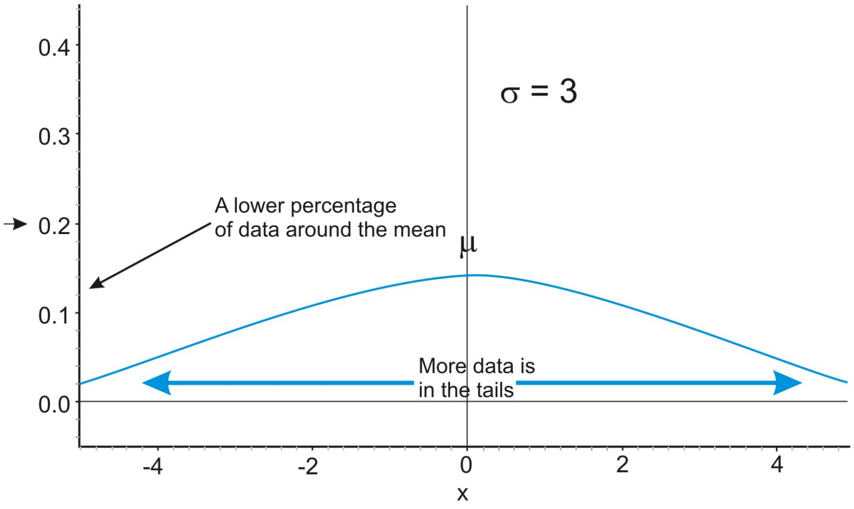
Let's go back to our popcorn example. The bag advertises a certain time, beyond which you risk burning the popcorn. From experience, the manufacturers know when most of the popcorn will stop popping, but there is still a chance that a rare kernel will pop after longer, or shorter periods of time. The 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 real, and not theoretical, there will be a time when it 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 the idealized normal distribution of a continuous random variable, the distribution continues infinitely in both directions.


Curve continues forever. Even if
it is very unlikely, there is always
a chance some data will be way
out here.
Curve gets infinitely
close to the axis without
touching it.

Because of this infinite spread, range would not be a possible statistical measure of spread. The most common way to measure the spread of a normal distribution then is using the standard deviation, or 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

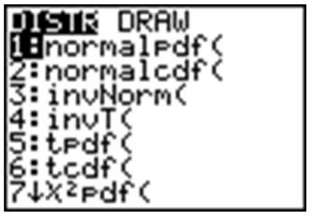


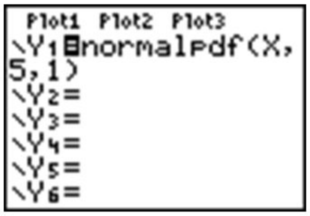
normal distribution, standard deviation = 3

The first distribution pictured above has a smaller standard deviation and so the bulk of the data is concentrated more heavily around the mean. There is less data at the extremes compared to the second distribution pictured above, which has a larger standard deviation and therefore the data is spread farther from the mean value with more of the data appearing in the tails.

Investigating the Normal Distribution on a TI-83/4 Graphing Calculator

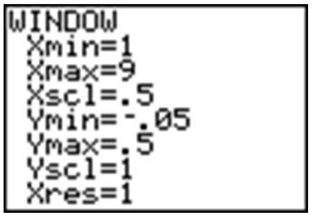
We can graph a normal curve for a probability distribution on the TI-83/4. Press [**y=**]. To create a normal distribution, we will draw an idealized curve using something called a density function. We will learn more about density functions in the next lesson. The command is called a probability density function 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. For this example, choose a mean of 5 and a standard deviation of 1.



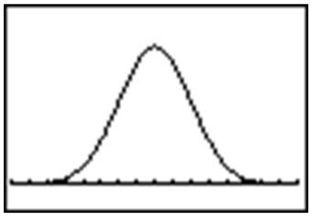


```
 Plot1  Plot2  Plot3
\Y1=normalpdf(X,
5,1)
\Y2=
\Y3=
\Y4=
\Y5=
\Y6=
```

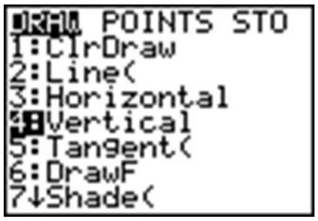
Adjust your window to match the following settings and press [**GRAPH**]

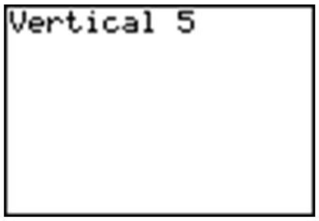


```
WINDOW
 Xmin=1
 Xmax=9
 Xscl=.5
 Ymin=-.05
 Ymax=.5
 Yscl=1
 Xres=1
```

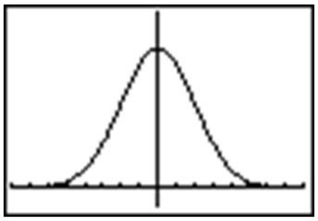


Choose [**2nd**] [**QUIT**] to go to the home screen. We can draw a vertical line at the mean to show it is in the center of the distribution by pressing [**2nd**] [**DRAW**] and choosing **VERTICAL**. Enter the mean (5) and press [**ENTER**]



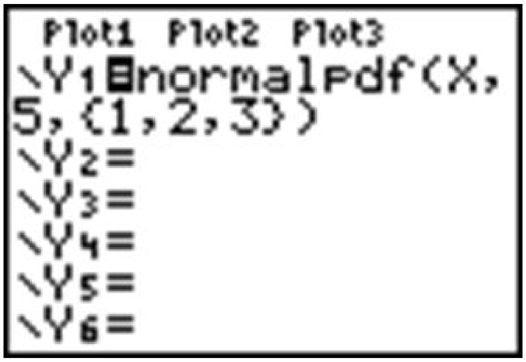


```
Vertical 5
```



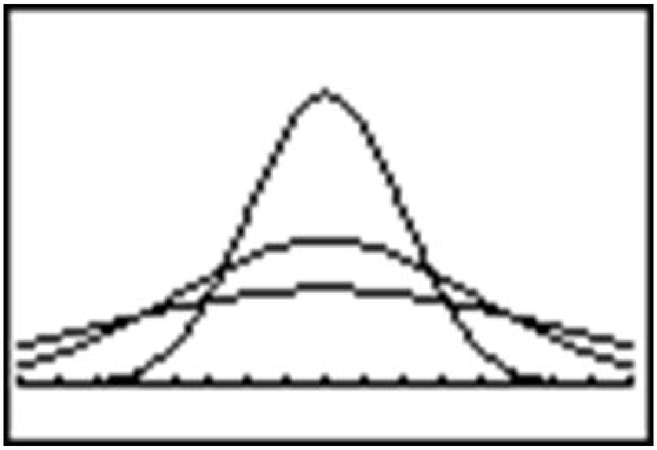
Remember that even though the graph appears to touch the $x-$axis it is actually just very close to it.

In your [**Y=**] Menu, make the following change to your normalpdf:



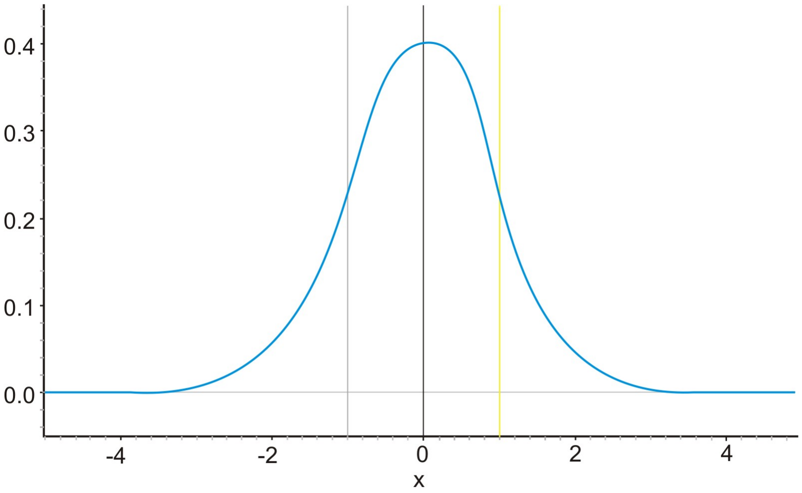
```
 Plot1 Plot2 Plot3
\Y1■normalpdf(X,
5,{1,2,3})
\Y2=
\Y3=
\Y4=
\Y5=
\Y6=
```

This will graph 3 different normal distributions with various standard deviations to make it easy to see the change in spread.



The Empirical Rule

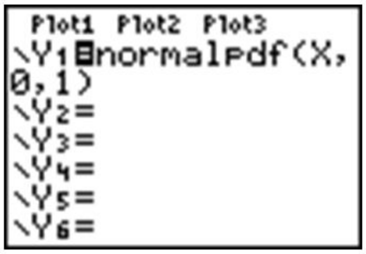
Because of the similar shape of all normal distributions we can measure the percentage of data that is a certain distance from the mean no matter what the standard deviation of the set is. The following graph shows a normal distribution with $\mu = 0$ **and** $\sigma = 1$. This curve is called a **standard normal distribution**. In this case, the values of x represent the number of standard deviations away from the mean.



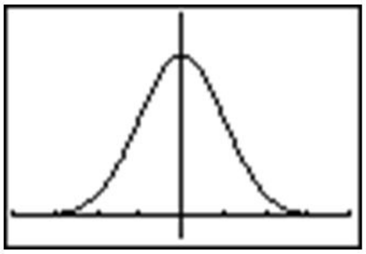
The Standard Normal Distribution

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 by the whole curve. What percentage would you estimate is between the two lines? It is a reasonable estimate to say it is about 2/3 of the total area.

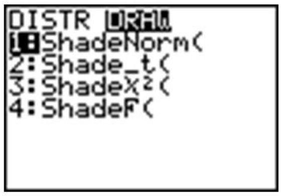
In a more advanced statistics course, you could use calculus to actually calculate this area. To help estimate the answer, we can use a graphing calculator. Graph a standard normal distribution over an appropriate window.



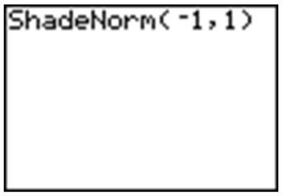
```
 Plot1  Plot2  Plot3
\Y1=normalpdf(X,
0,1)
\Y2=
\Y3=
\Y4=
\Y5=
\Y6=
```



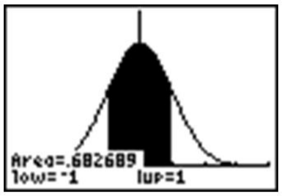
Now press [**2nd**] [**DISTR**] and choose **DRAW ShadeNorm**. Insert –1, 1 after the **ShadeNorm** command and it will shade the area within one standard deviation of the mean.



```
DISTR DRAW
1:ShadeNorm(
2:Shade_t(
3:Shadeχ²(
4:ShadeF(
```

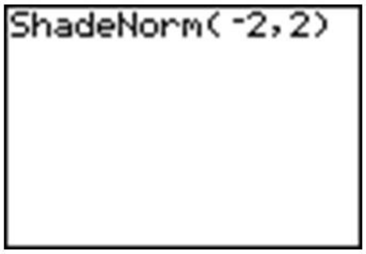


```
ShadeNorm(-1,1)
```

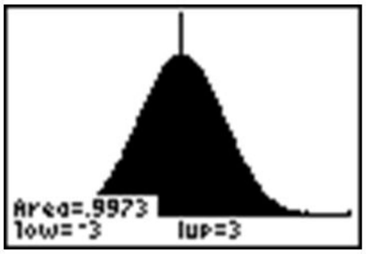


```
Area=.682689
low=-1       up=1
```

The calculator also gives a very accurate estimate of the area. We can see from this that approximately 68 percent of the area is within one standard deviation of the mean. If we venture two standard deviations away from the mean, how much of the data should we expect to capture? Make the changes to the **ShadeNorm** command to find out.

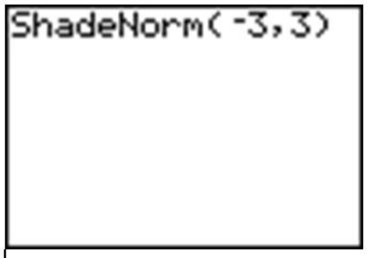


```
ShadeNorm(-2,2)
```

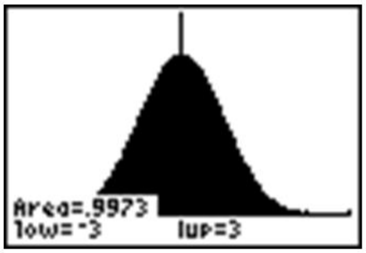


```
Area=.9973
low=-3       up=3
```

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 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.



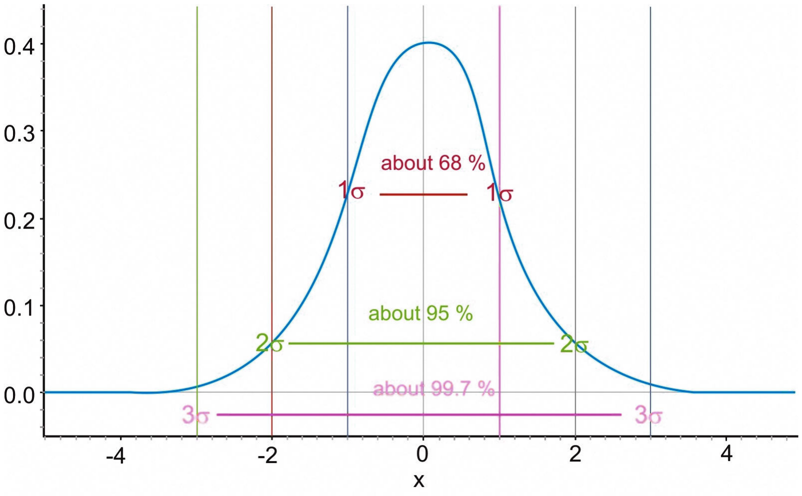
```
ShadeNorm(-3,3)
```



```
Area=.9973
low=-3       up=3
```

Notice that the shading of 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 useful for beginning statistics students and 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.



The Empirical Rule

Z-Scores

A $z-$**score** is a measure of the number of standard deviations a particular data point is away from the mean. For example, let's say the mean score on a test for your statistics class were 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 is exactly one standard deviation below the mean, and your $z-$score would be -1. To show that it is below the mean, we will assign it a $z-$score of negative one. All values that are below the mean will have negative $z-$scores. A $z-$score of negative two would represent a value that is exactly 2 standard deviations below the mean, or $82 - 14 = 68$ in this example.

To calculate a $z-$score in which the numbers are not so obvious, you take the deviation and divide it by the standard deviation.

$$z = \frac{\text{Deviation}}{\text{Standard Deviation}}$$

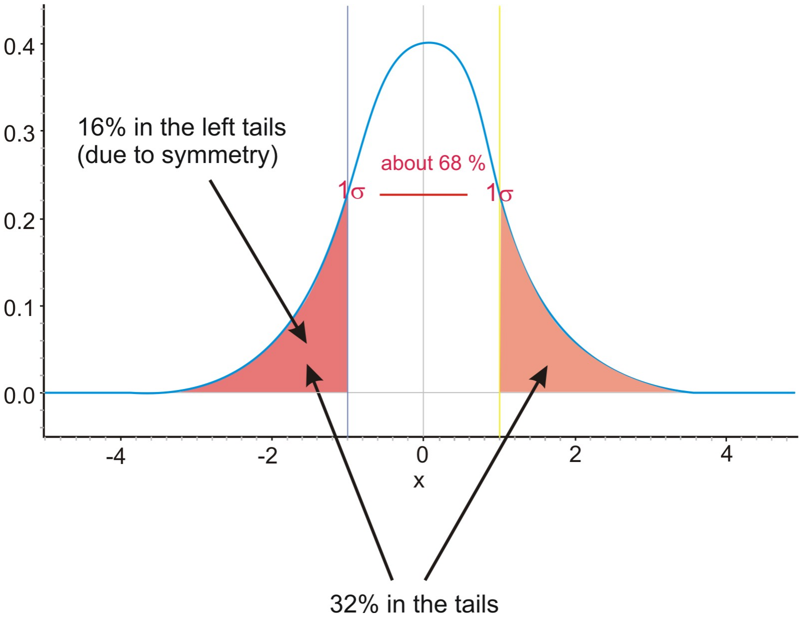
You may recall that deviation is the observed value of the variable, subtracted by the mean value, so in symbolic terms, the $z-$score would be:

$$z = \frac{x - \bar{x}}{sd}$$

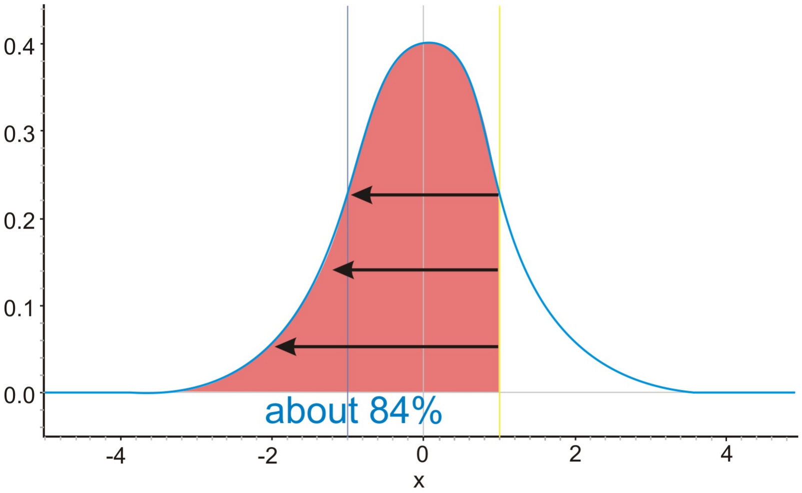
Ex. What is the $z-$score for an A on this test? (assume that an A is a 93).

$$z = \frac{x - \bar{x}}{sd}$$
$$z = \frac{93 - 82}{7}$$
$$z = \frac{11}{7} \approx 1.57$$

It is not necessary to have a normal distribution to calculate a $z-$score, but the $z-$score has much more significance when it relates to a normal distribution. For example, 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 percent of the students would have scored between a $z-$score of –1 and 1, or between a 75 and an 89. If 68% of the data is between those two values, then that leaves a remaining 32% in the tail areas. Because of symmetry, that leaves 16% in each individual tail.



If we combine the two percentages, approximately 84% of the data is below an 89 score. We typically refer to this as a **percentile**. A student with this score could conclude that he or she performed better than 84% of the class, and that he or she was in the 84^{th} percentile.

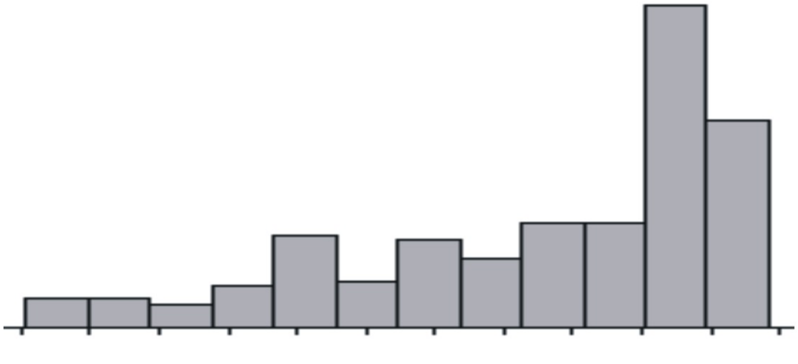


84th percentile corresponds to a z-score of 1

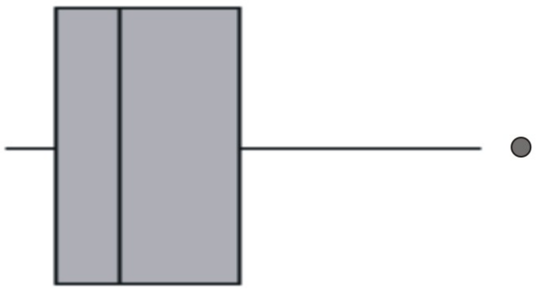
This same conclusion can be put in terms of a probability distribution as well. We could say that if a student from this class were chosen at random the probability that we would choose a student with a score of 89 or less is .84, or there is an 84% chance of picking such a student.

Assessing Normality

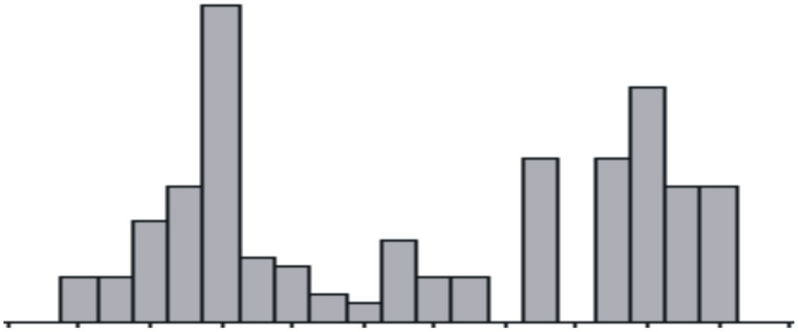
The best way to determine if a data set approximates a normal distribution is to look at a visual representation. Histograms and box plots can be useful indicators of normality, but are not always definitive. It is often easier to tell if a data set is *not* normal from these plots.



Skewed left distribution



Skewed right distribution with outliers

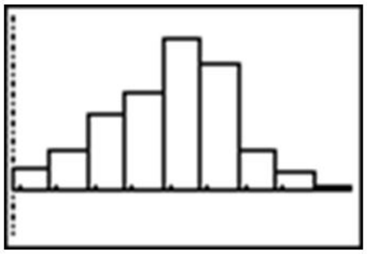


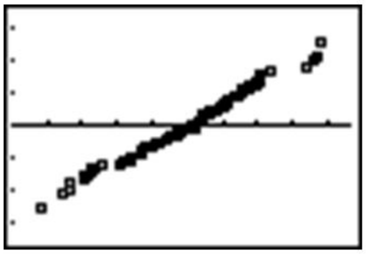
Bimodal Distribution

If a data set is **skewed right** it means that the right tail is significantly larger than the left. Likewise, **skewed left** means the left tail has more weight than the right. A **bimodal distribution** has two modes, or peaks, as if two normal distributions were added together. Multimodal distributions with two or more modes often reflect two different types. For instance, a histogram of the heights of American 30-year-old adults, you will see a bimodal distribution – one mode for males, one mode for females.

Now that we know how to calculate $z-$scores, there is a plot we can use to determine if a distribution is normal. If we calculate the $z-$scores for a data set and plot them against the actual values, this is called a **normal probability plot**, or a **normal quantile plot**. If the data set is normal, then this plot will be perfectly linear. The closer to being linear the normal probability plot is, the more closely the data set approximates a normal distribution.

Look below at a histogram and the normal probability plot for the same data.





The histogram is fairly symmetric and mound-shaped and appears to display the characteristics of a normal distribution. When the $z-$scores are plotted against the data values, the normal probability plot appears strongly linear, indicating that the data set closely approximates a normal distribution.

Example:

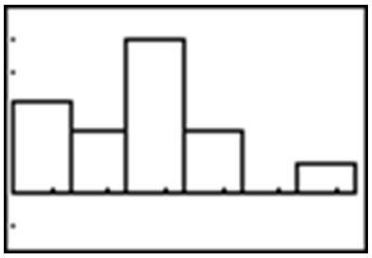
The following data set tracked high school seniors' involvement in traffic accidents. The participants were asked the following question: "During the last 12 months, how many accidents have you had while you were driving (whether or not you were responsible)?"

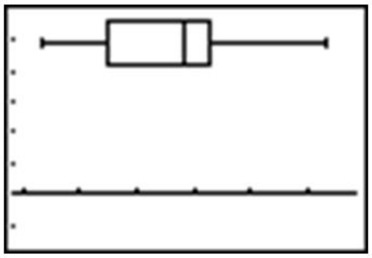
Table 5.1:

Year	Percentage of high school seniors who said they were involved in no traffic accidents
1991	75.7
1992	76.9
1993	76.1
1994	75.7
1995	75.3
1996	74.1
1997	74.4
1998	74.4
1999	75.1
2000	75.1
2001	75.5
2002	75.5
2003	75.8

Figure: Percentage of high school seniors who said they were involved in no traffic accidents. *Source:* Sourcebook of Criminal Justice Statistics: http://www.albany.edu/sourcebook/pdf/t352.pdf

Here is a histogram and a box plot of this data.





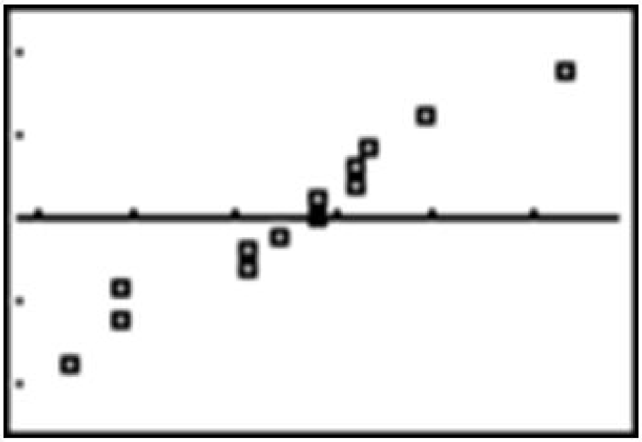
The histogram appears to show a roughly mound-shaped and symmetric distribution. The box plot does not appear to be significantly skewed, but the various sections of the plot also do not appear to be overly symmetric either. In the following chart the $z-$scores for this data set have been calculated. The mean percentage is approximately 75.35

Table 5.2:

Year	Percentage	$z-$score
1991	75.7	.45
1992	76.9	2.03
1993	76.1	.98
1994	75.7	.45
1995	75.3	$-.07$
1996	74.1	-1.65
1997	74.4	-1.25
1998	74.4	-1.25
1999	75.1	$-.33$
2000	75.1	$-.33$
2001	75.5	.19
2002	75.5	.19
2003	75.8	.59

Figure: Table of $z-$scores for senior no-accident data.

Here is a plot of the percentages and the $z-$scores, or the normal probability plot.



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

One additional clue about normality might be gained from investigating the empirical rule. Remember than in an idealized normal curve, approximately 68% of the data should be within one standard deviation of the mean. If we count, there are 9 years for which the $z-$scores are between -1 and 1. As a percentage of the total data, 9/13 is about 69%, or very close to the ideal value. This data set is so small that it is difficult to verify the other percentages, but they are still not unreasonable. About 92% of the data (all but one of the points) ends up within 2 standard deviations of the mean, and all of the data (Which is in line with the theoretical 99.7%) is located between $z-$scores of -3 and 3.

Lesson Summary

A **normal distribution** is a perfectly symmetric, mound-shaped distribution that appears in many practical and real data sets and is an especially important foundation for making conclusions about data called inference. A **standard normal distribution** is a normal distribution in 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:

$$z = \frac{x - \bar{x}}{sd}$$

$Z-$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. A distribution has $z-$scores regardless of whether or not it is normal in shape. If the distribution is normal, however, the $z-$scores are useful in explaining how much of the data is contained within a certain distance of the mean. The **empirical rule** is the name given to the observation that approximately 68% of the data is within 1 standard deviation of the mean, about 95% is within 2 standard deviations of the mean, and 99.7% of the data is within 3 standard deviations of the mean. Some refer to this as the $68 - 95 - 99.7$.

There is no straight-forward test for normality. You should learn to recognize the normality of a distribution by examining the shape and symmetry of its visual display. However, a **normal probability** or **normal quantile plot** is a useful tool to help check the normality of a distribution. This graph is a plot of the $z-$scores of a data set against the actual values. If the distribution is normal, this plot will be linear.

Points To Consider

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

Review Questions

1. Which of the following data sets is most likely to be normally distributed? For the other choices, explain why you believe they would not follow a normal distribution.
   (a) The hand span (measured from the tip of the thumb to the tip of the extended 5^{th} 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.
   (a) 65
   (b) 83
   (c) 93
   (d) 100
3. Assume that the mean weight of 1 year-old girls in the US is normally distributed with a mean of about 9.5 grams with a standard deviation of approximately 1.1 grams. Without using a calculator, estimate the percentage of 1 year-old girls in the US that

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 $\mu = 2.80$ and $\sigma = 1.34$[1]. What is the approximate $z-$score that corresponds to an exam score of 5 (The scores range from $1 - 5$).

(a) 0.786
(b) 1.46
(c) 1.64
(d) 2.20
(e) A $z-$score can not 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 United States 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:

$$241, 282, 258, 224, 133, 335, 322, 323, 354, 194, 332, 274, 233, 147, 213, 262, 227, 366$$

(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:

$\mu =$ ________________ $\sigma =$ ________________

Table 5.3:

Number of Sprinkles	Deviation	$Z-$scores
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.

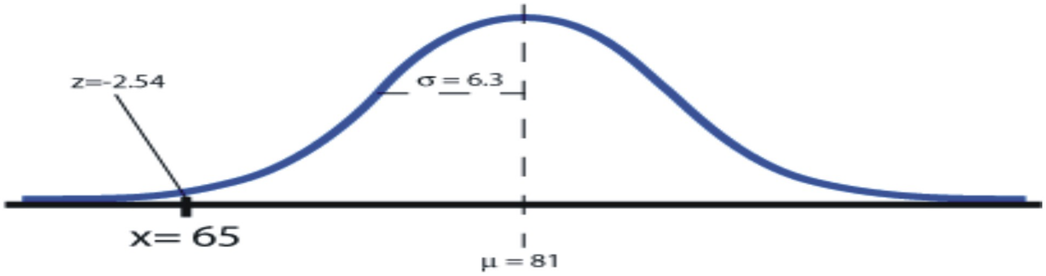
(d) Based on this plot, comment on the normality of the distribution of sprinkle counts on these doughnuts.

Open-ended Investigation: Munchkin Lab.

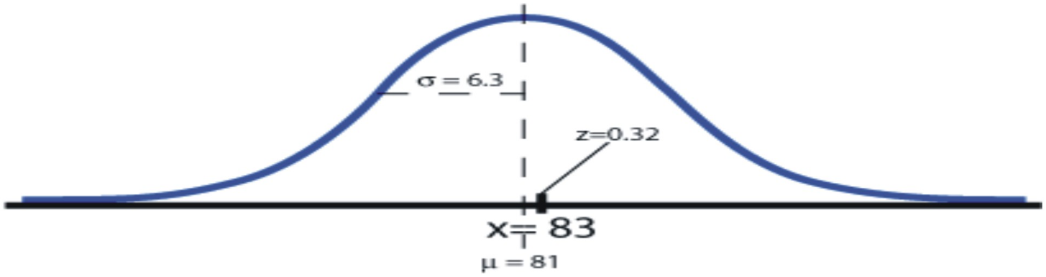
Teacher Notes: For this activity, obtain two large boxes of Dunkin Donuts' munchkins. Each box should contain only one type of munchkin. I have found students prefer the glazed and the chocolate, but the activity can be modified according to your preference. If you do not have Dunkin Donuts near you, the bakery section of your supermarket should have boxed donut holes or something similar you can use. You will also need an electronic balance capable of measuring to the nearest 10^{th} of a gram. Your science teachers will be able to help you out with this if you do not have one. I have used this activity before introducing the concepts in this chapter. If you remove the words "$z-$score", the normal probability plot and the last two questions, students will be able to investigate and develop an intuitive understanding for standardized scores and the empirical rule, before defining them. Experience has shown that this data very closely approximates a normal distribution and students will be able to calculate the $z-$scores and verify that their results come very close to the theoretical values of the empirical rule.

Review Answers

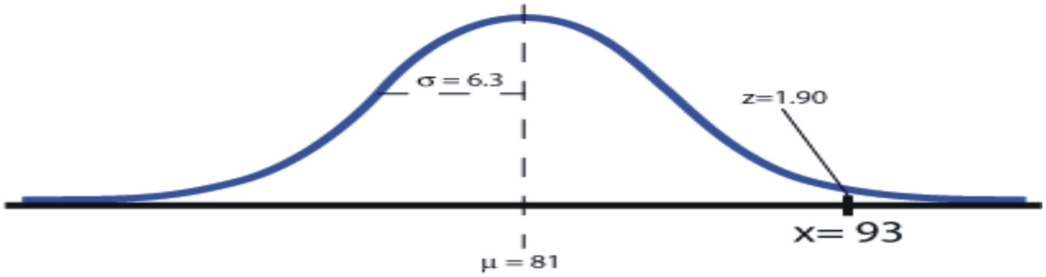
1. (a) You would expect this situation to vary normally with most students' hand spans centering around a particular value and a few students having much larger or much smaller hand spans.
   (b) Most employees could be hourly laborers and drivers and their salaries might be normally distributed, but the few management and corporate salaries would most likely be much higher, giving a skewed right distribution.
   (c) Many studies have been published detailing the shrinking, but still prevalent income gap between male and female workers. This distribution would most likely be bi-modal, with each gender distribution by itself possibly being normal.
   (d) You might expect most of the pennies to be this year or last year, fewer still in the previous few years, and the occasional penny that is even older. The distribution would most likely be skewed left.
2. (a) $z \approx -2.54$



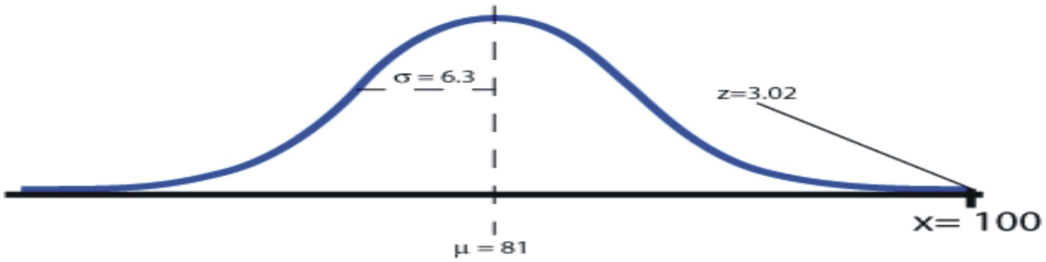
   (b) $z \approx 0.32$



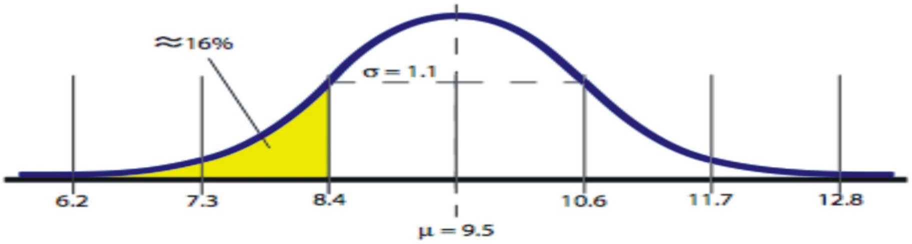
   (c) $z \approx 1.90$



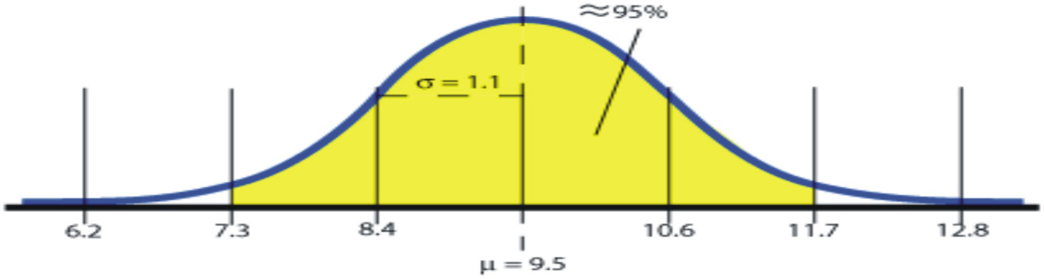
   (d) $z \approx 3.02$



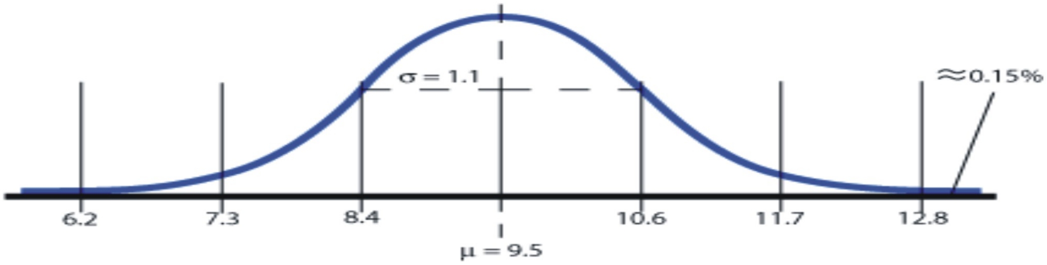
3. Because the data is normally distributed, students should use the $68 - 95 - 99.7$ rule to answer these questions.
   (a) about 16% (less than one standard deviation below the mean)



   (b) about 95% (within 2 standard deviations)



   (c) about 0.15% (more than 3 standard deviations above the mean)

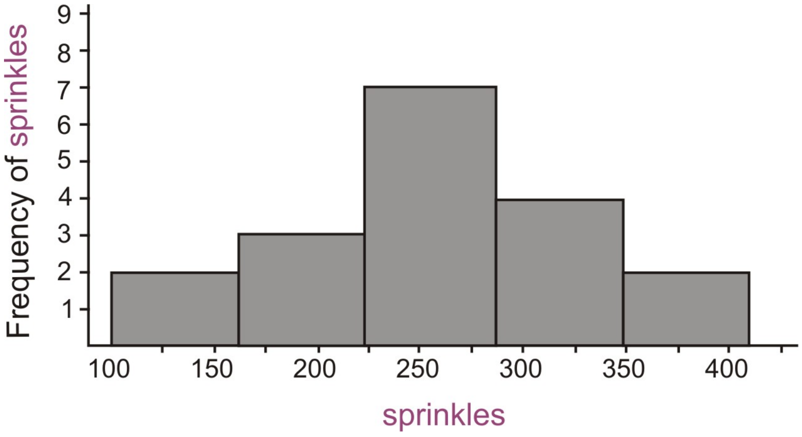


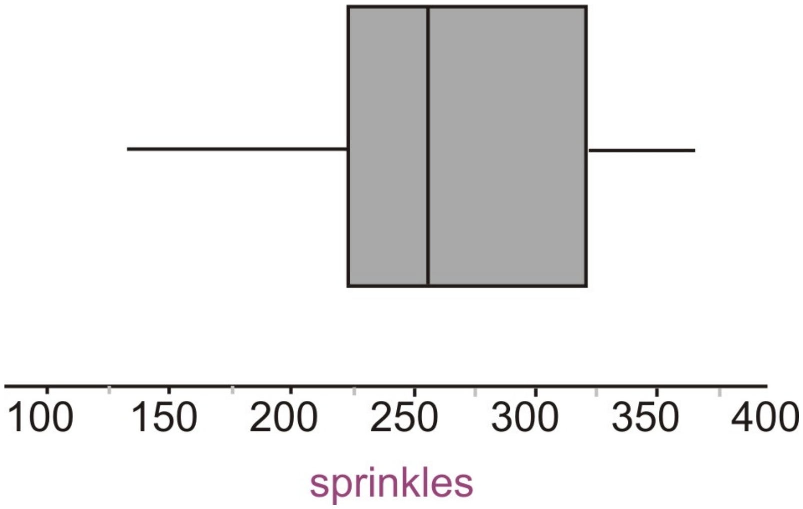
4. The standard normal curve has a mean of zero and a standard deviation of one, so all the values correspond to $z-$scores. The corresponding values are approximately:
   (a) 0.84

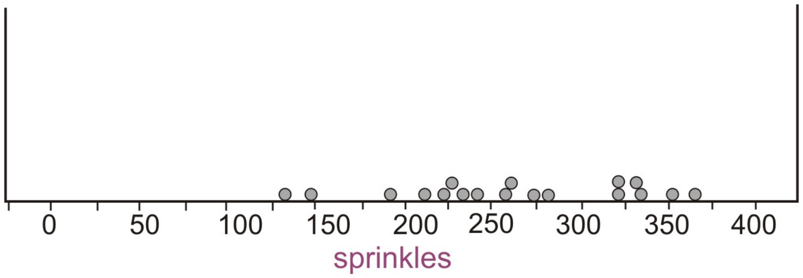
(b) 0.16
(c) 0
(d) 1
(e) 0.025

Therefore the correct order is: c, e, b, a, d

5. c
6. 0.025.157.7 is exactly 2 standard deviations above the mean height. According to the empirical rule, the probability of a randomly chosen value being within 2 standard deviations is about 0.95, which leaves 0.05 in the tails. We are interested in the upper tail only as we are looking for the probability of being above this value.
7. (a) Here are the possible plots showing a symmetric, mound shaped distribution.





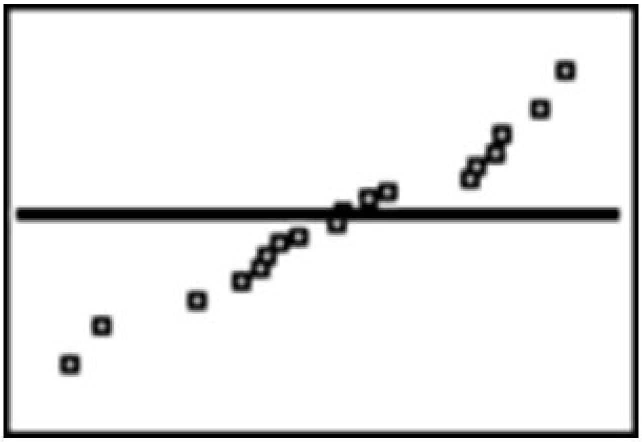


(b) $\mu = 262.222$ $s = 67.837$

Table 5.4:

Number of Sprinkles	Deviations	$Z-$scores
241	-21.2222	-0.313
282	19.7778	0.292
258	-4.22222	-0.062
223	-38.2222	-0.563
133	-129.222	-1.905
335	72.7778	1.073
322	59.7778	0.881
323	60.7778	0.896
354	91.7778	1.353
194	-68.2222	-1.006
332	69.7778	1.029
274	11.7778	0.174
233	-29.2222	-0.431
147	-115.222	-1.699
213	-49.2222	-0.726
262	-0.222222	-0.003
227	-35.2222	-0.519
366	103.778	1.530

(c)



(d) The normal probability plot shows a fairly linear pattern which is an indication that there are no obvious departures from normality in this distribution.

References

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

5.2 The Density Curve of the Normal Distribution

Learning Objectives

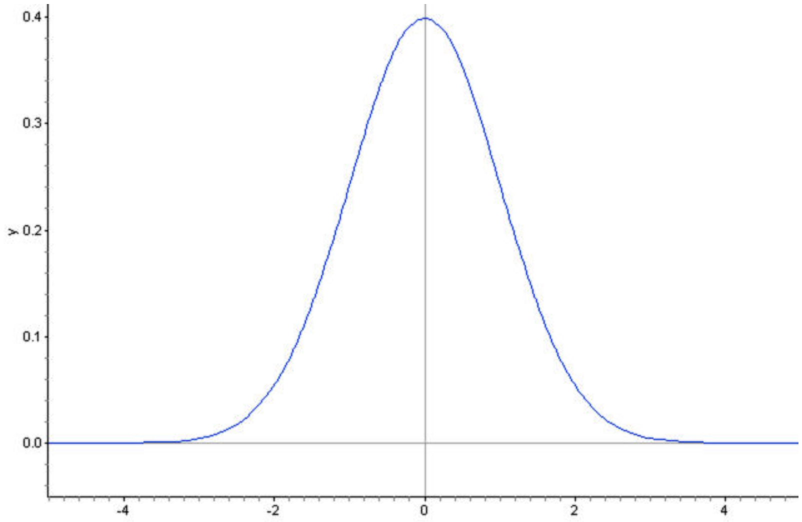
- Identify the properties of a normal density curve, and the relationship between concavity and standard deviation.
- Convert between $z-$scores and areas under a normal probability curve.
- Calculate probabilities that correspond to left, right, and middle areas from a left-tail $z-$score table.
- Calculate 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.

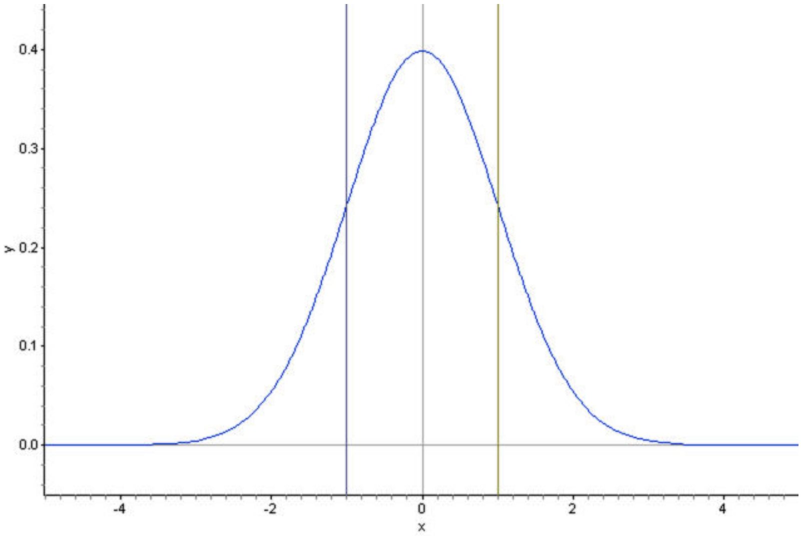
Density Curves

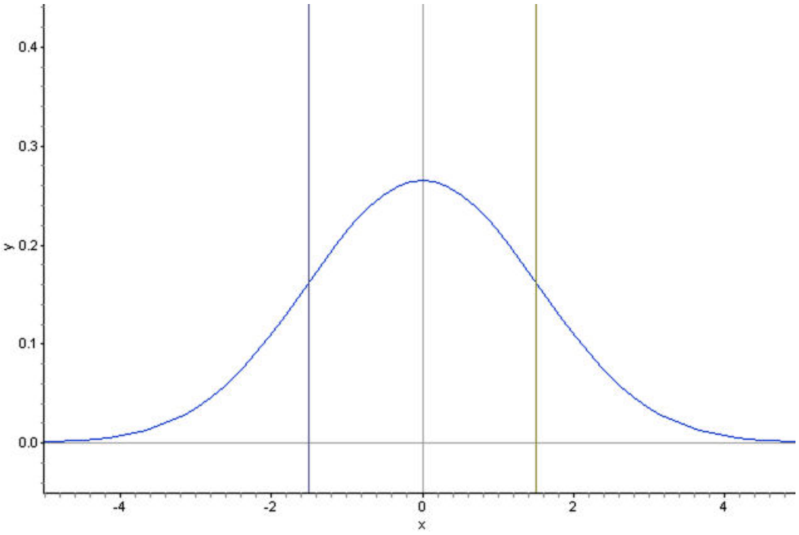
A **density curve** is an idealized representation of a distribution in which the area under the curve is defined to be 1. Density curves need not be normal, but the **normal density curve** will be the most useful to us.

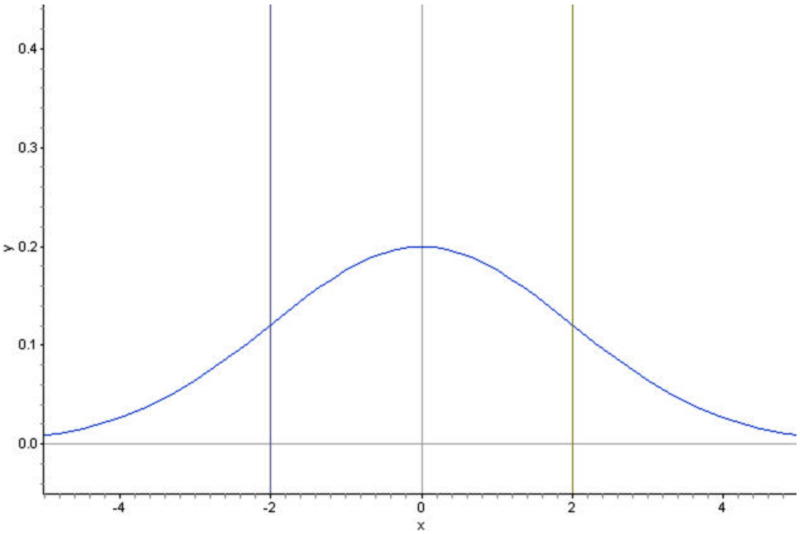


Inflection Points on a Normal Density Curve

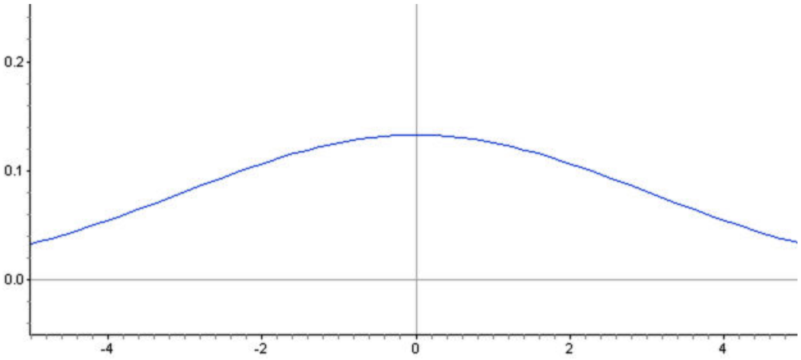
We already know from the empirical rule, that approximately 2/3 of the data in a normal distribution lies within 1 standard deviation of the mean. In a density curve, this means that about 68% of the total area under the curve is within $z-$scores of ± 1. Look at the following three density curves:





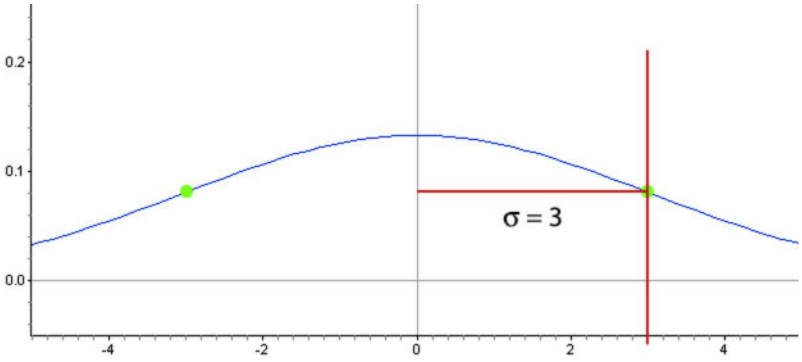


Notice that the curves are spread increasingly wider. Lines have been drawn to show the points one standard deviation on either side of the mean. Look at *where* this happens on each density curve. Here is a normal distribution with an even larger standard deviation.



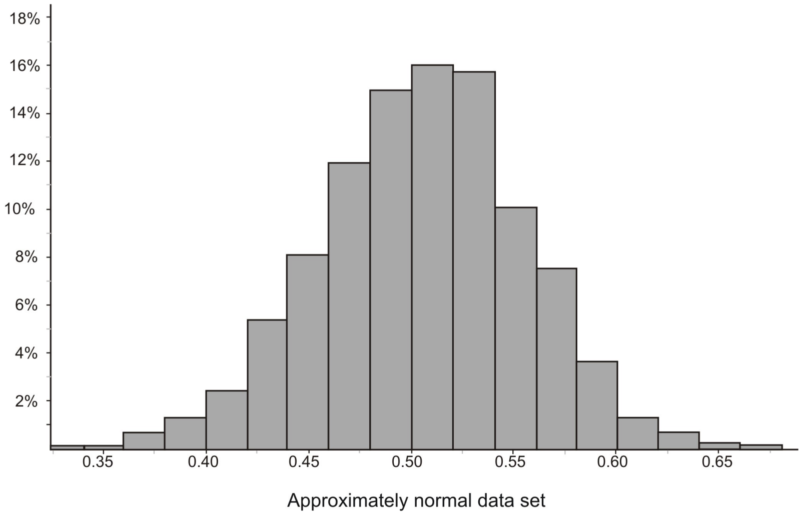
Could you predict the standard deviation of this distribution from estimating the point on the density curve?

You may notice that the density curve changes shape at this point in each of our examples. In Calculus, we learn to call this shape changing location an **inflection point**. It is the point 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 this point, the curve is concave up (it looks like a valley or ”u” shape). We will leave it to the Calculus students to prove it, but in a normal density curve, this inflection point is always exactly one standard deviation away from the mean.

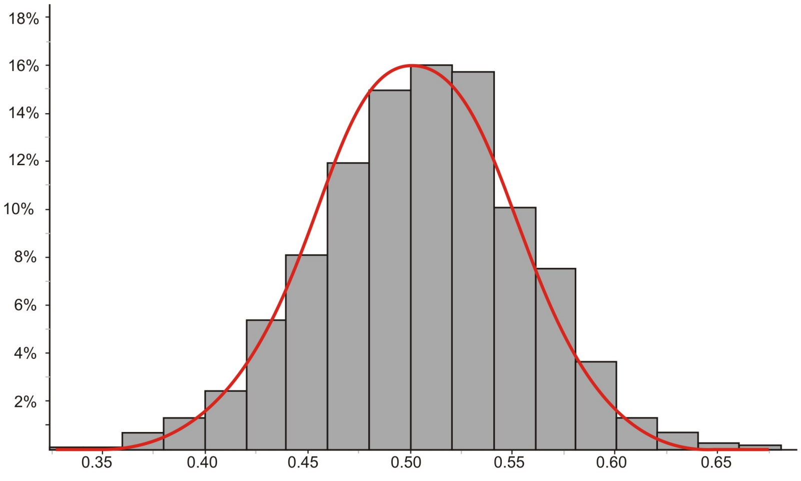


In this example, the standard deviation was 3 units. We can use these concepts to estimate the standard deviation of a normally distributed data set.

Can you estimate the standard deviation of the distribution represented by the following histogram?

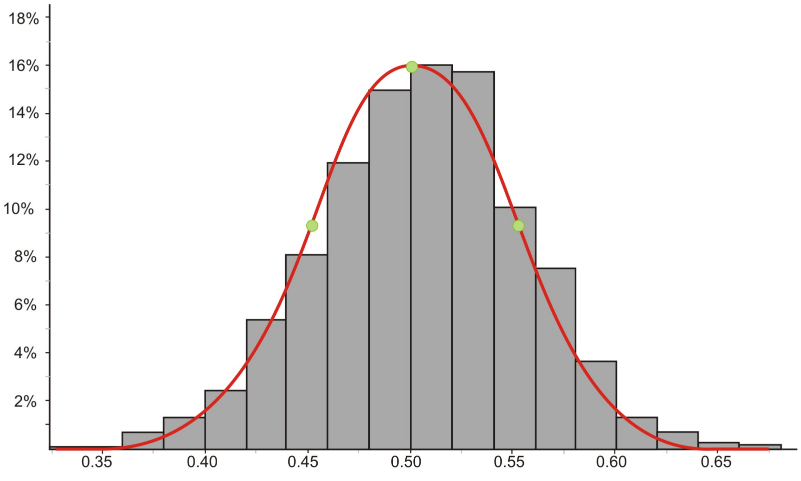


This distribution is fairly normal, so we could draw a density curve to approximate it as follows.



Density Curve approximating histogram

Now estimate the inflection points:



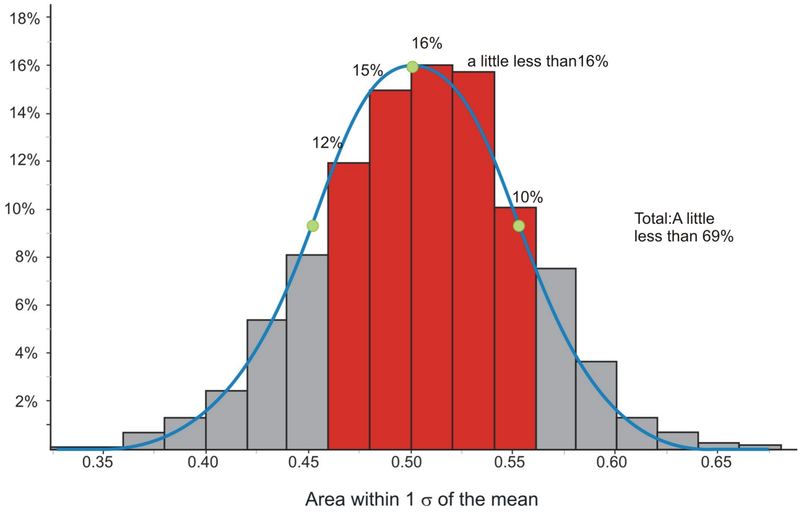
Density Curve with Inflection Points

It appears that the mean is about 0.5 and the inflection points are 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:

$$s \approx 0.04988$$
$$\bar{x} \approx 0.04997$$

We can verify this using expectations from the empirical rule. In the following graph, we have highlighted the bins that are contained within one standard deviation of the mean.



If you estimate the relative frequencies from each bin, they total remarkably close to 68%!

Calculating Density Curve Areas

While it is convenient to estimate areas using the empirical rule, we need more precise methods to calculate the areas for other values. In Calculus you study methods for calculating the area under a curve, but in statistics, we are not so concerned about the specific method used to calculate these areas. We will use formulas or technology to do the calculations for us.

Z-Tables

Before software and graphing calculator technology was readily available, it was common to use tables to approximate the amount of area under a normal density curve between any two given $z-$scores. We have included two commonly used tables at the end of this lesson. Here are a few things you should know about reading these tables:

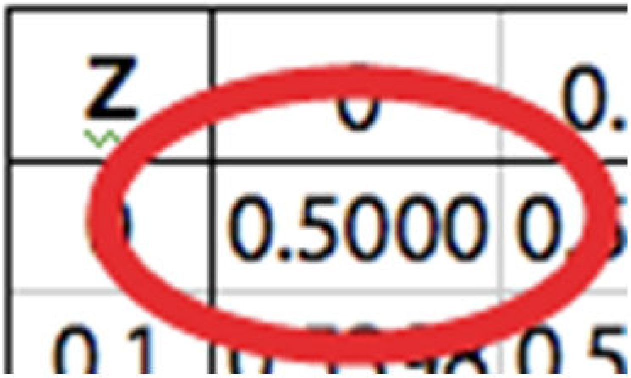
The values in these tables are all in terms of $z-$scores, or **standardized**, meaning that they correspond to a standard normal curve in which the mean is 0 and the standard deviation is 1. It is important to understand that the table shows the areas **below** the given $z-$score in the table. It is possible and often necessary to calculate the area **above**, or **between** $z-$scores as well. You could generate new tables to show these values, but it is just as easy to calculate them from the one table.

The values in these tables can represent areas under the density curve. For example, .500 means half of the area (because the area of the total density curve is 1). However, they are most frequently expressed as probabilities, e.g. .500 means the probability of a randomly chosen value from this distribution being in that region is .5, or a 50% chance.

$Z-$scores must be rounded to the nearest hundredth to use the table.

Most $z-$score tables do not go much beyond 3 standard deviations away from the mean in either direction because as you know, the probability of experiencing results that extreme in a normal distribution is very low.

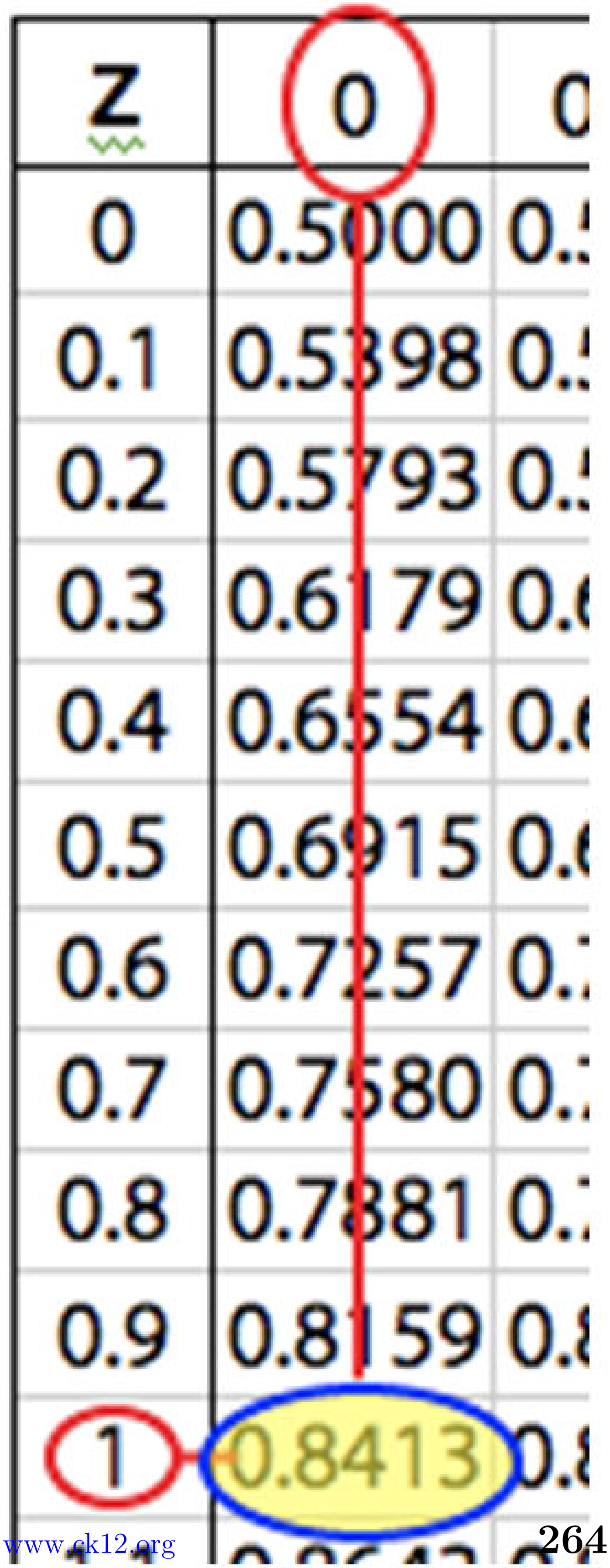
Table 5.5 shows those below the mean and Table 5.6 shows values of $z-$scores that are to the right of the mean. To help you understand how to read the table, look at the top left entry of Table 5.6. It reads .500.



Z	0	0.
0	0.5000	0.5
0.1	[illegible]	0.5

Think of the table as a stem and leaf plot with the stem of the $z-$scores running down the left side of the table and the leaves across the top. The leaves represent 100ths of a $z-$score. So, this value represents a $z-$score of 0.00. This should make sense because we are talking about the actual mean.

Let's look at another common value. In Table 5.6 find the $z-$score of 1 and read the associated probability value.

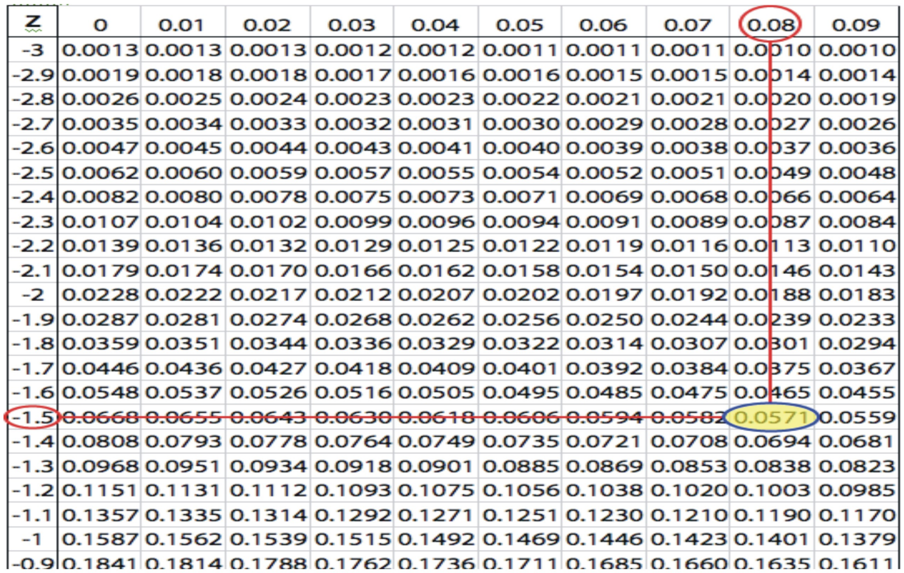


Z	0
0	0.5000
0.1	0.5398
0.2	0.5793
0.3	0.6179
0.4	0.6554
0.5	0.6915
0.6	0.7257
0.7	0.7580
0.8	0.7881
0.9	0.8159
1	0.8413

As we have already discovered, approximately 84% of the data is below this value (68% in the middle, and 16% in the tail). This corresponds to the probability in the table of .8413.

Now find the probability for a $z-$score of -1.58. It is often a good idea to estimate this value before using the table when you are first getting started. This $z-$score is between -2 and -1. We know from the empirical rule that the probability for $z = -1$ is approximately .16 and similarly, for -2 it is around .025, so we should expect to get a value somewhere between these two estimates.

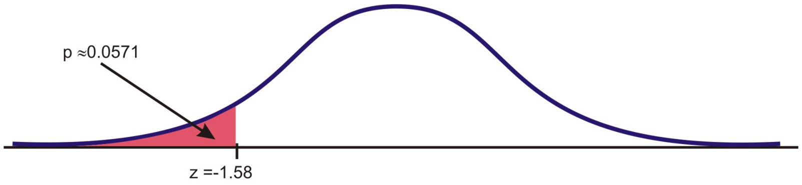
Locate the stem and the leaf for -1.58 on Table 5.5 and follow them across and down to the corresponding probability. The answer appears to be approximately 0.0571, or approximately 5.7% of the data in a standard normal curve is below a $z-$score of -1.58.



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

z-table value for z≤ -1.58

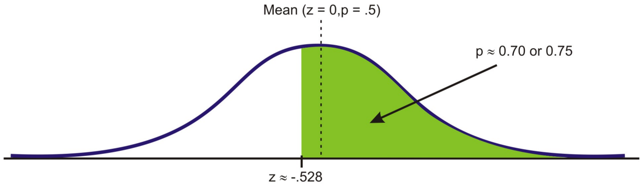
It is extremely important, especially when you first start with these calculations, that you get in the habit of relating it to the normal distribution by drawing a sketch of the situation. In this case, simply draw a sketch of a standard normal curve with the appropriate region shaded and labeled.



Sketch for z ≤ -1.58

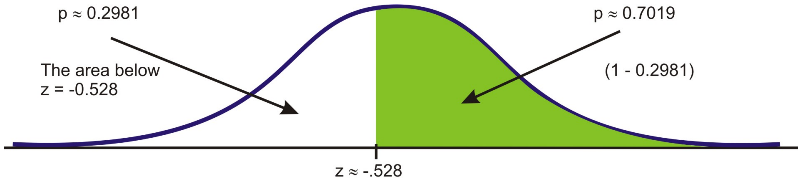
Let's try an example in which we want to find the probability of choosing a value that is **greater than** $z = -0.528$. Before even using the table, draw a sketch and estimate the probability. This $z-$score is just below the mean, so the answer should be more than 0.5.

The $z-$score of –0.5 would be half way between 0 and -1, but because there is more area concentrated around the mean, we could guess that there should be more than half of the 34% of the area in this section. If we were to guess about $20 - 25\%$, we would estimate an answer of between 0.70 and 0.75.



Estimated Sketch for z ≥ -0.528

First 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. This will slightly under-estimate the probability, but it is the best we can do using the table. The table returns a value of 0.2981 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 .7019.

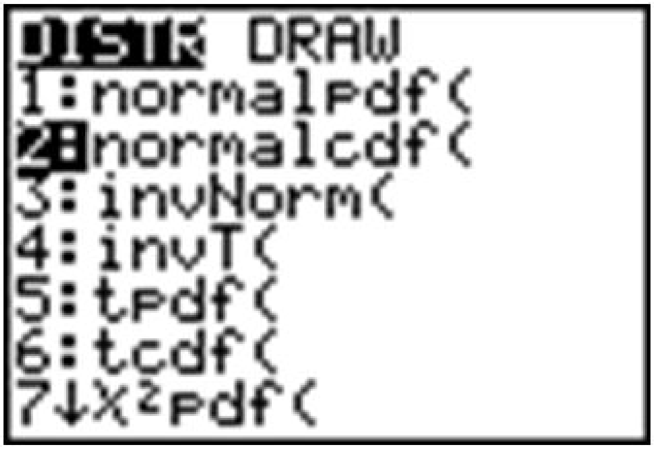


Final Sketch for z ≥ -0.528

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 that we will leave this for one of the homework exercises.

Using Graphing Calculators: The Normal CDF Command.

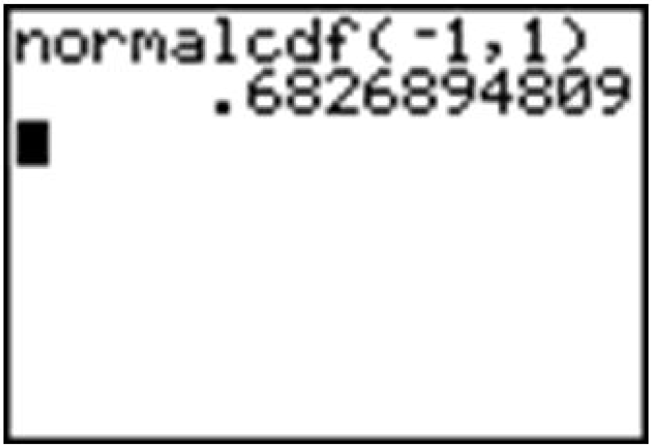
Your graphing calculator has already been programmed to calculate probabilities for a normal density curve using what is called a **cumulative density function** or **cdf**. This is found in the distributions menu above the VARS key.



Press **[2nd]** **[VARS]**, **[2]** to select the **normalcdf** (command. **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 68%.



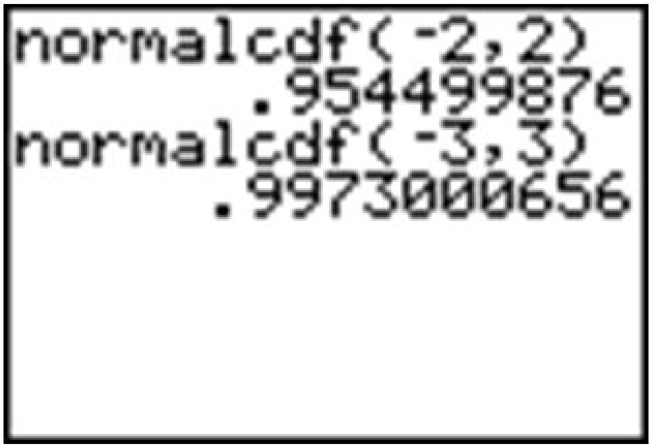
```
normalcdf(-1,1)
        .6826894809
```

Try to verify the other values from the empirical rule.

Summary:

Normalpdf $(x, 0, 1)$ gives values of the **probability density function**. It gives the value of the probability (vertical distance to the graph) at any value of x. This is the function we graphed in Lesson 5.1

Normalcdf $(a, b, 0, 1)$ gives values of the **cumulative density function**. It gives the probability of an event occurring between $x = a$ and $x = b$ (area under the probability density function curve and between two vertical lines).



```
normalcdf(-2,2)
        .954499876
normalcdf(-3,3)
       .9973000656
```

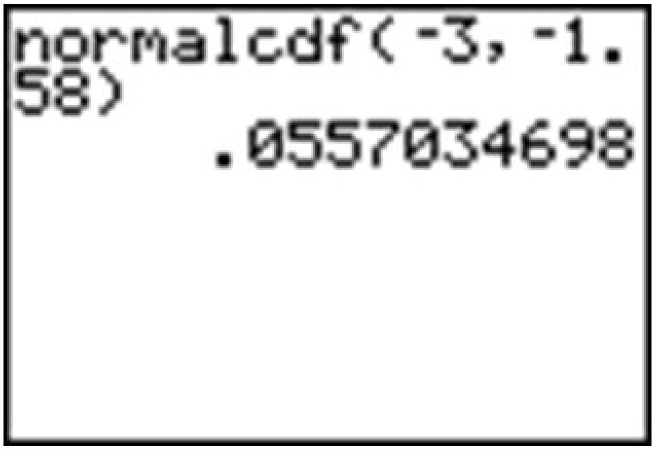
Let's look at the two examples we did in the last section using the table.

Example:

Find the probability for $x < -1.58$.

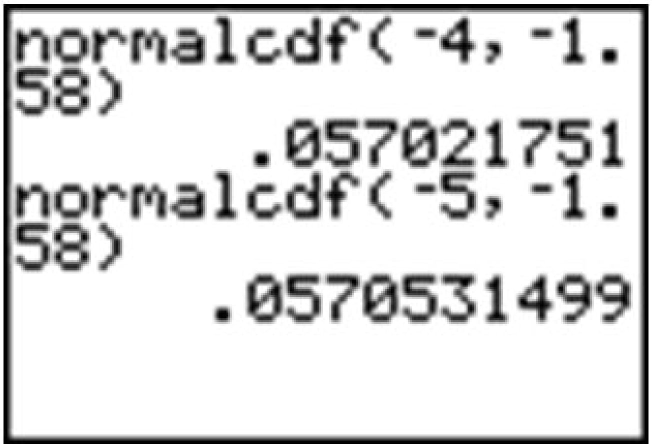
Solution:

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.



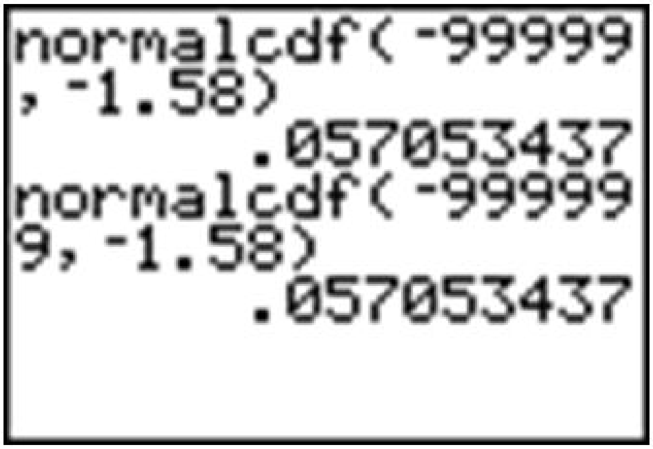
```
normalcdf(-3,-1.
58)
       .0557034698
```

The answer is accurate to the nearest 1%, but remember that there really still is some data, no matter how little, that we are leaving out if we stop at –3. In fact, if you look at Table 1, you will see that about 0.0013 has been left out. Try going out to -4 and -5.



```
normalcdf(⁻4,⁻1.
58)
      .057021751
normalcdf(⁻5,⁻1.
58)
     .0570531499
```

Notice that if we use -5, the answer is as accurate as the one in the table. 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.



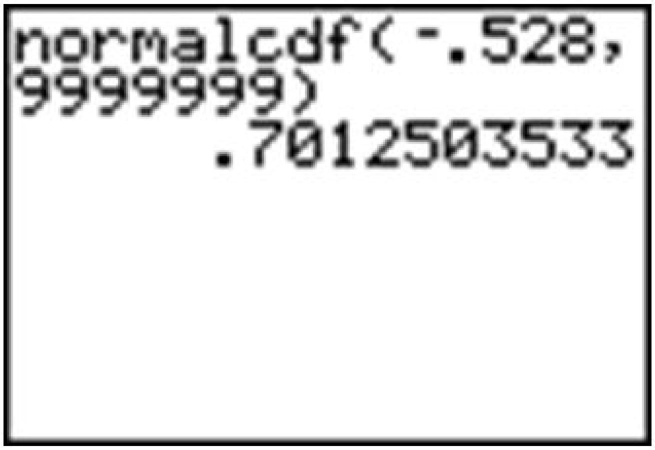
```
normalcdf(-99999
,⁻1.58)
       .057053437
normalcdf(-99999
9,⁻1.58)
       .057053437
```

Example:

Find the probability for $x \geq -0.528$.

Solution:

Right away we are at an advantage using the calculator because we do not have to round off the $z-$score. Enter a **normalcdf** command from -0.528 to "bunches of nines". This upper bound represents a ridiculously large upper bound that would insure a probability of missing data being so small that it is virtually undetectable.



```
normalcdf(-.528,
9999999)
          .7012503533
```

Remember that our answer from the table was slightly too small, so when we subtracted it from 1, it became too large. The calculator answer of about .70125 is a more accurate approximation than the table value.

Standardizing

In most practical problems involving normal distributions, the curve will not be standardized ($\mu = 0$ and $\sigma = 1$). When using a $z-$table, you will have to first 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?

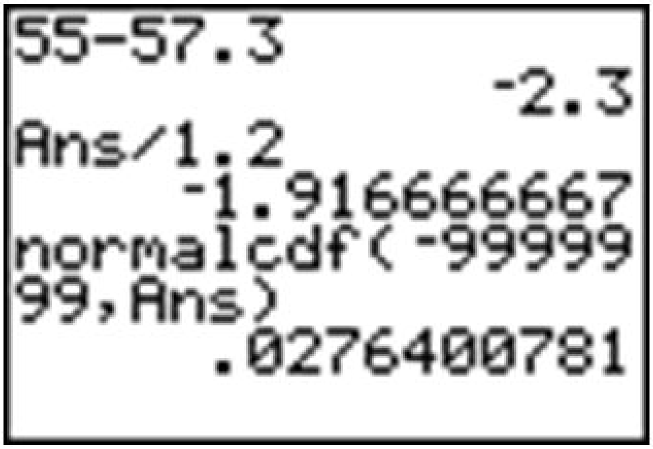
Solution:

Calculate the $z-$score for 55.

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

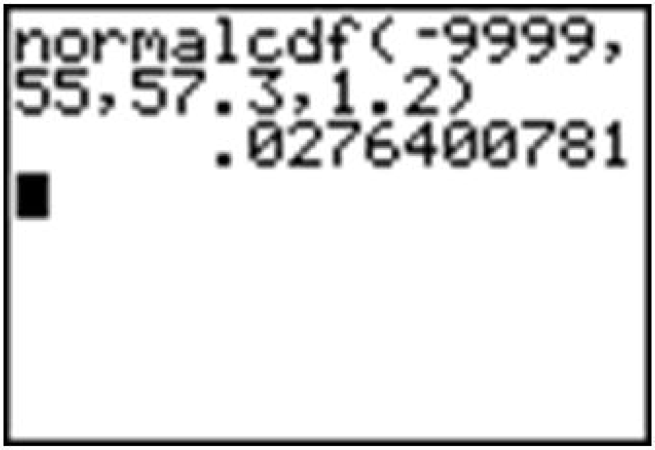
Using Table 5.5, the probability of experiencing a value this low is approximately 0.0274. In other words, there is about a 3% chance that you would get a bag of candy with 55 or fewer pieces, so Endy should feel cheated.

Using the graphing calculator, the results would look as follows (the ANS function has been used to avoid rounding off the $z-$score):



```
55-57.3
              -2.3
Ans/1.2
      -1.916666667
normalcdf(-99999
99,Ans)
       .0276400781
```

However, the advantage of using the 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.



```
normalcdf(-9999,
55,57.3,1.2)
       .0276400781
```

Lesson Summary

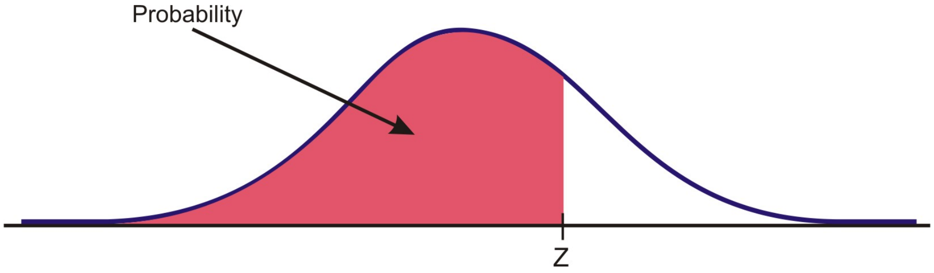
A **density curve** is an idealized representation of a distribution in which the area under the curve is defined as 1, or in terms of percentages, 100% of the data. A **normal density curve** is simply a density curve for a normal distribution. Normal density curves have two **inflection points**, which are the points on the curve where it changes concavity. Remarkably, 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 using a $z-$**table** or using the **normalcdf** command on the TI-83/84. The $z-$table provides the area less than a particular $z-$score for the standard normal density curve. The calculator command allows you to specify two values, either standardized or not, and will calculate the area between those values.

Points To consider

1. How do we calculate the areas/probabilities for distributions that are not normal?
2. How do we calculate the $z-$scores, mean, standard deviation, or actual value given the probability or area?

Tables

There are two tables here, Table 1 for $z-$scores less than 0 and one and Table 2 for $z-$scores greater than 0. The table entry for z is the probability of lying below z. Essentially, these tables list the area of the shaded region in the figure below for each value of z.



For example, to look up $P(z < -2.68) = 0.0037$ in the first table, find -2.6 in the left hand column, then read across that row until you reach the value in the hundredths place (8) to read off the value.

Using this same technique and the second table, you should find that $P(z < 1.42) = 0.92$.

Table 5.5: **Table of Standard Normal Probabilities for z < 0**

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019

Table 5.5: (continued)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table 5.6: **Table of Standard Normal Probabilities for z > 0**

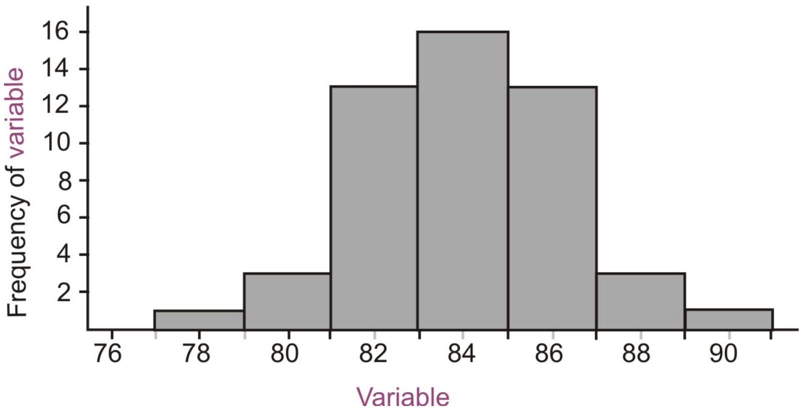
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549

Table 5.6: (continued)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Review Questions

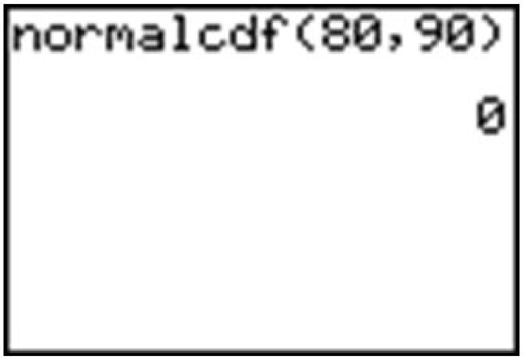
1. Estimate the standard deviation of the following distribution.



2. The $z-$table most commonly gives the probabilities ***below*** the given $z-$score, or what are sometimes referred to as ***left tail probabilities***. Probabilities above a certain $z-$score are complementary to those below, so all we have to do is subtract the table value from 1. To calculate the probabilities ***between*** two $z-$scores, calculate the left tail probabilities for both $z-$scores and subtract the left-most value from the right. Try these using the ***table only***!!

   (a) $P(z \geq -0.79)$
   (b) Use the table to verify the empirical rule value for: $P(-1 \leq z \leq 1)$. Show all work.
   (c) $P(-1.56 \leq z \leq 0.32)$

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
```

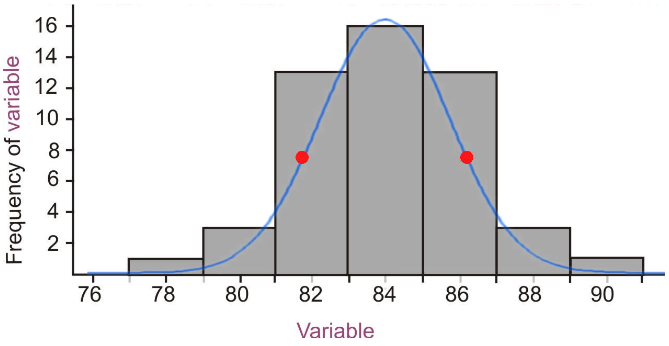
   (a) Explain her mistake and the resulting answer on the calculator.
   (b) Calculate the correct answer.

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 is 82. The standard deviation of $A's$ scores is 10 and the standard deviation of $B's$ 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.

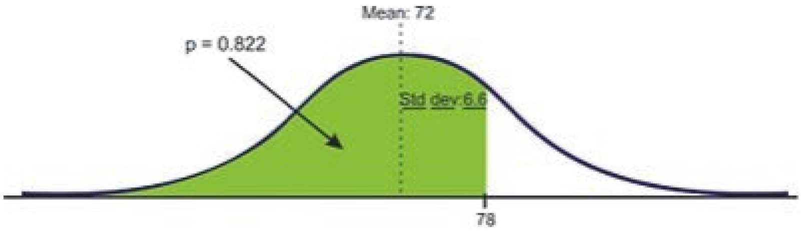
Review Answers

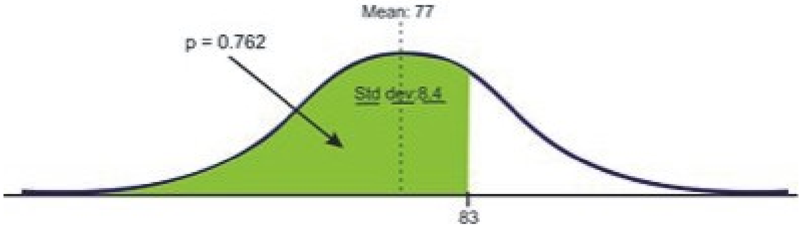
1. Here is the distribution with a density curve drawn and the inflection points estimated.



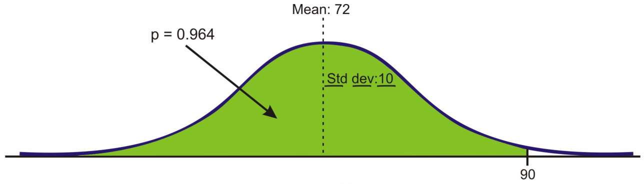
   The distribution appears to be normal. The inflection points appear to be a little more than two units from the mean of 84, therefore we would estimate the standard deviation to be a little more than two. Using the frequencies, the middle three bins contain 42 of the 50 values. Approximately 34 of the values should be within one standard deviation, which is consistent with our estimate.

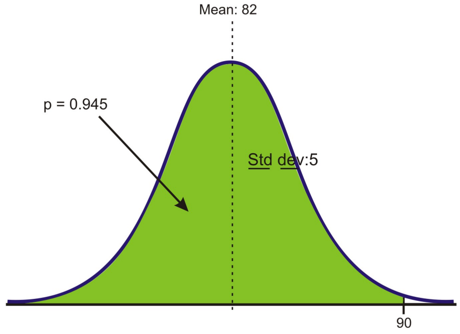
2. (a) $1 - 0.2148 = 0.7852$
   (b) $P(z \leq -1) = 0.1587, P(z \leq 1) = 0.8413, 0.8413–0.1587 = 0.6826$
   (c) $P(z \leq -1.56) = 0.0594, P(z \leq 0.32) = 0.6255, 0.6255–0.0594 = 0.5661$
3. (a) Brielle did not enter the mean and standard deviation. The calculator defaults to the standard normal curve, so the calculation she performed is actually explaining the percentage of data between the $z-$scores of 80 and 90. There is virtually 0 probability of experiencing data that is over 80 standard deviations away from the mean, especially given a test grade presumably out of 100.
   (b) 0.525 or about 53%
4. The 78 is a better grade. The percentile for that score is slightly higher, at about 82%, than the 77, which is only about the 76^{th} percentile.



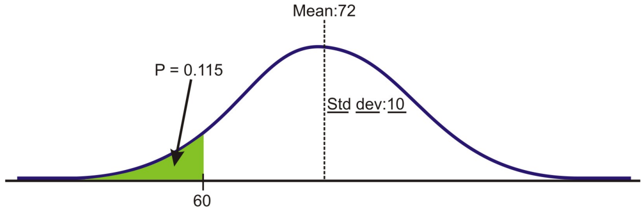


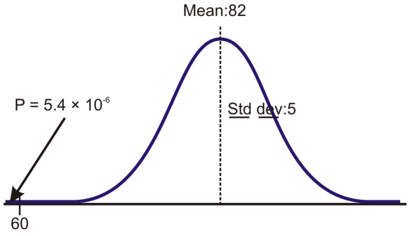
5. (a) $A90$ is a better score with teacher A.





(b) $A60$ is by far a much lower score with teacher B.





References

Java Applets that calculate standard normal probabilities:

- http://davidmlane.com/hyperstat/z_table.html

Some online normal probability tables:

- http://www.isixsigma.com/library/content/zdistribution.asp
- http://www.statsoft.com/textbook/stathome.html?sttable.html&#38;1
- http://www.math.unb.ca/~knight/utility/NormTble.htm

- http://math2.org/math/stat/distributions/z-dist.htm
- http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm

5.3 Applications of the Normal Distribution

Learning Objectives

- Apply the characteristics of the 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 truly understand the relationship between the area under a density curve and the mean, standard deviation, and $z-$score, 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. While perhaps not directly practical, it is the thorough understanding of these calculations that will lead to a high degree of comfort when a more relevant context is provided.

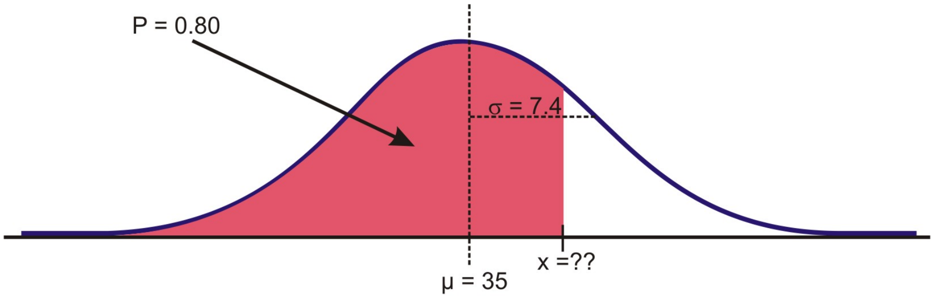
In the last lesson we found the probability, or area under the density curve. What if you are asked to find a value that gives a particular probability?

Example:

Given a 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 that is 80%?

Solution:

As suggested before, it is important and helpful to sketch the distribution.

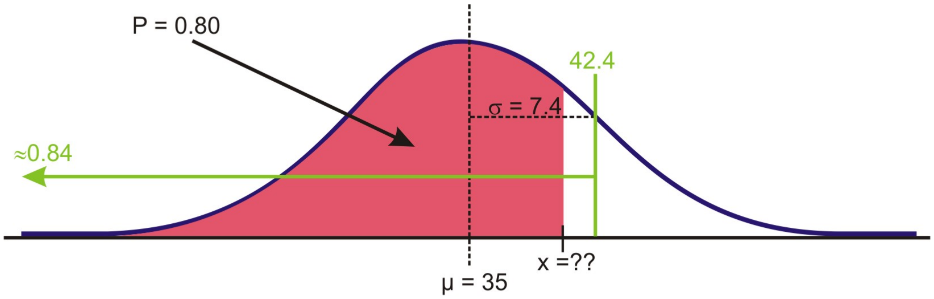


Sketch of distribution

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.

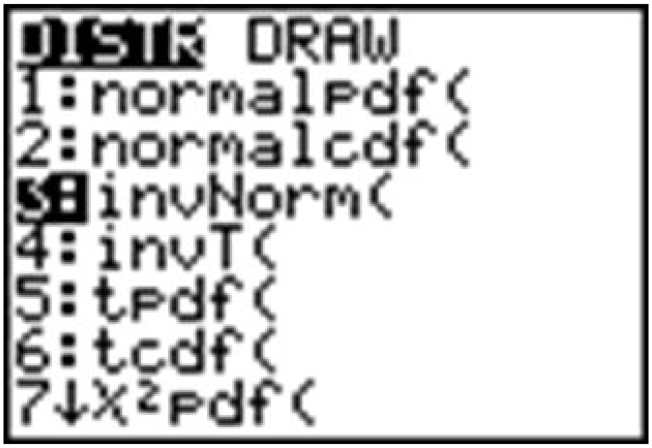
$$\mu + 1 \quad \sigma = 35 + 7.4 = 42.4$$

We expect the answer to be slightly below this value.



Estimating using the empirical rule

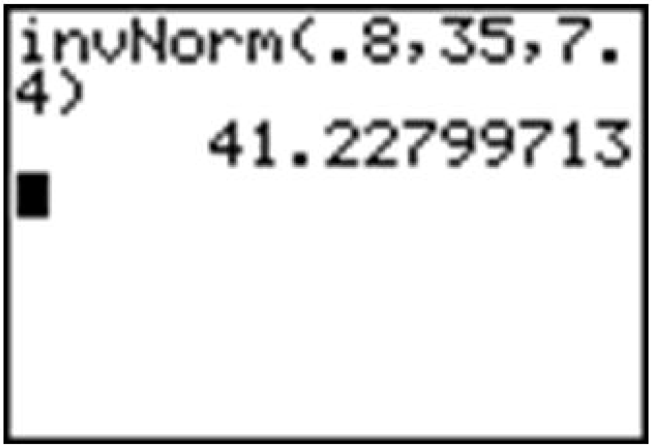
When we were given a value of the variable and were asked to find the percentage or probability, we used the $z-$table or a **normalcdf** command. But how do we find a value given the percentage? Again, the table has its limitations in this case and graphing calculators or computer software are much more convenient and accurate. The command on the TI-83/84 calculator is **invNorm.** You may have seen it already in the distribution menu.



The syntax for this command is:

InvNorm (percentage or probability to the left, mean, standard deviation)

Enter the values in the correct order:



```
invNorm(.8,35,7.
4)
      41.22799713
```

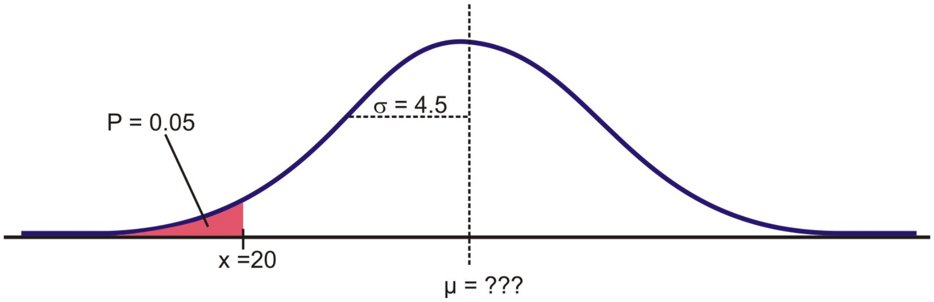
Unknown Mean or Standard Deviation

Example:

For a normally distributed random variable, $\sigma = 4.5, x = 20$, and $p = .05$, Estimate μ

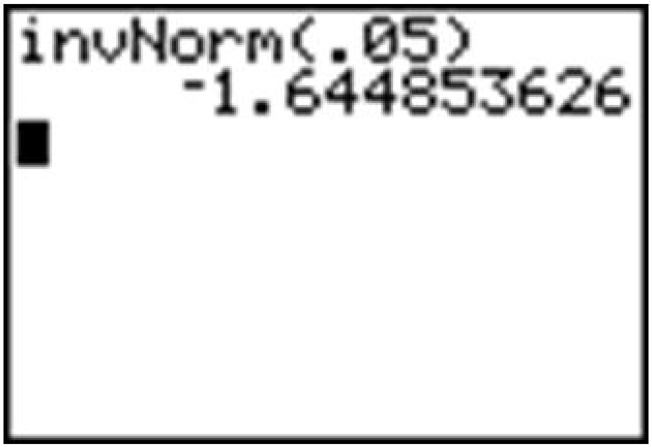
Solution:

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 should be less than 2 standard deviations from the mean.



```
invNorm(.05)
    -1.644853626
```

In one of the few instances in beginning statistics that we use algebra, plug in the known quantities into the $z-$score formula:

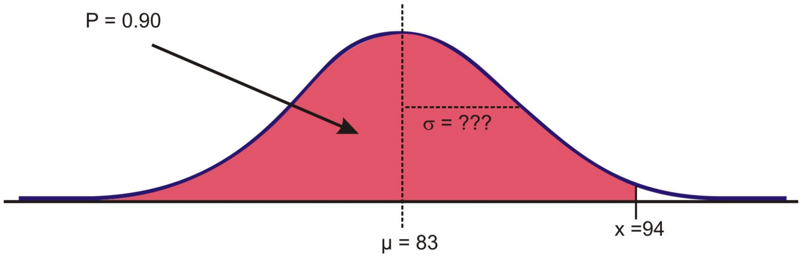
$$\begin{aligned}
z &= \frac{x-\mu}{\sigma}\\
-1.645 &\approx \frac{20-\mu}{4.5}\\
-1.645 * 4.5 &\approx 20-\mu\\
-7.402-20 &\approx -\mu\\
-27.402 &\approx -\mu\\
\mu &\approx 27.402
\end{aligned}$$

Example:

For a normally distributed random variable, $\mu = 83, x = 94$, and $p = .90$, find σ.

Solution:

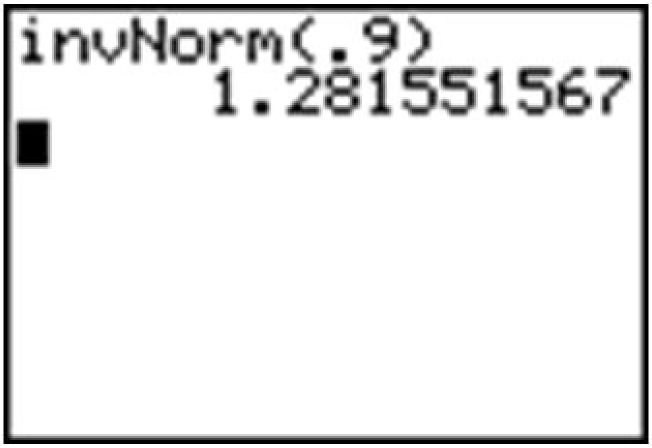
Again, let's first look at a sketch of the distribution.



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 and σ might be around 7 or 8.

Again, use invNorm to calculate the $z-$score. Remember that we are not entering a mean or standard deviation, so the result is from $\mu = 0$ and $\sigma = 1$.



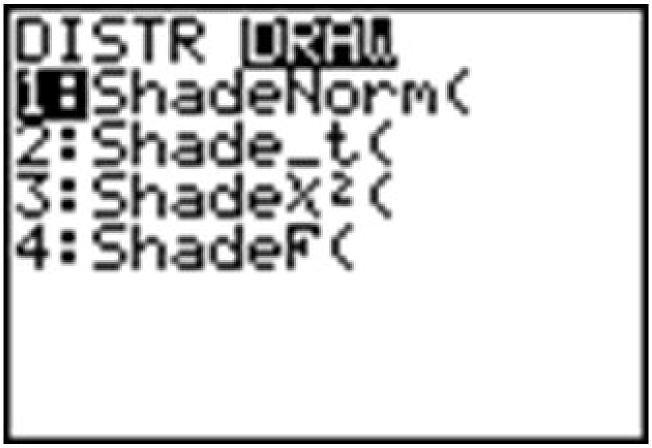
```
invNorm(.9)
        1.281551567
```

Use the $z-$score formula and solve for σ:

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

Drawing a Distribution on the Calculator

As you saw in Lesson 1 of this chapter, the TI-83/84 will also draw the distribution for you. But before doing that, 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 of the normal curve with $\mu = 83$ and $\sigma = 8.583$ below 94. 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.



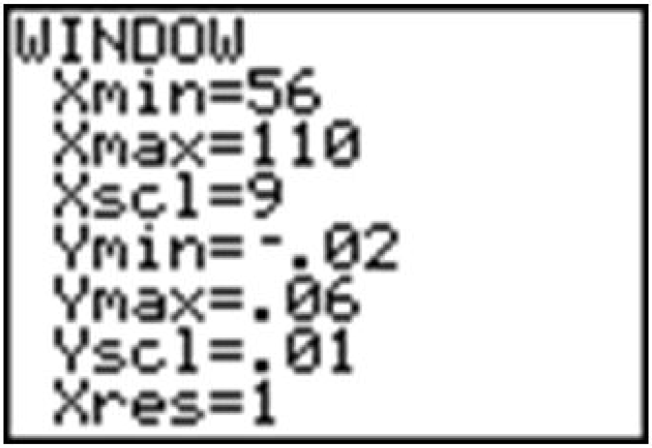
```
DISTR DRAW
1:ShadeNorm(
2:Shade_t(
3:ShadeΧ²(
4:ShadeF(
```

The reason that we went below the $x-$axis is to leave room for the text as you will see.

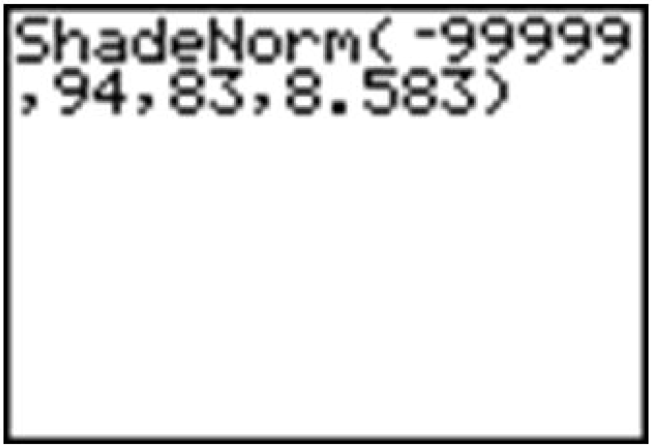
Now press [**2nd**] [**DISTR**]> and arrow over to the **Draw** option.

Choose the **ShadeNorm** command. You enter the values just as if you were doing a **normalcdf** calculation:

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

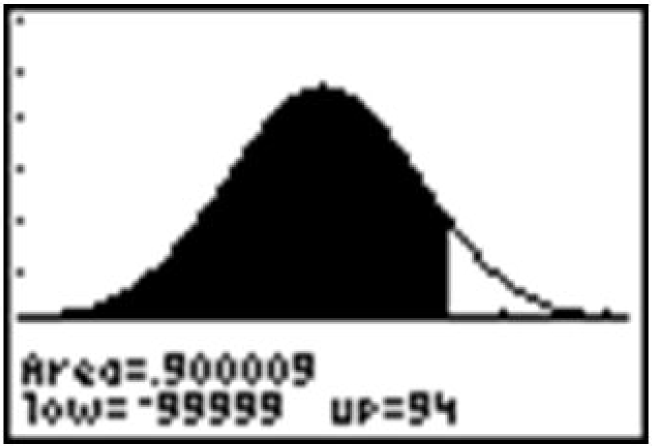


```
WINDOW
 Xmin=56
 Xmax=110
 Xscl=9
 Ymin=⁻.02
 Ymax=.06
 Yscl=.01
 Xres=1
```



```
ShadeNorm(-99999
,94,83,8.583)
```

Press [**ENTER**] to see the result.



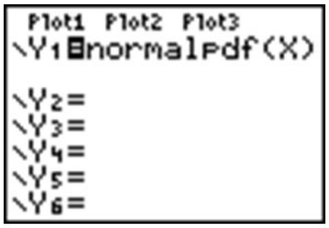
```
Area=.900009
low=-99999   up=94
```

Normalpdf on the Calculator

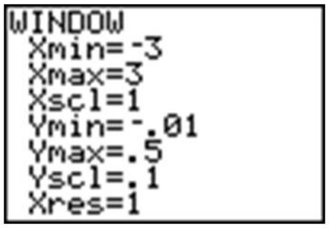
You may have noticed that the first option in the distribution menu is **Normalpdf,** which stands for a normal probability density function. It is the option you used in lesson 5.1 to draw the graph of the normal distribution. Many students wonder what this function is for and occasionally even use it by mistake to calculate what they think are cumulative probabilities. This function is actually the mathematical formula for drawing the normal distribution. You can find this formula in the resources at the end of the lesson if you are interested. The numbers this formula returns are not really useful to us statistically. The primary useful purpose for this function is to draw the normal curve.

As you did in Lesson 5.1, plot Y1=**Normalpdf** with the window shown below. Be sure to turn off any plots and clear out any functions. Enter x and close the parentheses. Because we did not specify a mean and standard deviation, we will draw the standard normal curve.

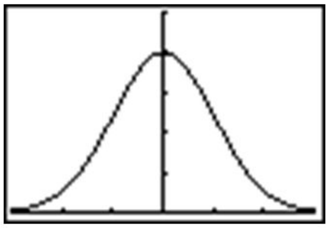
Enter the window settings necessary to fit most of the curve on the screen as shown below (think about the empirical rule to help with this).



```
 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

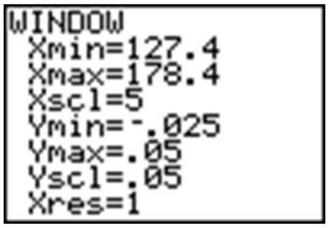
The foundation of collecting surveys, samples, and experiments is most often based on the normal distribution as you will learn in later chapters. Here are two examples.

Example:

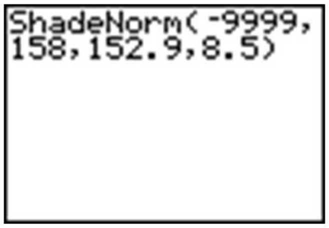
The Information Centre of the National Health Service in Britain collects and publishes a great deal of information and statistics on health issues affecting the population. One such comprehensive data set tracks information about the health of children[1]. According to their statistics, in 2006 the 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).

Part 1 If 12 year old Cecil is 158 cm, approximately what percentage of all 12 year-old boys in Britain is he taller than?

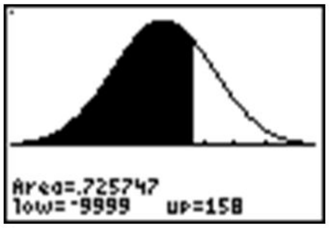
We first must assume that the height of 12 year-old boys in Britain is normally distributed. This seems a reasonable assumption to make. As always, the first step should be to draw a sketch and estimate a reasonable answer prior to calculating the percentage. In this case, let's use the calculator to 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 that value to/from the mean to find the horizontal extremes. Then enter the appropriate **ShadeNorm** command.



```
WINDOW
 Xmin=127.4
 Xmax=178.4
 Xscl=5
 Ymin=-.025
 Ymax=.05
 Yscl=.05
 Xres=1
```



```
ShadeNorm(-9999,
158,152.9,8.5)
```

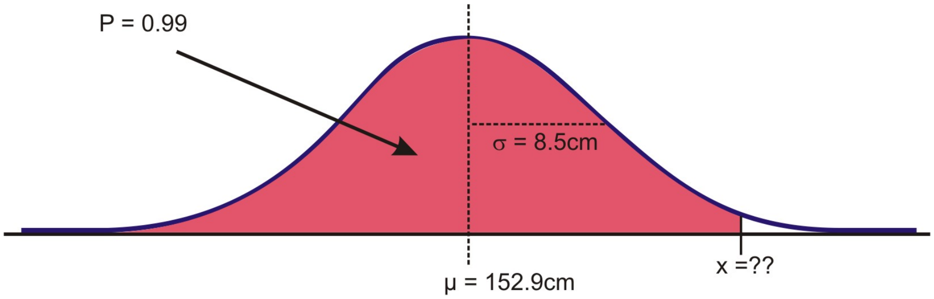


```
Area=.725747
low=-9999    up=158
```

From this data, we would estimate Cecil is taller than 73% of 12 year-old boys. We could also phrase this answer as follows: the probability of a randomly selected British 12 year-old boy being shorter than Cecil is 0.73. Often with data like this we use percentiles. We would say Cecil is in the 73^{th} percentile for height among 12 year-old boys in Britain.

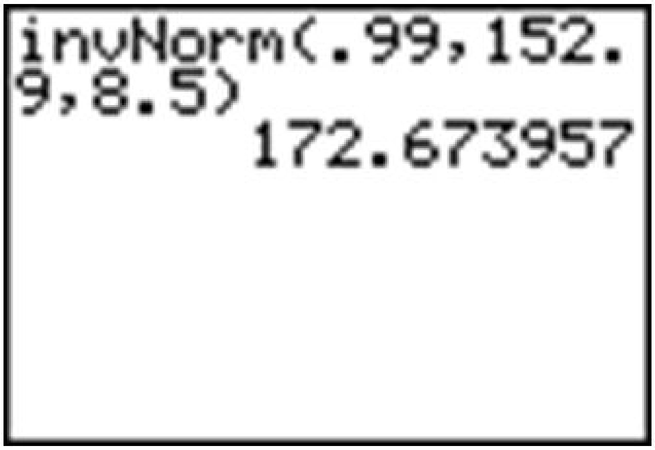
Part 2 How tall would Cecil need to be to be in the top 1% of all 12 year-old boys in Britain?

Here is a sketch:



Sketch for Height Problem

In this case we are given the percentage, so we need to use the **invNorm** command.



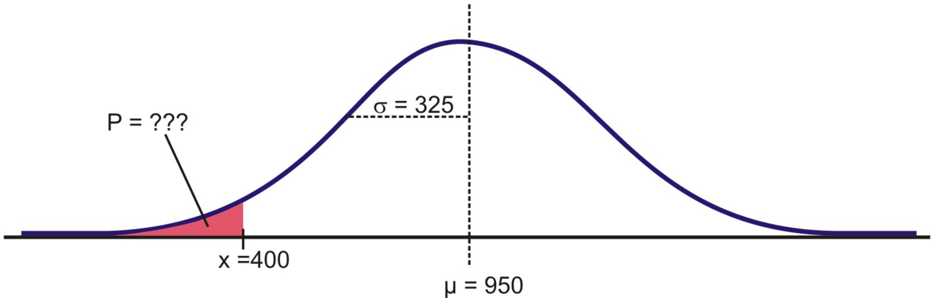
```
invNorm(.99,152.
9,8.5)
          172.673957
```

Cecil would need to be about 173 cm tall to be in the top 1% of 12 year-old boys in Britain.

Marine Iguanas in Puerto Villamil, Isabela Island, Galapagos, Ecuador
Photo by Larry Ottman

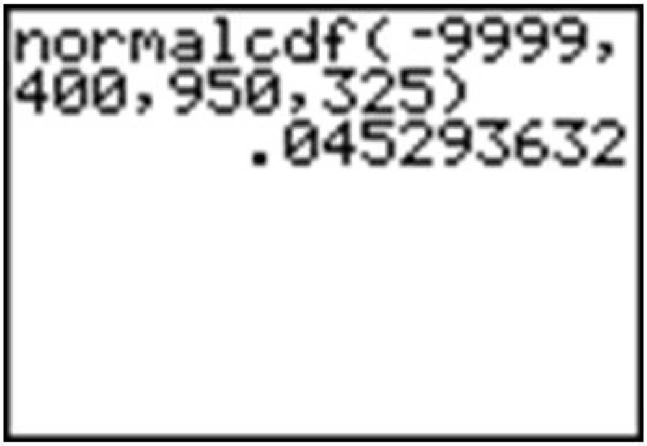
Example:

Suppose that the distribution of mass of female marine iguanas 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?



Solution:

Using the graphing calculator we need to approximate the probability of being less than 200 grams.



```
normalcdf(-9999,
400,950,325)
            .045293632
```

With a probability of approximately 0.045, we could say it is rather unlikely (only about 5% of the time) 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.

Points to Consider

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

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.

   (a)

mean	Standard deviation	x	P
85	4.5		0.68

   (b)

mean	Standard deviation	x	P
	1	16	0.05

   (c)

mean	Standard deviation	x	P
73		85	0.91

   (d)

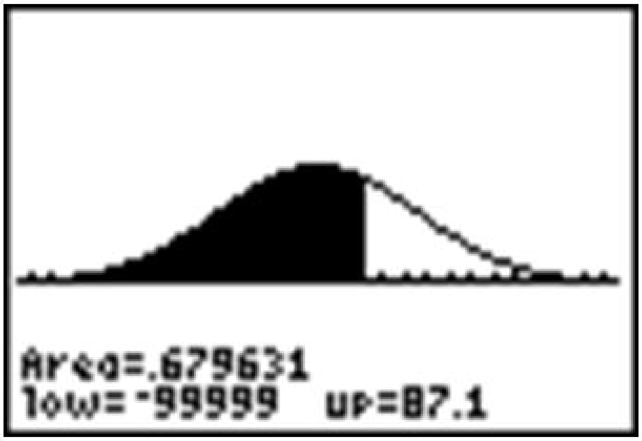
mean	Standard deviation	x	P
93	5		0.90

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 .0035 cm to 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 $1/10^{th}$ of 1% of the bearings that are made are ineffective. What should the standard deviation of the manufactured bearings 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 $\sigma = .04$ ounces.
   (a) What proportion of candy bars weigh less than the advertised weight?
   (b) What proportion of candy bars weight between 2.2 and 2.3 ounces?
   (c) What weight candy bar would be heavier than all but 1% of the candy bars out there?
   (d) If the manufacturer wants to adjust the production process so no more than 1 candy bar in 1000 weighs less than the advertised weight, what should the mean of the actual weights be? (Assuming 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??

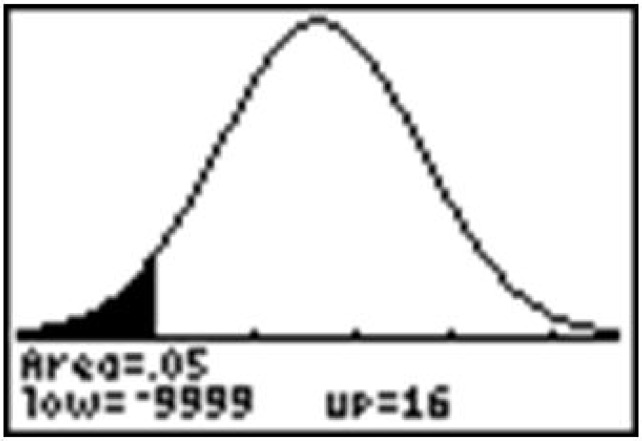
Review Answers

1. e
2. (a) 87.1



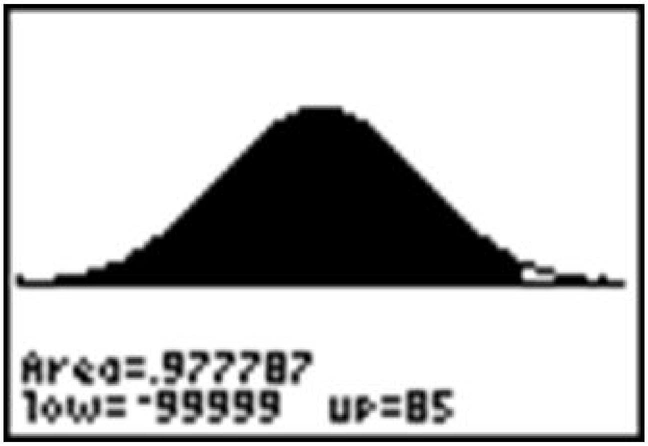
```
Area=.679631
low=-99999   up=87.1
```

(b) 17.64



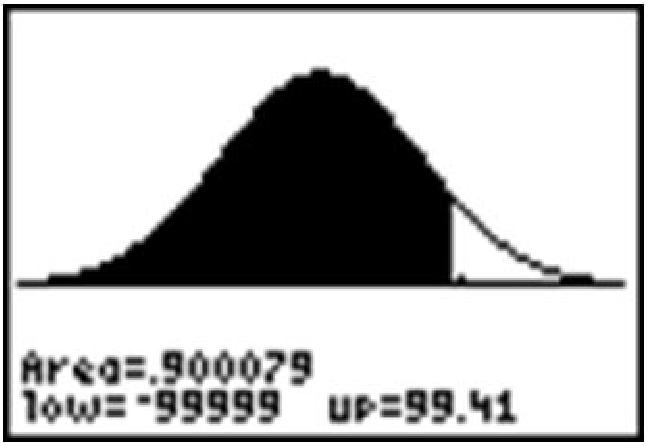
```
Area=.05
low=-9999    up=16
```

(c) 8.96



Area=.977787
low=-99999 up=85

(d) 99.41



Area=.900079
low=-99999 up=99.41

3. −0.674
4. $\sigma \approx 0.00106$
5. (a) ≈ 0.04
   (b) ≈ 0.49
   (c) ≈ 2.29 ounces
   (d) ≈ 2.254 ounces
   (e) ≈ 0.023 ounces

References

- www.ic.nhs.uk/default.asp?sID=1198755531686
- www.nytimes.com/2008/04/04/us/04poll.html

Other sites of interest

- This one contains the formula for the normal probability density function: http:

//davidmlane.com/hyperstat/A25726.html

- This one contains some background of the normal distribution, including a picture of Carl Friedrich Gauss, a German mathematician who first used the function: http://www.willamette.edu/~mjaneba/help/normalcurve.html
- This one is highly mathematical: http://en.wikipedia.org/wiki/Normal_distribution

Keywords

Normal Distribution

Density Curve

Standard Normal Curve

Empirical Rule

Z **Scores**

Normal Probability Plot (or Normal Quantile Plot)

Cumulative Density Function

Probability Density Function

Inflection Points

Image Sources

Chapter 6

Planning and Conducting an Experiment or Study

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 current issues, such as election results, approval ratings of current leaders, or opinions about economic or foreign policy issues. In an article that explains some of the details of a recent poll entitled "How the Poll Was Conducted" 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 the two different potential problems with opinion polls, surveys, observational studies, and experiments. In chapter 1, we identified some of the basic vocabulary of populations and sampling. In this lesson, we will review those ideas and investigate the sampling in more detail.

Census vs. Sample

In Chapter 1 we identified a **population** as the entire group that is being studied. A **sample** is a small, representative subset of the population. If a statistician or other researcher really 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, a smaller sample is used to generalize from. If we really wanted to know the true approval rating of the president, for example, we would have to ask every single American adult their opinion. There are some obvious reasons that 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. How would you even be sure that you could find every American adult? It would take an army of such workers and many hours to organize, interpret, and display this information. Even if all those problems could be overcome, how long do you think it would take? Being overly optimistic that 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 the results would be obsolete.

Another reason to avoid a census is when it is destructive to the population. 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 to sell! 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 US Census is probably the largest and longest running census. 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 new Census is scheduled for 2010 and in a report by the Government Accountability Office in 1994, was estimated to cost \$11 billion[2]. 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 US 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 population. As we learned in Chapter 1, 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**. In later chapters, you will learn the statistical theory behind these calculations. The second statement quoted from the New York Times article mentions the other problem with sampling. 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.

Flipping Coins

To help understand these ideas, let's look at a more theoretical example. A coin is considered "fair" if the probability, p, of the coin landing on heads is the same as the probability of landing on tails ($p = 0.5$). The probability is defined as the proportion of each result obtained from flipping the coin infinitely. A census in this example would be an infinite number of coin flips, which again is quite impractical. So instead, we might try a sample of 10 coin flips. Theoretically, you would expect the coin to land on heads 5 times. But it is very possible that, due to chance alone, we would experience results that differ from the actual probability. 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://shazam.econ.ubc.ca/flip/. You can see the random variation among samples by asking for the site to flip 100 coins five times. Our results for that experiment produced the following number of heads: $45, 41, 47, 45$, and 45, which seems quite strange, since the expected number is 50. 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

Sampling bias refers in general to the methods used in selecting the sample for a survey, observational study, or experiment. The **sampling frame** is the term we use to refer to the group or listing from which the sample is to be chosen. If we 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. The following are some of the more common sources of potential sampling bias.

Incorrect Sampling Frame

If the list from which you choose your sample does not accurately reflect the characteristics of the population, this is called **incorrect sampling frame**. A sampling frame error occurs when some group from the population does not have the opportunity to be represented in the sample. Surveys are often done over the telephone. You could use the telephone book as a sampling frame by choosing numbers from the phonebook. In addition to the many other potential problems with telephone polls, some phone numbers are not listed in the telephone book. Also, if your population includes all adults, it is possible that you are leaving out important groups of that population. For example, many younger adults especially tend to only use their cell phones or computer based phone services and may not even have traditional phone service. The sampling frame does not need to be an actual list. 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 bias.

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 undercovered.

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 that share common values, interests, and opinions. This could cause those opinions to be over-represented in relation to the true population. 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 under coverage.

Judgment Sampling

Judgment sampling occurs when an individual or organization, 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 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 they decide who to include? It would be left to their 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. Related 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 more or less 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 have a greater chance of being chosen than smaller ones. Suppose that we wanted to do a survey to find out the typical size of a student's math class at this school. The chances are greater that you would choose someone from a larger class. To understand this, let's use a very simplistic example. Say that you went to a very small school where there are only four math classes, one has 35 students, and the other three have only 8 students. If you simply choose a student at random, there are more students in the larger class, so it is more likely you will select students in your sample who will answer "35".

For example, people driving on an interstate highway tend 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 most often be true! Let's say though, 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. It could in fact be true that most of the people on the road at that time are going the same exact speed as our driver. 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 overrepresented. 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 over or under coverage. We could also cite incorrect sampling frame or convenience bias as potential problems in this example.

Response Bias

We will use the term **response bias** to refer in general terms to the types of 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. When people hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly past the interviewer on the street. We just don't know how those individuals beliefs and opinions reflect those of the general population and therefore almost all surveys could be prone to **non-response bias**.

Questionnaire Bias

Questionnaire bias occurs when the way in which the question is asked influences the response given by the individual. It is possible to ask the same question in two different ways that would lead individuals with the same basic opinions to respond differently. Consider the following two questions about gun control.

Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?

Do you believe that it is reasonable for the government to infringe on an individual's constitutional right to bear arms?

A gun rights activist might feel very strongly that the government should never be in the position of limiting guns in any way and would answer no to both questions. Someone who is very strongly against gun ownership would similarly answer no to both questions. However, individuals with a more tempered, middle position on the issue might believe in an individual's right to own a gun under some circumstances while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the wording of a question to obtain a certain response to a poll question. Questionnaire bias is not necessarily always a deliberate action. If a question is poorly worded, confusing, or just plain hard to understand it could lead to non-representative results. When you ask people to choose between two options, it is even possible that the order in which you list the choices may influence their response!

Incorrect Response Bias

A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individually 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. As the dangers of donated blood being tainted with diseases carrying a negative social stereotype developed in the 1990's, the Red Cross deals with this type of 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 them on through donated blood[4]. Screening for these 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

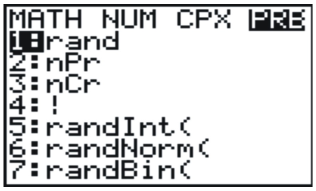
Reducing Bias: Randomization and other Techniques

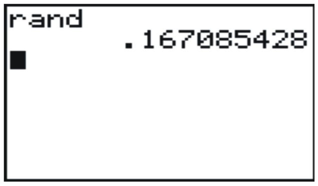
Randomization

The best technique for reducing bias in sampling is **randomization**. A **simple random sample** (commonly referred to as an **SRS**) is a technique in which all units 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, they could simply place the names of all the students in the class in a hat, mix them up, and choose one. More scientifically, we could assign each student in the class a number from 1 to say 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.

A note about "randomness"

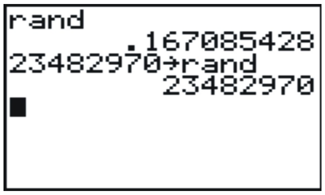
Your graphing calculator has a random number generator. Press [**MATH**] and move over to [**PRB**], which stands for probability. (Note: instead of pressing the right arrow three times, you can just use the left once!). Choose rand for the random number generator and press [**ENTER**] twice to produce a random number between 0 and 1. Press [**ENTER**] a few more times to see more results.





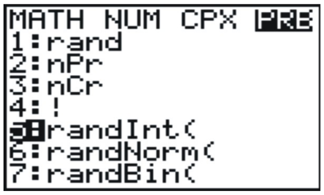
```
rand
          .167085428
■
```

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. Two brand new calculators will give the exact same sequence 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 someone else's, you need to seed your random number function using a number different from theirs. Type a unique sequence of digits on the homescreen and then press [**STO**], enter the rand function, and press [**ENTER**]. As long as the number you chose to seed the function is different, you will get different results.



```
rand
          .167085428
23482970→rand
            23482970
■
```

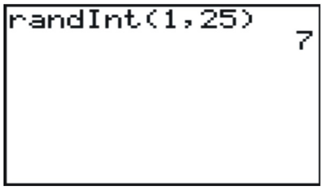
Now, back to our example, if we want to choose a student, at random, between 1 and 25, we need to generate a random integer between 1 and 25. To do this, press [**MATH**], [**PRB**], and choose the random integer function.



The syntax for this command is as follows:

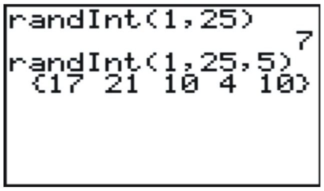
RandInt(starting value, ending value, number of random integers)

The default for the last field is 1, so if you only need a single random digit, you can enter:



```
randInt(1,25)
                 7
```

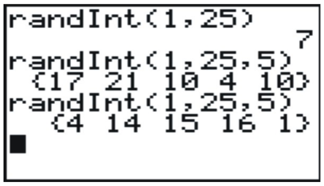
In this example, the student chosen would be student #7. If we wanted to choose 5 students at random, we could enter:



```
randInt(1,25)
                 7
randInt(1,25,5)
 {17 21 10 4 10}
```

However, because the probabilities of any digit being chosen each time are independent, it is possible to choose the same student twice.

What we can do in this case is ignore any repeated digits. Student 10 has already been chosen, so 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.

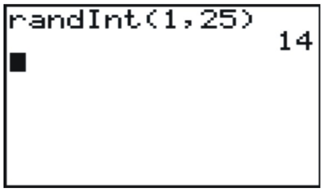


```
randInt(1,25)
                 7
randInt(1,25,5)
 {17 21 10 4 10}
randInt(1,25,5)
  {4 14 15 16 1}
■
```

In this example, student 4 was also already chosen, so we would select #14 as our fifth student.

Systematic Sampling

There are other types of samples that are not simple random samples. In **systematic sampling**, after choosing a starting point at random, subjects are selected using a **jump number** chosen at the beginning. 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 you wanted to choose five of them, we would choose every 5^{th} student. Generate a random number from 1 to 25.



```
randInt(1,25)
                  14
■
```

In this case we would start with student #14 and then generate every fifth student until we had five in all, and when we came to the end of the list, we would continue the count at number 1. Our chosen students would be: $14, 19, 24, 4, 9$. 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 to be 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 from random out of that group, or cluster into smaller subgroups, this is referred to as **multi-stage sampling**. 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 those 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 approximately equal numbers, then we could simply choose to take an *SRS* of size 25 from each school. If the numbers in each stratum are different, then it would be more appropriate to choose a fixed sample (100 students, for example) from each school and take a number from each school proportionate to the total school size.

Lesson Summary

If you collect information from every unit in a population, it is called a **census**. Because censuses are 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, 100% 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 choosing or surveying 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 sub group 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

1. How is the margin of error for a survey calculated?
2. What are the effects of sample size on sampling error?
3. Is the plural of census censuses, or censi?

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 underrepresented?
(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:
(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 1000^{th} 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 12^{th} 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?

Review Answers

1. (a) All high school soccer players.

(b) Each individual high school soccer player.

(c) A census.

(d) Boys, students from other areas of the country of different socio-economic or cultural backgrounds, if she is on a varsity team, perhps JV or freshman soccer players might have different preferences.

(e) There are multiple answers, which is why the explanation is very important. The two most obvious sources are:

Convenience bias, she asked the group that was most easily accessible to her, her own teammates.

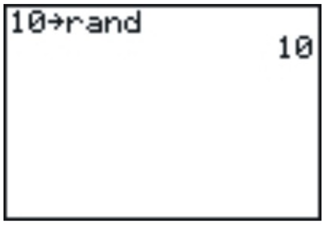
Incorrect Sampling frame, boys or some of the other undercovered groups mentioned in *d*, have no chance of being included in the sample.

(f) The sampling frame.

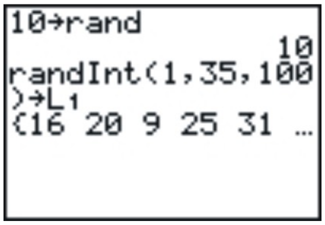
(g) Stratification.

2. This is incorrect response bias. You are intentionally answering the question incorrectly so as to not antagonize the giant talking snake!
3. The biologist is using her knowledge of moose behavior to choose an area and a time in which to estimate the population, this is judgment sampling. She has also selected one particular lake to estimate the entire region, which could be considered a form of cluster sampling.
4. Systematic sampling. The customer is selected based on a fixed interval.
5. Convenience bias. The first 30 riders is an easy group to access. Incorrect Sampling Frame. The first riders of the day are likely to be those who are most excited by high-thrill rides and may not have the same opinions as those who are less enthusiastic about riding.
6. Cluster sampling. A group is chosen because of a natural relationship that does not necessarily have any similarity of response, i.e. we have no reason to believe that people wearing a certain color would respond similarly, or differently, from anyone else in the population.

7. Voluntary response bias. Participants will self-select. Non-response bias. A large percentage of potential participants are not going to want to be bothered participating in a survey at the end of a long day at an amusement park.
8. There are several potential answers. Incorrect Response Bias. The chosen participants might not want to admit to being scared in front of the young lady. Questionnaire bias. The question is definitely worded in a manner that would encourage participants to answer in a particular way. This is also systematic sampling and someone used their judgment that only boys should be surveyed. A case could also be made for incorrect sampling frame as no girls or other age groups have a chance of being represented. All of these examples also eliminate the opinions of those in the park who do not choose to ride.
9. Stratification. It could be that people who ride at different times during the day have different opinions about thrill rides or are from different age groups. In this case, each hour is a stratum. For example, it could be that those riding early in the morning are more of the thrill seeker types, and the more hesitant folks might take some time to muster the courage to ride.
10. To make it easier to keep track of repeated choices, we have generated 100 numbers and stored them in $L1$.



```
10→rand
             10
```



```
10→rand
             10
randInt(1,35,100
)→L₁
{16 20 9 25 31 …
```

The chosen students are:

$$16, 20, 9, 31, 30, 29, 8, 10, 23, 33$$

In this example there were no repeated digits.

References

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

- http://www.gao.gov/cgi-bin/getrpt?GAO-04-37
- http://www.cnn.com/2008/TECH/04/03/census.problems.ap/
- http://en.wikipedia.org/wiki/Literary_Digest

6.2 Experimental Design

Learning Objectives

- Identify the important characteristics of an experiment.
- Distinguish between confounding and lurking variables.
- Use a random number generator to randomly assign experimental units to treatment groups.
- Identify experimental situations in which blocking is necessary or appropriate and create a blocking scheme for such experiments.
- Identify experimental situations in which a matched pairs design is necessary or appropriate and explain how such a design could be implemented.
- Identify the reasons for and the advantages of blind experiments.
- Distinguish between correlation and causation.

Introduction

A recent study published by the Royal Society of Britain[1] concluded that there is a relationship between the nutritional habits of mothers around the time of conception and the gender of their child. 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 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 may still have an effect on the other variables involved is called a **lurking variable**. For example, perhaps 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 are observed but it cannot be distinguished which one is actually causing the change in the response variable. This study also mentions 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 ender 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 rate.

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 that "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 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 and 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 layer very softly playing Mozart string quartets while they slept 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, the other group will not. When one of the treatment groups is actually withholding the treatment of interest, 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.

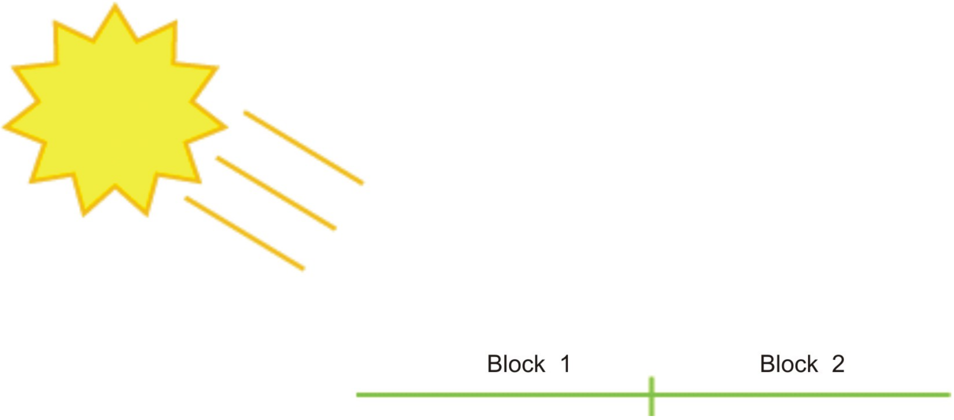
Placebos and Blind Experiments

In medical studies, the treatment group is usually receiving 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** (pronounce Pluh-see-bo). 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**. For 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. 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 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 the 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.

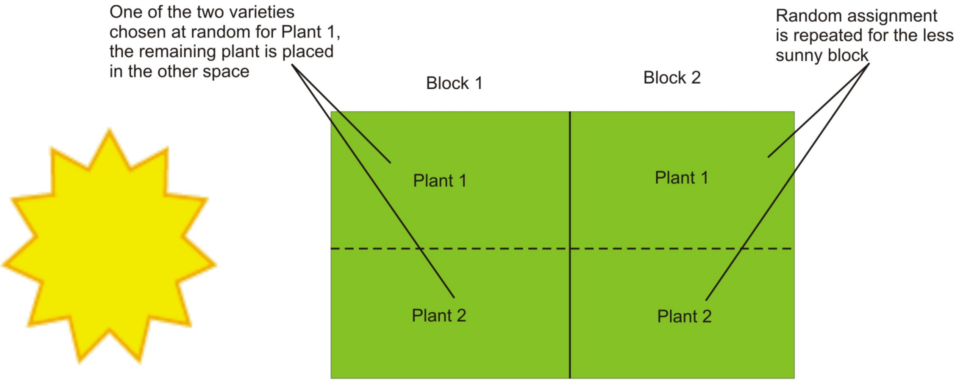
Blocking

Blocking in an experiment serves a similar purpose to stratification in a survey. 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. 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.

For 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. So 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, randomly assign the variety to one of the two positions.


One of the two varieties chosen at random for Plant 1, the remaining plant is placed in the other space
Random assignment is repeated for the less sunny block
Block 1
Block 2
Plant 1
Plant 1
Plant 2
Plant 2

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. For example, let's say we were interested in the effectiveness of two different types of running shoes. We 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 is chosen. 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 would be compared.

In the previous example, there may be some potential confounding influences. Things 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 their 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 still could be important. 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 that amount of wear on the shoe. To minimize this, we would 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. These elements are the only effective method for establishing meaningful cause and effect relationships. An experiment attempts to isolate, or **control** other potential variables to 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, it is called a **placebo.** If the participants are not aware of the treatment they are receiving, it is called a **blind experiment.** When neither the participant nor the researcher are 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 which 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 when there are two treatments. The researcher creates blocks of size two 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 it's own matched pair by applying both treatments and comparing the results.

Points to Consider

1. What are some other ways that researchers design more complicated experiments?
2. When one treatment seems to result in a notable difference, how do we know if that difference is statistically significant?
3. 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 along with the same species of ordinary flies. When released together in an environment with high competition for food, the ordinary flies survived by a significantly higher percentage than the intelligent flies.

   (a) Identify the population of interest and the treatments.
   (b) Based on the information given, 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 the 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 grade in the class. At a point during the test, you will ask the

students to stop for a moment and measure their pulse rate. You measure the at rest pulse rate of all the students in your class.

Here are those pulse rates in beats per minute:

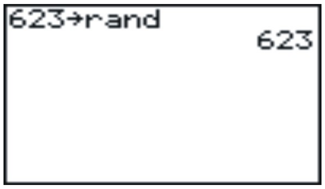
Table 6.1:

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

$$46, 72, 64, 66, 82, 44, 56, 76, 60, 62, 54, 76$$

(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.

(a) Identify any potential lurking variables in this experiment.

(b) 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.”[1]

(a) What is the population in this example?
(b) One group began the diet right away, 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. 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 using only organic fertilizer. They also 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 is 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.

Review Answers

1. (a) The population is all fruit flies of this species. The treatment is breeding for intelligence. The other treatment is really a control group. The second group of flies were not bred for any special quality.
(b) By the strict definition, this is an observational study as the subjects (fruit flies) are not randomly assigned to the treatment. A group of fruit flies was selectively bred for intelligence.
(c) Because the treatments were not randomly assigned the results are susceptible to lurking variables. It is possible that some other trait not observed in the population of intelligent fruit flies led to their lower survival rate. It is also questionable to generalize the behavior of fruit flies to the larger population of all animals. We have no guarantee that other animals will not behave differently than fruit flies. Without reading the study completely, it is difficult to determine how many of these concerns were addressed by the scientists performing the study. You can read more at:
http://www.nytimes.com/2008/05/06/science/06dumb.html?ref=science
2. (a) This is a repeated measures design. Each student becomes their own matched pair as they are sampling both colas.
(b) Students may have a preconceived idea of which cola they prefer for many possible reasons. You could have the colas already poured into identical unmarked cups,

or hide the label of the bottle. This would be an example of a blind experiment.

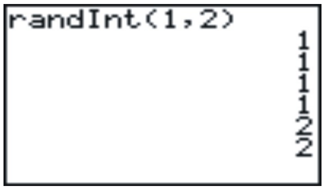
(c) It is possible that the taste of the first cola might affect the taste of the second. In general, the order in which they taste the colas could affect their perception in a number of ways. To control for this, we could randomly assign the order in which the colas are sampled. Assign one of the colas to be 1 and the other to be 2, then use your calculator to choose 1 or 2 randomly for each subject. If the student is given the two cups and given the option of choosing which one to drink first, we could randomly assign the position of each cup (right or left).

3. (a) Because students with lower pulses may react differently than students with higher pulses, we will block by pulse rate. Place the students in order from lowest to highest pulse rate, then take them two at a time.

Table 6.2:

Pair Number	Students
1	6, 1
2	11, 7
3	9, 10
4	3, 4
5	2, 8
6	12, 5

(b) The calculator would generate the following 6 random ones and twos.



```
randInt(1,2)
                 1
                 1
                 1
                 1
                 2
                 2
```

the order in which the students appear in the table as their number, the students could be assigned by placing the chosen student for each pair into treatment 1, and the remaining student to treatment 2:

Treatment 1 (black ink)	6, 11, 9, 3, 8, 5
Treatment 2 (red ink)	1, 7, 10, 4, 2, 12

(c) It is possible that different students react to testing taking and other situations differently and it may not affect their pulse directly. Some students might be better test takers than others. The level of mathematics ability or previous success on the subject matter being tested could also affect the stress level. Perhaps amount of sleep, diet, and amount of exercise may also be lurking variables.

(d) A repeated measures design would help control for individual differences in pulse rate. Each student would have to take both a black ink and red ink test. A second test would have to be carefully designed that was similar to the first, but with different color ink. If you just gave the students the same test twice, their stress level might be significantly lower when they take it the second time.

4. (a) The population is children with epilepsy who have not responded to other traditional medications.
   (b) We need assurances that the children were randomly assigned to the treatment and control groups.
   (c) The treatment is starting on the high fat diet immediately, the control group is the group who started the diet 3 months later. Notice in this case, researchers did not completely withhold the treatment from the control group for ethical reasons. This treatment has already shown some effectiveness in non-clinical trials.
   (d) We would conclude that the high fat diet is effective in treating seizures among children with epilepsy who do not respond to traditional medication.
5. We will need at least 6 blocks to impose the various treatments, which are: Organic fertilizer, 1 inch
   Chemical fertilizer, 1 inch
   Organic fertilizer, 2 inches
   Chemical fertilizer, 2 inches
   Organic fertilizer, 4 inches
   Chemical fertilizer, 4 inches
   Assign the plots numbers from 1 to 6.

1	2	3
4	5	6

   Then randomly generate a number from 1 to 6, without replacement, until all six treatments are assigned to a plot.
   In this example, the random number generator was seeded with 625, repeated digits were ignored, and the assignments were as follows:
   fertilizer, 1 inch PLOT 6
   Chemical fertilizer, 1 inch PLOT 2
   Organic fertilizer, 2 inches PLOT 1
   Chemical fertilizer, 2 inches PLOT 5
   Organic fertilizer, 4 inches PLOT 4
   Chemical fertilizer, 4 inches PLOT 3

Further reading:

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

References

- http://journals.royalsociety.org/content/w260687441pp64w5/
- http://www.fhi.no/eway/default.aspx?pid=238&#38;trg=Area_5954&#38;MainLeft_5812=5954:0:&#38;Area_5954=5825:68516::0:5956:1:::0:0

6.3 Chapter Review

Questions

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 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 book 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. Those students current averages are recorded. At the end of the year. the grades for each group are compared and the agenda book group overall 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. The response variable is using an agenda book
   (a) The explanatory variable is grades.
   (b) This is an experiment because the participants were chosen randomly.
   (c) The school should have stratified by gender.
   (d) 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.

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=156231&#38;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 the 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 is given in the following table:

Table 6.3:

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, they 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?

Answers

Multiple Choice:

1. c
2. b
3. .
4. e
5. c

Open-Ended Questions

1. Incorrect response bias. The main focus of the piece, and an issue in exit polling in general is that there is no guarantee that, for many possible reasons, subjects in an exit poll will answer truthfully. The pollsters also ask the questions in a variety of rude, unethical, and inappropriate ways that would manipulate the responses. Even though a real pollster would never actually engage in this type of behavior, it could be considered questionnaire bias.
2. (a) Randomly assign each tortoise a number from 1-9 using a random number generator, then incubate the eggs from tortoises 1-3 at 25.50 C, 4-6 at 29.50 C, and 7-9 at 33.50 C. When the tortoises hatch, observe and compare the ratio of female and male tortoises (which is not easy to do) at the various temperatures. The results of this study did confirm that the ratio of females is higher found that 29.50 C is the optimum temperature for a higher female ratio and good survival rate, and 280 C is the best temperature to insure more males (source: Restoring the Tortoise Dynasty, Godfrey Merlin, Charles Darwin Foundation, 1999.)

   (b) This would be a blocking design. We would block on species and temperature, so there would be 9 blocks, 3 of each species, and three at each incubation temperature. There really would not be any randomization in this design.

3. (a) There are two treatments, the new medication, and the existing medication. All the subjects could be told that they were receiving a new treatment, and then only some would be given the new treatment and the rest would be given their original medication. The resulting differences in skin condition between the two groups would be compared.

(b) You could assign the subjects a different numbering from 1 to 12, but this time generating the assignments at random. Then subjects 1-6 would be given the new treatment, and subjects 7-12 would be given the original medication. Compare the results.

(c) In blocking for condition, each block should be homogeneous for that trait. So, you would create three blocks: all 4 mild subjects, all 4 moderate subjects, and all 4 severe subjects. Then, within each block, randomly assign two subjects to receive the new treatment, and two to receive the original. Compare the results.

Table 6.4:

Pair Number	Gender	Age
1	**F**	13
1	**F**	13
2	**F**	15
2	**F**	16
3	**F**	17
3	**F**	18
4	**M**	14
4	**M**	14
5	**M**	15
5	**M**	16
6	**M**	17
6	**M**	18

Place the females in chronological order, then group the two youngest, the next two, and the last two. Repeat the same procedure with the males. This way we have pairs that are similar in both age and gender. One of the subjects would be chosen at random for the new treatment and the other would receive the traditional medication.

(d) Repeated measures are not a good idea with medication studies as it would be hard to distinguish if the effects from the repeated treatment are not in fact from still occurring from the presence of the first medication that was given.

Image Sources

Chapter 7

Sampling Distributions and Estimations

7.1 Sampling Distribution

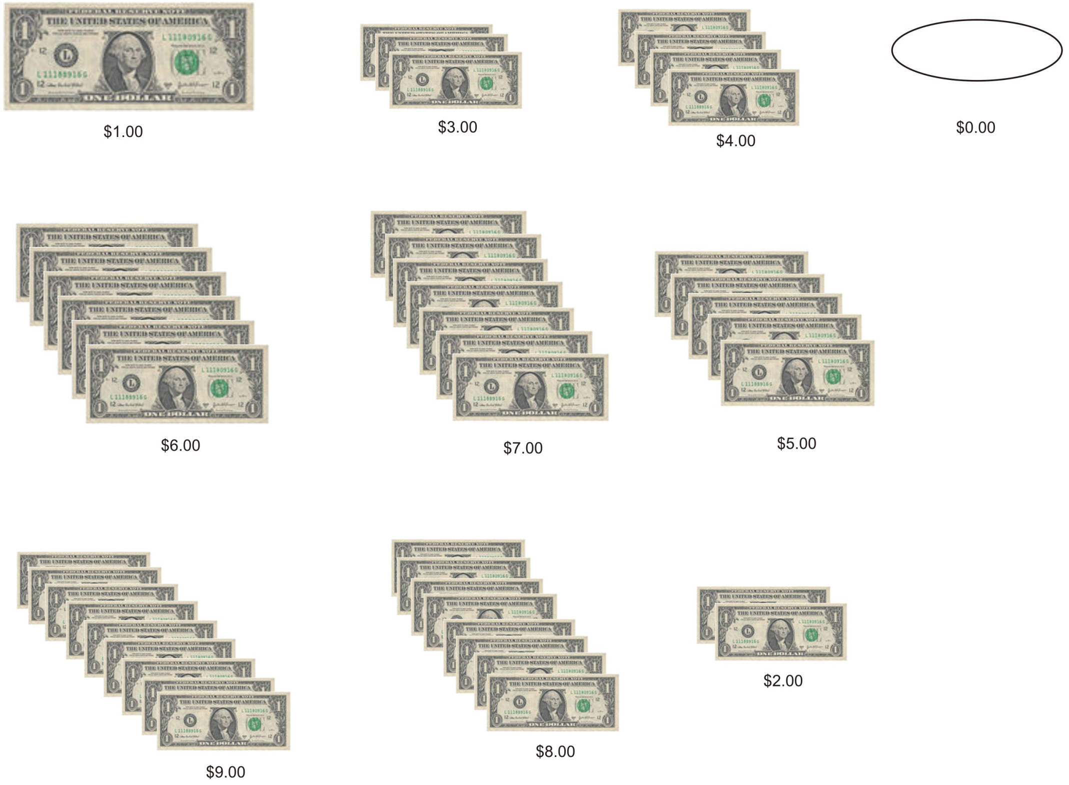
Learning Objectives

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

Introduction

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

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



Sampling Distribution

In previous chapters, you have examined methods that are good for exploration and description of data. In this section we will discuss how collecting data by random sample helps us to draw more rigorous conclusions about the data.

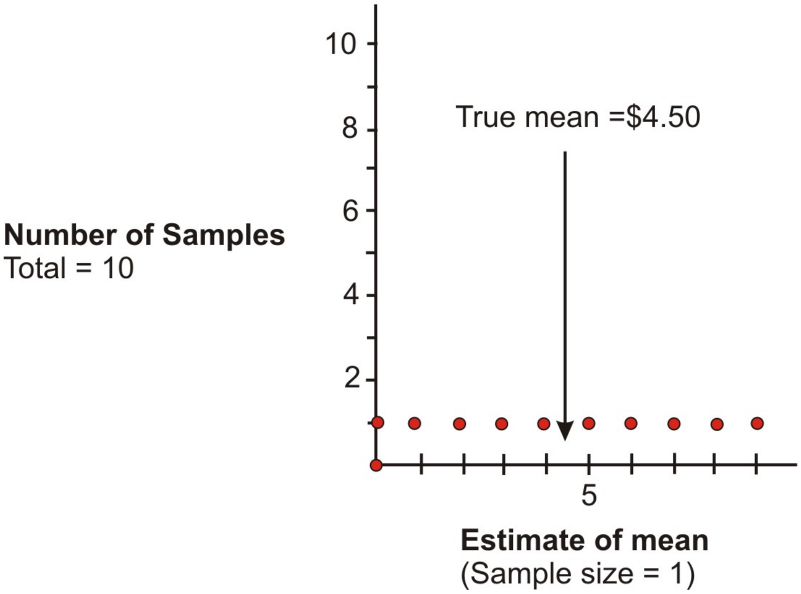
The ultimate purpose of sampling is to select a set of units or **elements** from a population that represents the parameters of the total population from which the elements were selected. Random sampling is one special type of what is called probability sampling. The reasons for using random sampling are that it erases the danger of a researcher, whether conscious or unconscious, to be biased when selecting cases. In addition, the choice of random selection allows us to use tools from probability theory that provide the bases for estimating the characteristics of the population as well as estimates the accuracy of 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. Probability theory also helps statisticians estimate the parameters of a population. A parameter is the summary description of a given variable in a population. Some examples of parameters of a population are the distribution of ages within that population, or the distribution of income levels. 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 the estimates will accurately represent the actual parameters in 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 parameter. 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.

It is now time to examine an example of sampling distribution to see how this all works. In the scenario that was presented in the introduction to this lesson, the assumption was made that in a case of size ten, one person had no money, another had $1.00, another had $2.00 etc. until we reach the person that had $9.00.

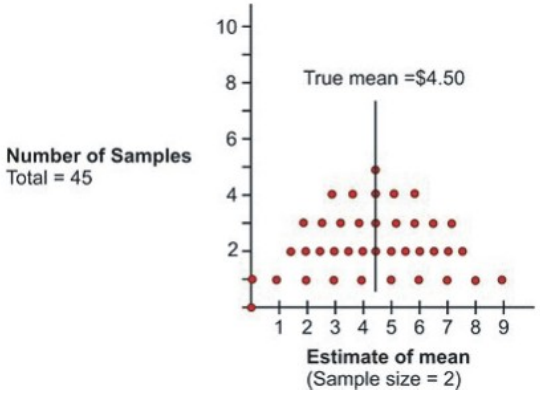
The purpose of the task is to determine the average amount of money in this population. If you total the money of the ten people, you will find that the sum is $45.00, thus yielding a mean of $4.50. To complete the task of determining the mean number of dollars of this population, it is necessary to 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 that 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 called the ***sampling distribution***. As can be concluded, 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? In other words, from a population of 10, in how many ways can two be selected if the order of the two does not matter? The sample size is now 2 and these are being randomly selected from our population. This is referred to in mathematics as a combination and can be readily obtained by using the graphing calculator.

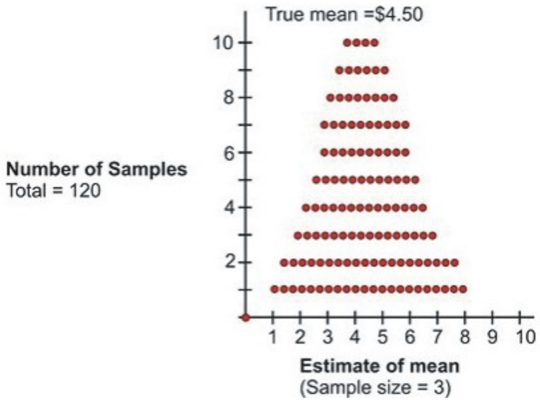
10 Math → PRB ↓ 3. $_nC_r$ 2 **enter** 45



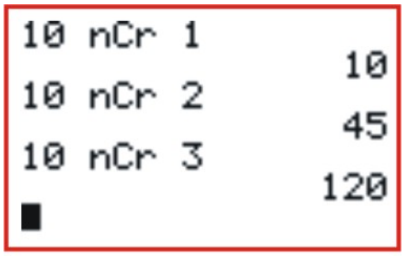
Increasing the sample size has improved your estimations. There are now 45 possible samples: and some of [$0, $1], [$0, $2], ...[$7, $8], [$8, $9]. Some of these samples produce the same means. For example [$0, $6], [$1, $5] and [$2, $4] all produce means of $3.00. The three dots above the $3.00 mean 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 that deviate by as much as $4.00. Five of the samples yield the true estimate of $4.50 and another eight deviate by only 50 cents (plus or minus).

If three are randomly selected from the population of 10, there are 120 samples.

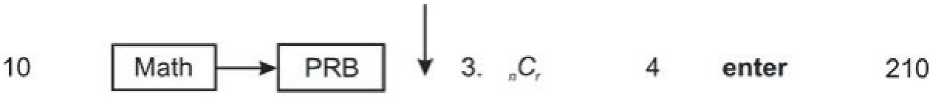
10 Math → PRB ↓ 3. $_nC_r$ 3 **enter** 120



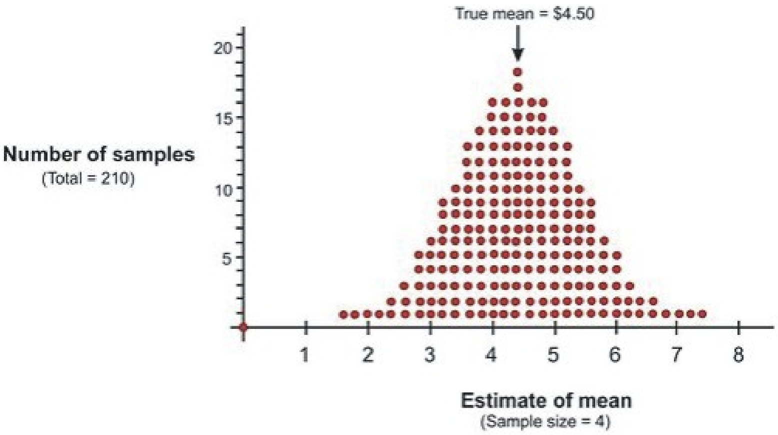
Here is a screen shot from the graphing calculator for the results of randomly selecting $1, 2$ and 3 from the population of 10. The $10, 45$ and 120 represent the total number of possible samples that are generated from increasing the sample size by 1.



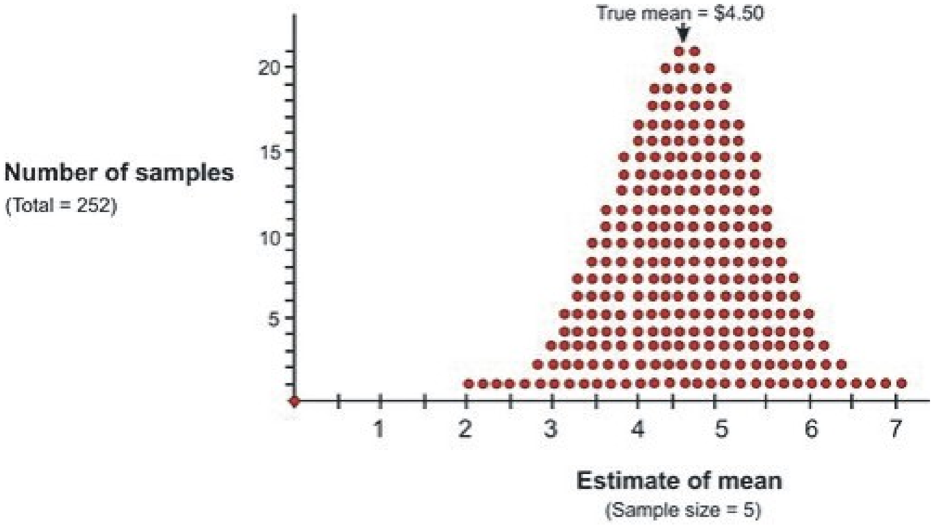
```
10 nCr 1
               10
10 nCr 2
               45
10 nCr 3
              120
█
```



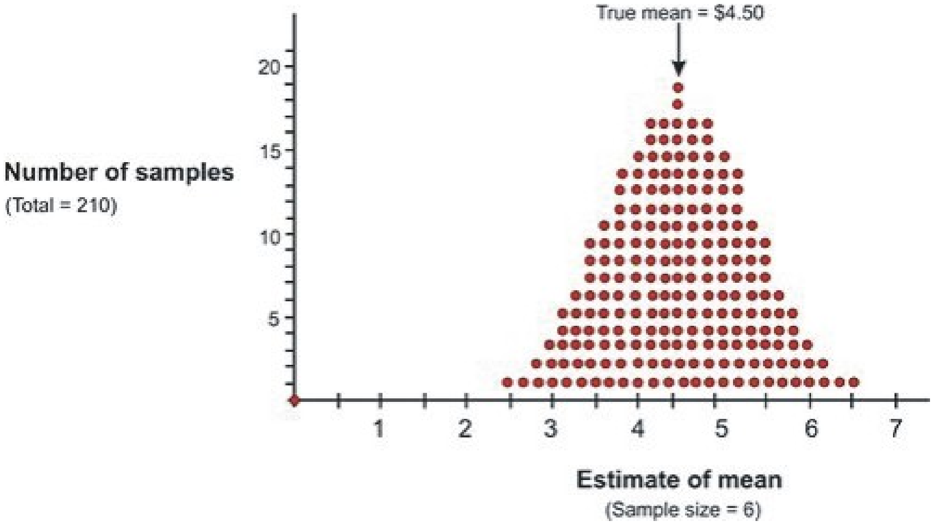
10 Math → PRB ↓ 3. $_nC_r$ 4 **enter** 210



10 $\boxed{\text{Math}} \rightarrow \boxed{\text{PRB}} \downarrow$ 3. $_nC_r$ 5 **enter** 252



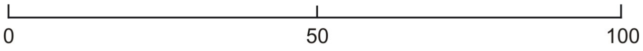
10 $\boxed{\text{Math}} \rightarrow \boxed{\text{PRB}} \downarrow$ 3. $_nC_r$ 6 **enter** 210



From the above graphs, it is obvious that increasing the sample size chosen from each sample of size 10 resulted in a distinct improvement in the distribution of estimates of the mean. If a sample size of 10 were selected, there would be only one possible sample, and it would yield the true mean of \$4.50. The sampling distribution of the sample means is approximately normal because it has the bell shape of the normal curve.

Now that you have been introduced to sampling distribution and how the sample size affects the distribution of the sampling mean, 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 population will be the $18,000$ students that attend the school. The elements will be the individual students. A random sample of 100 students will be selected for the purpose of estimating the entire student body. Attitudes toward the dress code will be the variable under consideration. For simplicity sake, assume that the attitude variable has two attributes: approve and disapprove. As you know from the last chapter, in a scenario such as this when a variable has two attributes it is called **binomial**.

The following figure shows the range of possible sample study results. The horizontal axis presents all possible values of the parameter in question. It represents the range from 0 percent to 100 percent of students approving of the dress code. The number 50 on the axis represents the midpoint, 50 percent, of the students approving the dress code and 50 percent disapproving. Since the sample size is 100, half of the students are approving and the other half are disapproving.

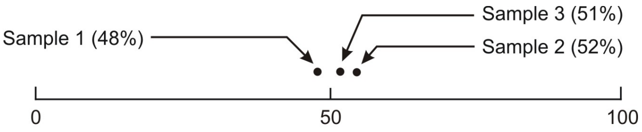


Percent of students approving of the dress code

To randomly select the sample of 100, every student is presented with a number (from 1 to $18,000$) and the sample is randomly selected from a drum containing all of the numbers.

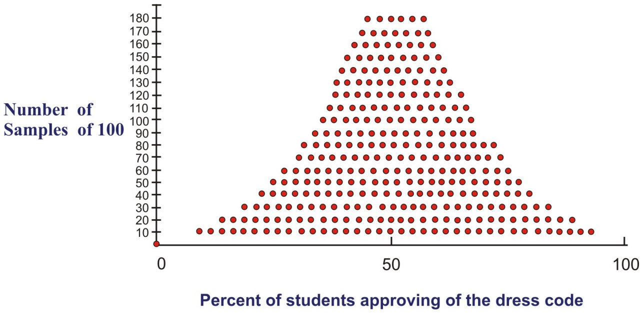
Each member of the sample is then asked whether they approve or disapprove of the dress code. If this procedure gives 48 students who approve of the code and 52 who disapprove, the result is recorded on the horizontal axis by placing a dot at 48%. This percentage describes the variable and is called a statistic.

Let's assume that the process was repeated again and this result in 52 students approving the dress code. A third sample resulted in 51 students approving the dress code.



Percent of students approving of the dress code

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



By increasing the number of samples of 100, the range of estimates provided by the sampling process has increased. It looks as if the problem in attempting to guess the parameter in the population has also become more complicated. However, probability theory provides an explanation of these results.

First, the sample statistics resulting from the samples are distributed around the population parameter. Although there is a wide range of estimates, more 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 – the degree of error expected for a given sample design. The formula $s = \sqrt{\frac{P \cdot Q}{n}}$ contains three factors: the parameters (P and Q), the sample size (n), and the standard error (s)

The symbols P and Q in the formula equal the population parameters for the binomial: If 60 percent of the student body approves of the dress code and 40% disapprove, P and Q are 60% and 40% respectively, or 0.6 and 0.4. Note that $Q = 1 - P$ and $P = 1 - Q$. The square root of the product of P *and* Q is actually the population standard deviation. The symbol n equals the number of cases in each sample, and s is the standard error.

If the assumption is made that the true population parameter is 50 percent approving the dress code and 50 percent disapproving the dress code while selecting samples of 100, the standard error obtained from the formula equals 5 percent or .05.

$$\begin{aligned} Q &= 1 - P \\ Q &= 1 - 0.50 \\ Q &= 0.50 \end{aligned} \qquad \begin{aligned} P &= 1 - Q \\ P &= 1 - 0.50 \\ P &= 0.50 \end{aligned}$$

$$s = \sqrt{\frac{P \cdot Q}{n}} \qquad s = \sqrt{\frac{(0.50).(0.50)}{100}} = 0.05 \text{ or } 5\%$$

$$\begin{aligned} \sigma &= \sqrt{P \cdot Q} \\ \sigma &= \sqrt{(0.50).(0.50)} \\ \sigma &= 0.050 \text{ or } 50\% \longrightarrow \end{aligned}$$

This is the assumption that was made
as being the true population parameter.

This 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.

Probability theory indicates that certain proportions of the sample estimates will fall within defined increments- each equal to one standard error-from the population parameter. Approximately 34 percent of the sample estimates will fall within one standard error increment above the population parameter and another 34 percent will fall within one standard error increment below the population parameter. In the above example, you have calculated the standard error increment to be 5 percent, so you know that 34% of the samples will yield estimates of student approval between 50% (the population parameter) and 55% (one standard error increment above). Likewise, another 34% of the samples will give estimates between 50% and 45% (one standard error increment below the parameter). Therefore, you know that 68% of the samples will give estimates within ±5 percent of the parameter. In addition, probability theory says that 95% of the samples will fall within ± two standard errors of the true value and 99.9% will fall within ± three standard errors. With reference to this example, you can say that only one sample out of one thousand would give an estimate below 35 percent or above 65 percent approval.

The size of the standard error is a function of the population parameter and the sample size. By looking at this formula, $s = \sqrt{\frac{P \cdot Q}{n}}$ it is obvious that the standard error will increase as a function of an increase in the quantity P *times* Q. Referring back to our example, the maximum quantity for P *times* Q occurred when there was an even split in the population. $P = .5$ so $P \ \times Q \ = .25$; If $P = .6$, then $P \ \times Q \ = .24$; if $P = .8$, then $P \ \times Q \ = .16$. If P is either 0.0 or 1.0 (none or all of the student body approve of the dress code) then the standard error will be 0. This means that there is no variation and every sample will give the same estimate.

The standard error is also a function of the sample size. As the sample size increases, the standard error decreases. This is an inverse function. As the sample size increases, the samples will be clustered closer to the true value. The last point about that formula that is obvious is noted by the square root operation. The standard error will be reduced by one-half if the sample size is quadrupled.

$$s = \sqrt{\frac{P \cdot Q}{n}}$$

$$s = \sqrt{\frac{(0.50).(0.50)}{400}} = 0.025 \quad \text{or} \quad 2.5\%$$

Lesson Summary

In this lesson we have learned about probability sampling which is the key sampling method used in controlled survey research. In the example presented above, the elements were chosen for study from a population on the basis of random selection. The sample size had a direct result on the distribution of estimates of the mean. The larger the sample size the more normal the 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 the 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 your pennies.
3. Calculate the mean of the ages of the pennies.
4. Have each student in the group randomly select a sample size of 5 pennies from the 100 coins and calculate the mean of the five ages on the chosen coins. The mean is then to be recorded on a number line. Have the students repeat this process until all of the coins have been chosen. How does the mean of the samples compare to the mean of the population(100 ages)?
5. Repeat step 4 using a sample size of 10 pennies. (As before, allow the students to work in groups)
6. What is happening to the shape of the sampling distribution of the sample means?

Review Answers

1. Many students may guess normal, but in reality the distribution is likely to be skewed toward the older pennies. (Remember that this means there are more newer pennies.)
2. The histogram will probably show the distribution skewed toward the older ages.
3. Answers will vary
4. The mean of the sampling distribution should be the same as the mean of the population.
5. .
6. The shape of the sampling distribution becomes approximately normal as the sample size increases.

Note: This activity would work very well done with an entire class. Each student could use 20 coins and the sample means could be an accumulation of sample means from each student.

Vocabulary

Element The unit may be selected in a sample. These units comprise a population.

Normal Distribution A useful and common probability distribution that has a symmetrical, upside - down U-shape or bell shape.

Parameter The summary description of a given variable in a population.

Population The entire set of the elements in a study.

Random Selection In sampling, a method of choosing representative elements where each element has an equal chance selection independent of any other event in the selection process.

Sample The set of units selected for study from a population.

Sampling Distribution The distribution of a sample statistic such as a sample mean that is the result of probability sampling.

Sampling Error The degree of error to be expected for a given probability sample design. The value of the sampling error will show how closely the sample statistics cluster around the true value of population.

Statistic The summary description of a variable in a sample.

7.2 The z-Score and the Central Limit Theorem

Learning Objectives

- Calculate the $z-$score of a mean distribution of a random variable in problem situations.
- 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 normal approximation of the binomial 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 mean for various sample sizes. It is now time to define some properties of the sampling distribution of the sample mean and to examine what we can conclude about the entire population based on it.

All normal distributions have the same basic shape and therefore rescaling and recentering can be implemented to change any normal distributions to one with a mean of zero and a standard deviation of one. This configuration is referred to as standard normal distribution. In this distribution, the variable along the horizontal axis is called the $z-$score. This score is another measure of the performance of an individual score in a population. The $z-$score measures how many standard deviations a score is away from the mean. The $z-$score of a term x in a population distribution whose mean is μ and whose standard deviation 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 zero means that the term has the same value as the mean. For the normal standard distribution, where $\mu = 0$, if we let $x = \sigma$, then $z = 1$. If we let $x = 2\sigma$, $z = 2$. Thus, a value of z tells 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?

Solution:

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

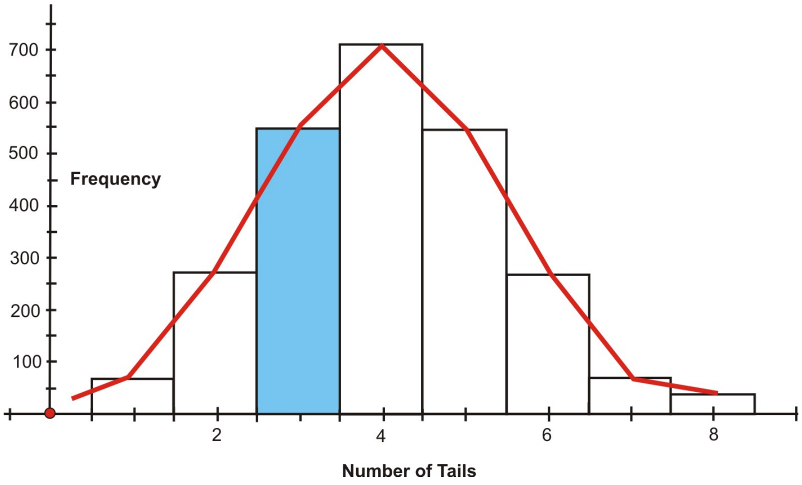
Example: On a college entrance exam, the mean was 70 and the standard deviation was 8. If Helen's $z-$score was -1.5, what was her exam mark?

Solution:

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ \therefore z \cdot \sigma &= x - \mu \\ X &= \mu + z \cdot \sigma \\ X &= (70) + (-1.5)(8) \\ X &= 58 \end{aligned}$$

Now you will see how $z-$scores are used to determine the probability of an event.

Suppose you were to toss 8 coins 2560 times. The following figure shows the histogram and the approximating normal curve for the experiment. The random variable represents the number of tails obtained.



The blue section of the graph represents the probability that exactly 3 of the coins turned up tails. One way to determine this is by the following

$$
\begin{aligned}
P(3\text{ tails}) &= \frac{{}_8C_3}{2^8}\\
P(3\text{ tails}) &= \frac{56}{256}\\
P(3\text{ tails}) &\cong 0.2186
\end{aligned}
$$

Geometrically this probability represents the area of the blue shaded bar divided by the total area of the bars. The area of the 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 is seldom used today. Below is an example of a table of $z-$scores and a brief explanation of how it works.

As shown in the illustration below, 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 0 and 2.36, look in the intersecting cell for the row labeled 2.30 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 on the 1.3 row and the 0.00 column (the area is 0.4032).

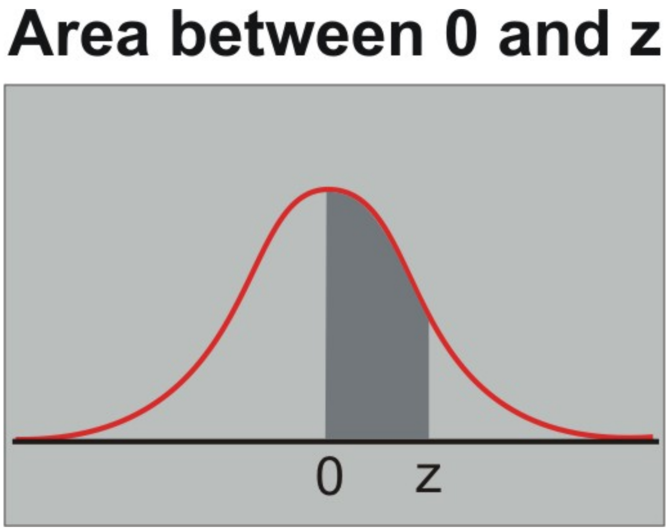


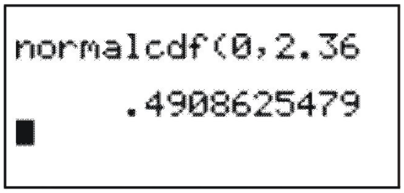
Table 7.1:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916

Table 7.1: (continued)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

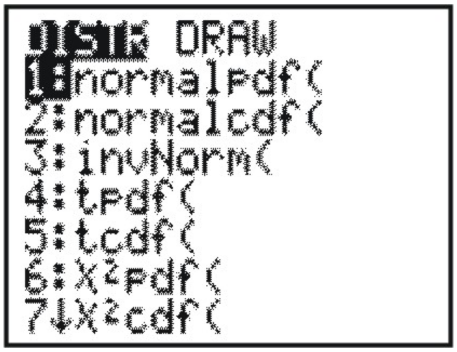
The graphing calculator will give greater accuracy in finding the proportion of values that lie between two specified values in a standard normal distribution.



```
normalcdf(0,2.36
      .4908625479
```

To use the TI-83 calculator for this operation is quite simple. Follow these steps.

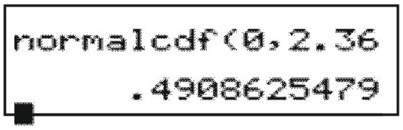
2^{nd} Vars – This will access the distribution function



Scroll down to 2: **normalcdf(enter** ⟶

This screen appears ⟶ normalcdf(

Type in the numbers $(0, 2.36$ **enter** ⟶



```
normalcdf(0,2.36
      .4908625479
```

The calculator has given an answer that is more accurate than that given in the chart.

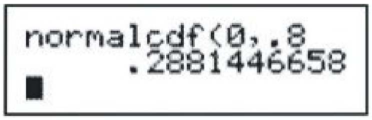
However, if the answer is rounded to the nearest ten-thousandth, then both answers would be the same. Using the calculator is a more efficient method of obtaining the $z-$score since you all have them on hand.

Example: For a normal distribution curve based on values of $\sigma = 5$ and $\mu = 20$, find the area between $x = 24$ and $x = 32$.

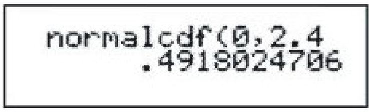
Solution:

$$\begin{array}{lcl} z = \dfrac{x-\mu}{\sigma} & \text{and} & z = \dfrac{x-\mu}{\sigma} \\ z = \dfrac{24-20}{5} & \text{and} & z = \dfrac{32-20}{5} \\ z = 0.8 & \text{and} & z = 2.4 \end{array}$$

Using the TI-83



```
normalcdf(0,.8
     .2881446658
```



```
normalcdf(0,2.4
     .4918024706
```

The area for $z = 0.8$ is 0.2881 and for $z = 2.4$ is 0.4918. Therefore the area between $x = 24$ and $x = 32$ is:

$$0.4918 - 0.2881 = 0.2037$$

This means that the relative frequency of the values between $x = 24$ and $x = 32$ is 20.37%.

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 number of trials of a random variable, the distribution of the sample trials 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 the sample mean. These symbols are shown in the table below:

Table 7.2:

	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	

In the previous lesson, you discovered that the standard error is the standard deviation of the sampling distribution and this value was calculated by using the formula $s = \sqrt{\frac{P \cdot Q}{n}}$. By making a few substitutions, this formula can be rewritten using the symbols from the chart above. The formula $s = \sqrt{\frac{P \cdot Q}{n}}$ can be expressed as the quotient of two radical expressions $s = \frac{\sqrt{P \cdot Q}}{\sqrt{n}}$. The square root of the product of the parameters P and Q is actually the standard deviation of the population (σ). When this value is divided by square root of the sample size, the result is the standard error (s), also known as the standard deviation of the sampling distribution ($S_{\bar{x}}$). Therefore $s = \sqrt{\frac{P \cdot Q}{n}}$ can be written as $S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ This frequency distribution only approximates the true sampling distribution of the sample mean because a finite number of sample means were used. If, hypothetically, an infinite number of sample means were used, the resulting distribution would be the desired sampling distribution and the following would be true:

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

The notation $\sigma_{\bar{x}}$ reminds you that this is the standard deviation of the sample mean ($\bar{x}$) and not the standard deviation (σ) 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 mean ($\bar{x}$) approximates a normal distribution as n increases.
- The mean of this sampling distribution approximates the population mean as n becomes large:

$$\mu \approx \mu_{\bar{x}}$$

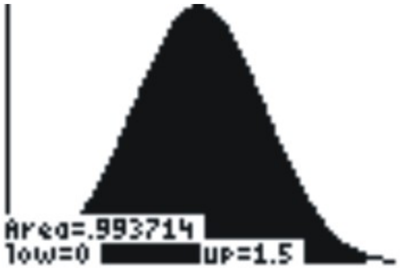
- The standard deviation of the sample mean is approximately equivalent to the following

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

These properties of the sampling distribution of the mean can be applied to determining probabilities. The sampling distribution of the sample mean can be assumed to be approximately normal, even if the population is not normally distributed. Now that it has been clarified that the sampling distribution of the mean is approximately normal, let's see how these properties work. 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 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{20}} \approx 0.27$

Using technology, a sketch of this problem is

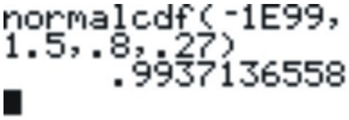


```
Area=.993714
low=0         up=1.5
```

The shaded area shows the probability that the sample mean is less than 1.5.

The $z-$ score for the value 1.5 is $z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \approx \frac{1.5-0.8}{0.27} \approx 2.6$

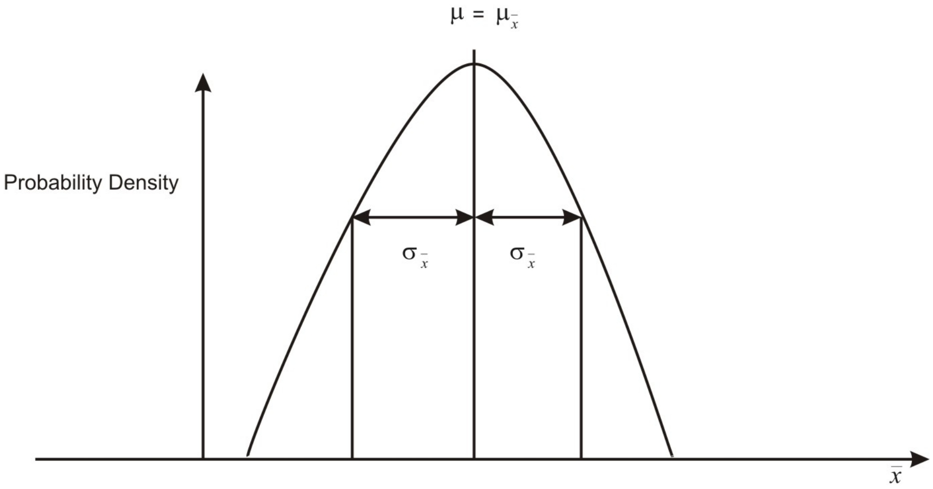
As shown above, the area under the standard normal curve to the left of 1.5 (a $z-$score of 2.6) is approximately 0.9937. This value can also be determined by using the graphing calculator



```
normalcdf(-1E99,
1.5,.8,.27)
     .9937136558
```

The probability that the sample mean will be below 1.5 is 0.9937. In a random sample of 20 families, it is almost definite that the average number of pets per family will be less than 1.5.

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



The vertical axis now reads *probability density* rather than *frequency*. Frequency can only be used when you are dealing with a finite number of sample means, as it is the number of selections divided by the total 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 the sample mean to be drawn.

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

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

c) What is the standard deviation of the sample mean?

Solution:

The question indicates that an infinite number of samples of size 40 are being collected from a known population, an infinite number of sample means are being calculated and then the sampling distribution of the sample mean is being studied. Therefore, an understanding of the Central Limit Theorem is necessary to answer the question.

a) The sampling distribution of the sample mean will be bell-shaped.

b) The sampling distribution of the sample mean will be centered about the population mean of 23.5

c)

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

A sample with a sample size of 40 is taken from a known population where $\mu = 25$ and $\sigma = 4$. The following chart displays the collected data:

24	23	30	17	24	22	23	21	29	25
26	25	29	28	29	29	32	22	27	28
24	32	21	29	30	18	21	24	30	24
25	26	25	27	26	25	27	24	24	25

a) What is the population mean?

b) Determine the sample mean using technology.

c) What is the population standard deviation?

d) Using technology, determine the sample standard deviation.

e) If an infinite number of samples of size 40 were collected from this population, what would be the value of the sample means?

f) If an infinite number of samples of size 40 were collected from this population, what would be the value of the standard deviation of the sample means?

Solution:

a) $\mu = 25$ The population mean of 25 was given in the question.

b) $\bar{x} = 25.5$ The sample mean is 25.5 and is determined by using $1-$ *Vars Stat* on the TI-83.

c) $\sigma = 4$ The population standard deviation of 4 was given in the question.

d) $S_x = 3.47$ The sample standard deviation is 3.47 and is determined by using $1-$ *Vars Stat* on the TI-83.

e) $\mu_{\bar{x}} = 25$ A property of the Central Limit Theorem.

f)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$\sigma_{\bar{x}} = \frac{4}{\sqrt{40}}$$
$$\sigma_{\bar{x}} = 0.63 \quad \text{Central Limit Theorem}$$

Lesson Summary

For approximately normal distributions, the mean and the standard deviation are used as the measure of center and spread. If these two values are known, the z-scores and the calculator can be used to find the percentage of values in any interval.

The Central Limit Theorem confirms the intuitive notion that with a large enough number of trials that are performed on a random variable, the sample means will begin to approximate a normal distribution.

Points to Consider

- What is a binomial experiment?
- What is the difference between a population proportion and a sample proportion?
- How does sample size affect the variation in sample results?

Review Questions

1. The lifetimes of a certain type of calculator battery are normally distributed. The mean lifetime is 400 days with a standard deviation of 50 days. For a sample of 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.
2. A sample with a sample size of 40 is taken from a known population where $\mu = 25$ and $\sigma = 4$ The following chart displays the collected data:

24	23	30	17	24	22	23	21	29	25
26	25	29	28	29	29	32	22	27	28
24	32	21	29	30	18	21	24	30	24
25	26	25	27	26	25	27	24	24	25

   (a) What is the population mean?

(b) Determine the sample mean using technology.
(c) What is the population standard deviation?
(d) Using technology, determine the sample standard deviation.
(e) If an infinite number of samples of size 40 were collected from this population, what would be the value the mean of the sample means?
(f) If an infinite number of samples of size 40 were collected from this population, what would be the value of the standard deviation of the sample means?

Review Answers

1. (a)

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} && \text{and} && z = \frac{x-\mu}{\sigma} \\ z &= \frac{360-400}{50} && \text{and} && z = \frac{460-400}{50} \\ z &= -0.8 && \text{and} && z = 1.2 \end{aligned}$$

Using the graphing calculator the area for $z = -0.8$ is 0.2881 and for $z = 1.2$ is 0.3849

$$\text{Area is: } 0.2881 + 0.3849 = 0.6730$$
$$(.6730)(6000) = 4038$$

This means that 67.3% of the 6000 batteries lasted between 360 and 460 days.
Note: For $z = -0.8$, the area is to the left of the mean. However, the curve is symmetrical about the mean and the value of the area for $z = 0.8$ is used and added to the area of $z = 1.2$.
(a) 4038 batteries will last between 360 and 460 days.
(b)

$$\begin{aligned} z &= \frac{x-\mu}{\sigma} \\ z &= \frac{320-400}{50} \\ z &= -1.6 \end{aligned}$$

The area for $z = 1.6$ is 0.4452.

$$0.4452 + 0.5000 = 0.9452$$
$$(.9452)(6000) = 5671$$

This means that 94.52% of the 6000 batteries lasted more than 320 days.

Note: For $z = -1.6$, the total area to the right of the mean is needed. Since the total area under the curve is one, the total area on either side of the mean is 0.5000. This area must be added to the area 0.4452

(b) 5671 batteries will last more than 320 days

(c)

$$\begin{aligned}
z &= \frac{x-\mu}{\sigma}\\
z &= \frac{280-400}{50}\\
z &= -2.4
\end{aligned}$$

The area for $z = 2.4$ is 0.4918.

$$0.5000 - 0.4918 = 0.0082$$
$$(.0082)(6000) = 49$$

This means that 0.82% of the 6000 batteries lasted less than 280 days.

Note: Since the total area to the left of $z = -2.4$ is required, the area for $z = 2.4$ is subtracted from 0.5000

(c) 49 batteries will last less than 280 days

2. (a) $\mu = 25$ The population mean of 25 was given in the question.
   (b) $\bar{x} = 25.5$ The sample mean is 25.5 and is determined by using 1− *Vars Stat* on the TI-83.
   (c) $\sigma = 4$ The population standard deviation of 4 was given in the question.
   (d) $S_x = 3.47$ The sample standard deviation is 3.47 and is determined by using 1− *Vars Stat* on the TI-83.
   (e) $\mu_{\bar{x}} = 25$ A property of the Central Limit Theorem.
   (f)

$$\begin{aligned}
\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}}\\
\sigma_{\bar{x}} &= \frac{4}{\sqrt{40}}\\
\sigma_{\bar{x}} &= 0.63 \quad \text{Central Limit Theorem}
\end{aligned}$$

Vocabulary

Central Limit Theorem An important result in statistics, stating that the shape of the sampling distribution of the sample mean becomes more normal as n increases.

Standard Normal Distribution A normal distribution with a mean of zero and a standard deviation of one.

$Z-$ **score** The variable along the horizontal axis of a normal distribution.

7.3 Binomial Distribution and Binomial Experiments

Learning Objectives

- Apply techniques to estimate the probability of a population proportion of outcomes in a survey and experiment.
- Understand the conditions needed to conduct a binomial experiment.

Introduction

A probability distribution shows the possible numerical outcomes of a chance process and from this the probability of any set of possible outcomes can be deduced. As seen in previous lessons, for many events probability distribution can be modeled by the normal curve. One type of probability distribution that is worth examining is a binomial distribution.

In probability theory and statistics, the **binomial distribution** is the discrete probability distribution of the number of successes in a sequence of "n" independent yes/no experiments, each of which yields success with probability "p" (such experiments are called Bernoulli experiments).

To conduct a **binomial experiment** a random sample (n independent trials) must be chosen, and the number of successes (x) determined. Then the sample proportion $\hat{p}$ can be found to predict the population proportion (probability of success, p).Many experiments involving random variables are simply exercises in counting the number of successes in n independent trials, such as

- The number of people with type O blood in a random sample of 10 people (a person either has type O blood or doesn't.)
- The number of doubles in eight rolls of a pair of dice (doubles either show up on each roll or they don't.)
- The number of defective light bulbs in a sample of 30 bulbs (either the bulb is defective or it isn't.)

Binomial Experiments

These events are called binomial because each one has two possible outcomes. 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. The following list of 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 7.3:

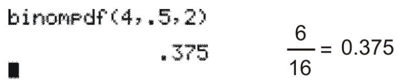
Number Who Correctly Identify Percolated Coffee	Outcomes C (correct), I (incorrect)	Number of Outcomes
0	$IIII$	1
1	$ICII$ $IIIC$ $IICI$ $CIII$	4
2	$ICCI$ $IICC$ $ICIC$ $CIIC$ $CICI$ $CCII$	6
3	$CICC$ $ICCC$ $CCCI$ $CCIC$	4
4	$CCCC$	1

Using the Multiplication Rule for Independent Events, you know that the probability of getting a certain outcome when two people guess correctly, like, $CICI$, is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$. The table shows six outcomes where two people guessed correctly so the probability of getting two people who correctly identified the percolated coffee is $\frac{6}{16}$. 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, or "4 choose 2":

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

A graphing calculator can also be used to calculate binomial probabilities.

2^{nd} **[DISTR]** ↓ 0:binompdf (This command calculates the binomial probability for k successes out of n trials when the probability of success on any one trial is p)



```
binompdf(4,.5,2)
                .375
```

$\frac{6}{16} = 0.375$

A random sample can be treated as a binomial situation if the sample size, n, is small compared to the size of the population. A rule of thumb to use here is making sure that the sample size is less than 10% of the size of the population.

A binomial experiment is a probability experiment that satisfies the following conditions:

1. Each trial can have only two outcomes – one known as "success" and the other "failure."
2. There must be a fixed number, n, of trials.
3. The outcomes of each trial must be independent of each other. The probability of each a "success" doesn't change regardless of what occurred previously.
4. The probability, p, of a success must remain the same for each trial.

The distribution of the random variable X that counts the number of successes is called a binomial distribution. The probability that you get exactly $X = k$ successes is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

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):

k	$P(X = k)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

The expected value for the above distribution is:

$$E(X) = 0(1/16) + 1(4/16) + 2(6/16) + 3(4/16) + 4(1/16)$$

$E(X) = 2$ In other words, you expect half of the four to guess correctly when given two equally, likely choices. $E(X)$ can be written as $4 \cdot \frac{1}{2}$ which is equivalent to np. For a random variable X having a binomial distribution with n trials and probability of successes p, the expected value (mean) and standard deviation for the distribution can be determined by:

$$E(X) = np = \mu_x \qquad \text{and} \qquad \sigma_x = \sqrt{np(1 - p)}$$

The graphing calculator will now be used to graph and compare binomial distributions. The 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 secondly the list of binomial probabilities.

Sequence of integers:

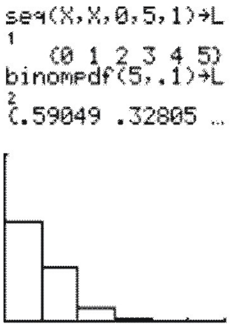
2^{nd} **[LIST]** → OPS ↓ 5.seq ( ) STO → 2^{nd} 1

Binomial Probabilities:

2^{nd} DISTR 0:binompdf(

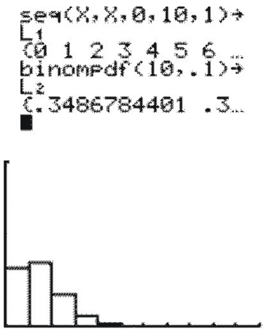
Horizontally, the following are examples of binomial distributions where n increases and p remains constant. Vertically, the examples display the results where n remains fixed and p increases.

$n = 5$ and $p = 0.1$



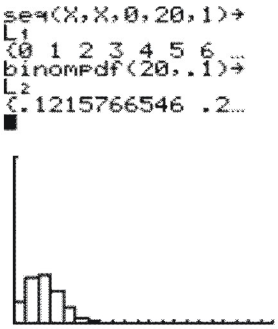
```
seq(X,X,0,5,1)→L
1
   {0 1 2 3 4 5}
binompdf(5,.1)→L
2
{.59049 .32805 …
```

$n = 10$ and $p = 0.1$



```
seq(X,X,0,10,1)→
L1
{0 1 2 3 4 5 6 …
binompdf(10,.1)→
L2
{.3486784401 .3…
■
```

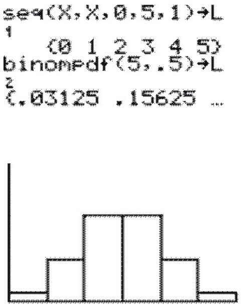
$n = 20$ and $p = 0.1$



```
seq(X,X,0,20,1)→
L1
{0 1 2 3 4 5 6 …
binompdf(20,.1)→
L2
{.1215766546 .2…
■
```

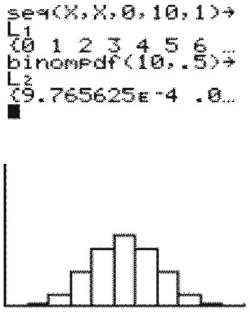
For the small value of p, the binomial distributions are skewed toward the higher values of x. As n increases, the skewness decreases and the distributions gradually move toward being more normal.

$n = 5$ and $p = 0.5$



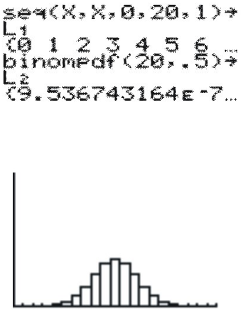
```
seq(X,X,0,5,1)→L
1
   {0 1 2 3 4 5}
binompdf(5,.5)→L
2
{.03125 .15625 …
```

$n = 10$ and $p = 0.5$



```
seq(X,X,0,10,1)→
L1
{0 1 2 3 4 5 6 …
binompdf(10,.5)→
L2
{9.765625E-4 .0…
■
```

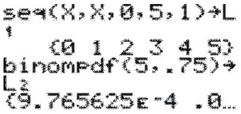
$n = 20$ and $p = 0.5$



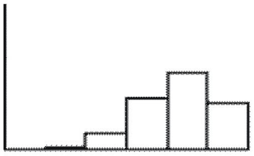
```
seq(X,X,0,20,1)→
L1
{0 1 2 3 4 5 6 …
binompdf(20,.5)→
L2
{9.536743164E-7…
```

As p increases to 0.5, the skewness disappeared and the distributions achieved perfect symmetry. The symmetrical, mound-shaped distribution remained the same for all values of n.

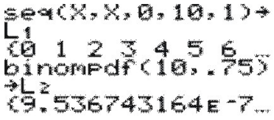
$n = 5$ and $p = 0.75$



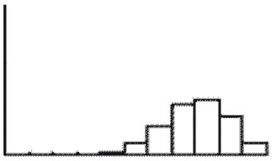
```
seq(X,X,0,5,1)→L
1
   {0 1 2 3 4 5}
binompdf(5,.75)→
L2
{9.765625E-4 .0…
```



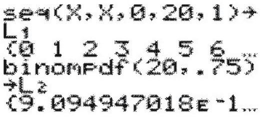
$n = 10$ and $p = 0.75$



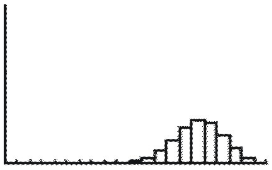
```
seq(X,X,0,10,1)→
L1
{0 1 2 3 4 5 6…
binompdf(10,.75)
→L2
{9.536743164E-7…
```



$n = 20$ and $p = 0.75$



```
seq(X,X,0,20,1)→
L1
{0 1 2 3 4 5 6…
binompdf(20,.75)
→L2
{9.094947018E-1…
```



For the 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 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.

Example: Suppose you flip a fair coin 12 times.

a) What is the probability that you will get exactly 5 heads?

b) Exactly 25% heads?

c) At least 10 heads?

Solution: Let X represent the number of heads from 12 flips of the coin. Exactly 5 heads:

a)

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ \binom{12}{5} &= \frac{12!}{5!(12-5)!} \\ \binom{12}{5} &= 792\end{aligned} \qquad P(X = 5) = 792(0.5)^{12} = 0.1933$$

b) 25% of 12 heads means getting 3 heads.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad P(X=3) = 220(0.5)^{12} = 0.0537$$
$$\binom{12}{3} = \frac{12!}{3!(12-3)!}$$
$$\binom{12}{3} = 220$$

c) At least 10 heads means getting 10 heads, 11 heads or 12 heads.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{12}{10} = \frac{12!}{10!(12-10)!} \qquad \binom{12}{10} = 66$$
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{12}{11} = \frac{12!}{11!(12-11)!} \qquad \binom{12}{11} = 12$$
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{12}{12} = \frac{12!}{12!(12-12)!} \qquad \binom{12}{12} = 1$$

$$P(X=10) = 66(0.5)^{12} \approx 0.0161$$
$$P(X=11) = 12(0.5)^{12} \approx 0.0029$$
$$P(X=12) = 1(0.5)^{12} \approx 0.0002$$

At least 10 heads: $0.0161 + 0.0029 + 0.0002 \approx 0.0192$

Example:

Since the closing of the two main industries, a small town has experienced a decrease in the population of teenagers. According to the statistics of the local high school, approximately 7.6% of teens ages $14-19$ are no longer registered. Suppose you randomly chose four people who were enrolled at the high school before the population dropped.

a) What is the probability that none of the four are registered?

b) What is the probability that at least one is not registered?

c) Create a probability distribution table for this situation.

Solution:

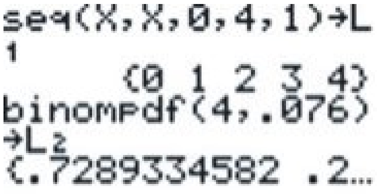
a)

$$P(X = k) = \binom{n}{k} P^k (1 - p)^{n-k}$$
$$P(X = 0) \binom{4}{0} (0.076)^0 (1 - 0.076)^4 = 0.7289$$

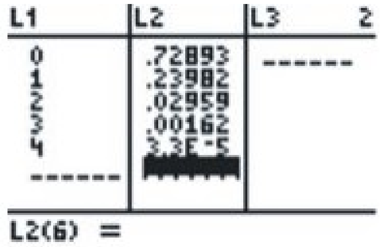
b)

$$P(X \geq 1) = 1 - [P(X = 0)]$$
$$P(X \geq 1) \approx 1 - 0.7289 = 0.2711$$

c) Using technology:



```
seq(X,X,0,4,1)→L
1
      {0 1 2 3 4}
binompdf(4,.076)
→L2
{.7289334582 .2…
```



```
L1      L2       L3     2
0       .72893   ------
1       .23982
2       .02959
3       .00162
4       3.3E-5
------
L2(6) =
```

If you so desire, you can transfer this data into your own table:

Table 7.4:

Number Not Registered	Probability
0	0.72893
1	0.23982
2	0.02959
3	0.00162
4	0.00003

Lesson Summary

In this lesson you have learned that the random variable X has a binomial distribution if X represents the number of "successes" in n independent trials. In each of these trials, the probability of success is p. You have also learned that a binomial distribution has these important features:

- The probability of getting exactly $X = k$ successes is given by the formula

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The mean or expected value is represented by

$$E(X) = np = \mu_x$$

- The standard deviation is

$$\sigma_x = \sqrt{np(1-p)}$$

Review Questions

1. 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?

Review Answers

1. (a) The operative word in this problem is 'median'; if the median income is $56,400, then this indicates that one-half of the income falls below $56,400 and one-half of the income lies above it. Therefore, the chance of a randomly selected income being below the median income is 0.5. Let X represent the number of households with incomes below the median in a random sample of size 24. X has a binomial distribution with $n = 24$ and $p = 0.5$.

$$\begin{aligned} E(X) &= np = \mu_x \\ E(X) &= (24)(0.5) = 12 \end{aligned}$$

   (b) The standard deviation of households with incomes less than $56,400 is

$$\begin{aligned} \sigma_x &= \sqrt{np(1-p)} \\ \sigma_x &= \sqrt{12(1-0.5)} \\ \sigma_x &= 2.25 \end{aligned}$$

(c) $P(X \geq 18) \approx$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$
$$\binom{24}{18} = \frac{24!}{18!(24-18)!} \qquad \binom{24}{19} = \frac{24!}{19!(24-19)!} \qquad \binom{24}{20} = \frac{24!}{20!(24-20)!}$$
$$\binom{24}{18} = 134596 \qquad \binom{24}{19} = 42504 \qquad \binom{24}{20} = 10626$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$
$$\binom{24}{21} = \frac{24!}{21!(24-21)!} \qquad \binom{24}{22} = \frac{24!}{22!(24-22)!} \qquad \binom{24}{23} = \frac{24!}{23!(24-23)!}$$
$$\binom{24}{21} = 2024 \qquad \binom{24}{22} = 276 \qquad \binom{24}{23} = 24$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad P(X = 18) = 134596(0.5)^{24} \approx 0.0080$$
$$\binom{24}{24} = \frac{24!}{24!(24-24)!} \qquad P(X = 19) = 42504(0.5)^{24} \approx 0.0025$$
$$\binom{24}{24} = 1 \qquad P(X = 20) = 10626(0.5)^{24} \approx 0.0006$$
$$P(X = 21) = 2024(0.5)^{24} \approx 0.0001$$
$$P(X = 22) = 276(0.5)^{24} \approx 0.000016$$
$$P(X = 23) = 24(0.5)^{24} \approx 0.0000014$$
$$P(X = 24) = 1(0.5)^{24} \approx 0.00000006$$

The sum of these probabilities gives the answer to the question: 0.01121746.

7.4 Confidence Intervals

Learning Objectives

- Calculate the point estimate of a sample to estimate a population proportion.
- Construct a confidence interval for a population proportion based on a sample population.
- Calculate the margin of error for proportions as a function of sample proportion and size.

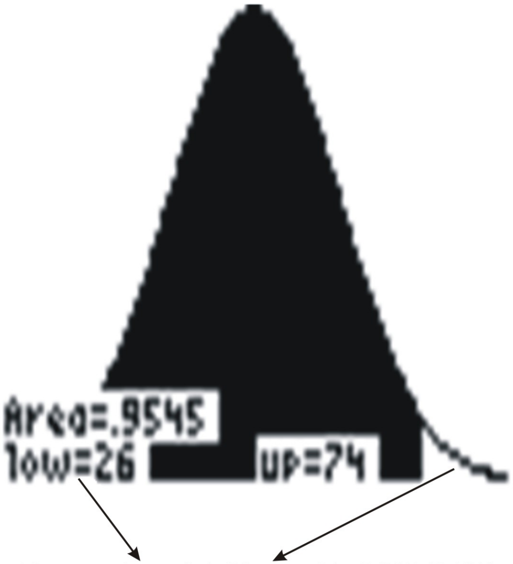
- Understand the logic of confidence intervals as well as the meaning of confidence level and confidence intervals.

Introduction

The objective of inferential statistics is to use sample data to increase knowledge about the corresponding entire population. 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 called a point estimate because this single number is used as a plausible value of the population mean. 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 p might take, controlling the probability that μ 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.

Normal distribution specifies that 68 percent of data will fall within one standard error of the parameter. This logic can be turned around to state that any single random sample has a 68 percent chance of falling within that range. Likewise, we may say that we are confident that in 95 percent of samples, sample statistics are within plus or minus two standard errors of the population parameter. As the confidence interval is expanded for a given statistic, 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 of the normal density curve. For example, a 95% confidence interval covers 95% of the normal curve – the probability of observing a value outside of this area is less than 5%. Because the normal curve is symmetric, half of the area is in the left tail of the curve, and the other half of the area is in the right tail of the curve. This means that 2.5% of the area is in each tail.



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

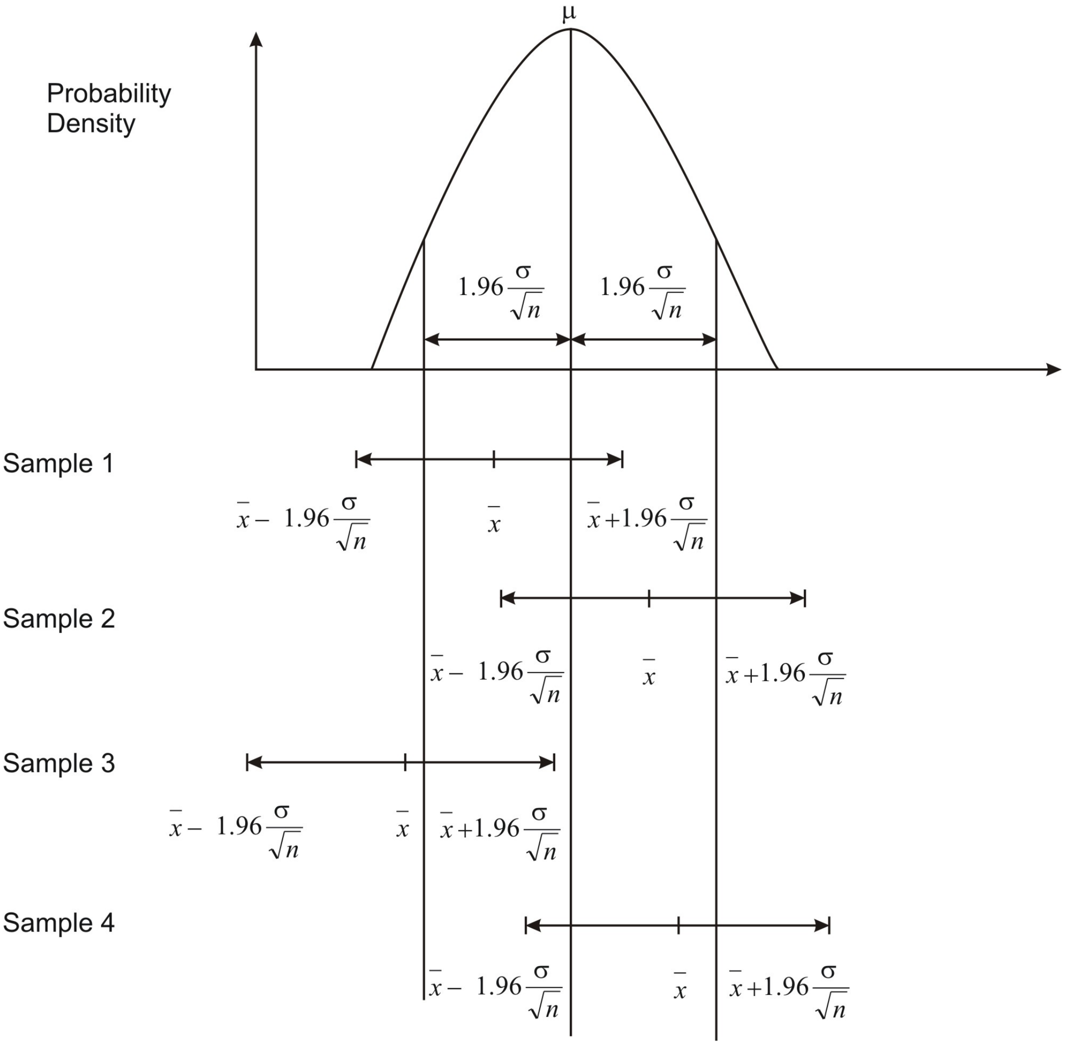
This graph was made using the TI-83 and shows a normal distribution curve for a set of data that has a mean of ($\mu = 50$) and a standard deviation of ($\sigma = 12$). A 95% confidence interval for the standard normal distribution, then, is the interval $(-1.96, 1.96)$, since 95% of the area under the 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 proportion that is added and subtracted from the mean to construct the confidence interval. For a 95% confidence interval, the margin of error equals $\pm 1.96\ \sigma$

The following example will help you to understand these terms and their meaning.

Example:

Jenny randomly selected 60 muffins from one company line 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 (interval estimator). Jenny reported that the number of grams of fat in each muffin is between 10.3 grams and 11.2 grams with 95% confidence.

The population mean refers to the unknown population mean. 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 known, 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 *does not* mean that there is a 0.95 probability that the population mean falls in the interval from 10.3 to 12.1. 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 capture the true population mean. If you formed 100 intervals in this manner, 95% of them would contain the population mean.


μ
Probability Density
$1.96\frac{\sigma}{\sqrt{n}}$
$1.96\frac{\sigma}{\sqrt{n}}$
Sample 1
$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$
Sample 2
$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$
Sample 3
$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$
Sample 4
$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

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?

Solution:

a) The sample mean ($\bar{x}$) for Sample 3 is not within $1.96\frac{\sigma}{\sqrt{n}}$ of the population mean. It does not fall within the two vertical lines on the left and right of the sampling distribution of the sample mean.

b) The confidence interval for Sample 3 does not enclose the population mean. This interval is just to the left of the population mean (μ), which is labeled as the vertical line found in the middle of the sampling distribution of the sample mean.

c) 95%

d) 95%

When the sample size is large ($n \geq 30$), the confidence interval for the population mean is calculated as shown below:

$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ where z is 1.96 for a 95% confidence interval; 1.645 for a 90% confidence interval and 2.56 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	112	109	115

a) For each sample, determine the 90% confidence interval?

b) Do all four confidence intervals enclose the population mean? Explain.

Solution:

a)

$$\begin{aligned} &\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \\ &107 \pm 1.645\frac{19}{\sqrt{60}} \\ &107 \pm 4.04 \\ &\text{from } 102.96 \text{ to } 111.04 \end{aligned}$$

$$\begin{aligned} &\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \\ &112 \pm 1.645\frac{19}{\sqrt{60}} \\ &112 \pm 4.04 \\ &\text{from } 107.96 \text{ to } 116.04 \end{aligned}$$

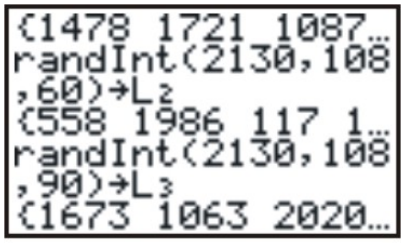
$$\begin{aligned} &\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \\ &109 \pm 1.645\frac{19}{\sqrt{60}} \\ &109 \pm 4.04 \\ &\text{from } 104.96 \text{ to } 113.04 \end{aligned}$$

$$\begin{aligned} &\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \\ &115 \pm 1.645\frac{19}{\sqrt{60}} \\ &115 \pm 4.04 \\ &\text{from } 110.96 \text{ to } 119.04 \end{aligned}$$

b) Three of the confidence intervals enclose the population mean. The interval from 110.96 to 119.04 do not enclose the population mean.

Example:

Now it is time to use the graphing calculator to simulate the collection of three samples of different sizes –30, 60, 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.



```
{1478 1721 1087…
randInt(2130,108
,60)→L₂
{558 1986 117 1…
randInt(2130,108
,90)→L₃
{1673 1063 2020…
```

Math → PBR → 5. randInt(

randInt(μ, σ, n) store in L_1 Sample size $= 30$

randInt(μ, σ, n) store in L_2 Sample size $= 60$

randInt(μ, σ, n) store in L_3 Sample size $= 90$

The lists of numbers can be viewed by **[Stat]** enter. The next step is to calculate the mean of each of these samples.

[List] → **[Math]** → mean(L_1) 1309.6 Repeat this for (L_2)1171.1 and (L_3)1077.1.

The three confidence intervals are:

$$\begin{array}{lll}
\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}} & \bar{x} \pm z\dfrac{\sigma}{\sqrt{n}} & \bar{x} \pm z\dfrac{\sigma}{\sqrt{n}} \\
1309.6 \pm 1.96\dfrac{108}{\sqrt{30}} & 1171.1 \pm 1.96\dfrac{108}{\sqrt{60}} & 1077.1 \pm 1.96\dfrac{108}{\sqrt{90}} \\
1309.6 \pm 38.65 & 1171.1 \pm 27.33 & 1077.1 \pm 22.31 \\
\text{from } 1270.95 \text{ to } 1348.25 & \text{from } 1143.77 \text{ to } 1198.43 & \text{from } 1054.79 \text{ to } 1099.41
\end{array}$$

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.

In all of the examples shown above, you calculated the confidence intervals for the population mean using the formula $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$. However, to use this formula, the population standard deviation (σ) had to be known in order to calculate the interval. If this value is unknown and if the sample size is large ($n \geq 30$), the population standard deviation can be replaced with the sample standard deviation. Thus, the formula $\bar{x} \pm z\frac{S_x}{\sqrt{n}}$ can be used as an interval estimator. An interval estimator of the population mean is called a confidence interval. This formula is valid only for simple random samples. Since $z\frac{S_x}{\sqrt{n}}$ is actually 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, Grade 12 students would score an average grade of 65% with accuracy within 3%, 9 times out of 10.

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.

Solution:

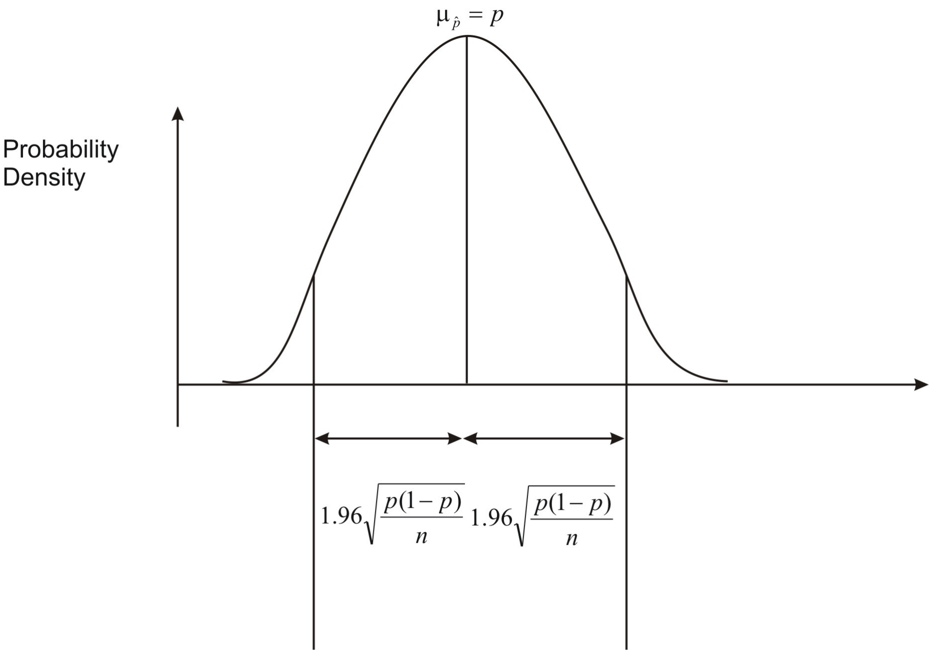
a) You are dealing with a 90% confidence level. This is indicated by 9 times out of 10.

b) The margin of error is 3%.

c) The confidence interval is $\bar{x}\pm$ the margin of error which is 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)

The calculation of a confidence interval for a population proportion is similar to that explained above for a sample mean. For a confidence interval of 95%, the sampling distribution of the sample proportions is approximately normal with large sample sizes ($n \geq 30$). From this statement you can say that 95% of the sample proportions from a population are within two standard deviations (more accurately 1.96 standard deviations) of the population proportion. This is shown in the diagram below:



Therefore, if a single sample proportion is within $1.96\sqrt{\frac{p(1-p)}{n}}$ of the population proportion, then the interval $\hat{p}-1.96\sqrt{\frac{p(1-p)}{n}}$ to $\hat{p}+1.96\sqrt{\frac{p(1-p)}{n}}$ will capture the population proportion. This will happen for 95% of all possible samples. If you look at the above formulas, you should notice that the population proportion (p) and the sample proportion ($\hat{p}$) are both used to calculate the confidence interval. However, in real-life situations, the population proportion is seldom known. Therefore,(p) is most often replaced with ($\hat{p}$) in the formulas above so that they now become:

$\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ or in a more standard form $p \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ There are two restrictions that apply to this formula: 1) $np \geq 5$ and 2) $n(1-p) \geq 5$.

As before, the margin of error is $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and the confidence interval is $\hat{p}\pm$ the margin of error.

Example:

A large grocery store has been recording data regarding the number of shoppers that use savings coupons at their outlet. Last year it was reported that 77% of all shoppers used coupons, and these results were considered accurate within 2.9%, 19 times out of 20.

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.

Solution:

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 2.9%.

c) The confidence interval is simply $\hat{p}\pm$ the margin of error.

$$77\% - 2.9\% = 74.1\% \qquad\qquad 77\% + 2.9\% = 79.9\%$$

The confidence interval is from 74.1% to 79.9%.

d) The 95% confidence interval from 74.1% to 79.9% 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.

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 $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$. The values of z are 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.

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?

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 6250 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 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.

Review Answers

1. (a)

$$\begin{aligned} \hat{p} &= \frac{x}{n} & \qquad \hat{p} &\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ \hat{p} &= \frac{142}{250} & \qquad 0.568 &\pm 2.56\sqrt{\frac{0.568(1-0.568)}{250}} \\ \hat{p} &= 0.568 & \qquad 0.568 &\pm 0.080 \end{aligned}$$

The interval is from 0.488 **to** 0.648 **OR from** 48.8% **to** 64.8%

(b) The 99% confidence interval could be narrowed by increasing the sample size from 250 to a larger number.

2. (a)

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$0.61 \pm 1.645\sqrt{\frac{0.61(1-0.61)}{300}}$$
from 0.564 to 0.656

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$0.61 \pm 2.56\sqrt{\frac{0.61(1-0.61)}{300}}$$
from 0.555 to 0.665

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$0.61 \pm 1.96\sqrt{\frac{0.61(1-0.61)}{300}}$$
from 0.538 to 0.682

(b) The confidence interval got wider as the confidence level increased. To increase the probability of enclosing the population proportion a wider confidence interval must be chosen.

(c) Yes, all three confidence intervals would capture the population proportion if it were 0.58.

Vocabulary

Binomial Experiment A type of survey or experiment in which there is a fixed number of trials that have one of only two outcomes. The probability of success for any trial is equal to the population proportion and remains the same for every trial. The outcomes for each trial are independent of one another and the binomial random variable, x, is the number of successes observed in n trials.

Confidence Interval An interval of plausible values for a population parameter. Any of the values in the interval could be used to define a population for which the defined sample statistic would be a likely outcome.

Confidence Level The probability that the method used to calculate the confidence interval will produce an interval that will enclose the population parameter.

Interval Estimator Another name for a confidence interval

Point Estimate A single number, such as a single sample mean, that is used as a plausible value of the population mean. This can also be another single value that represents a population parameter.

Population Proportion A fraction of the population that possesses a certain characteristic or probability of an event occurring. The characteristic or event is called a success.

Sample Proportion The ratio of successes x to sample size n.

7.5 Sums and Differences of Independent Random Variables

Learning Objectives

- Construct probability distributions of random variables.
- Calculate the mean and standard deviation for sums and differences of independent random variables.

Introduction

In previous lessons, you learned that sampling is a way of estimating a parameter of a population by selecting data from that population and a way of computing the chance of obtaining a specified outcome from a sample given a specific population. In Fort McMurray, Alberta, better known as "Fort McMoney" because of the oil industry, housing is becoming a rare commodity. Suppose you are a recent winner of a lottery and you decide to invest your winnings in a housing project for this city. Your plan is to build 500 single-family homes. Before you begin, there are some facts that you need to know so that the houses will be a quick sell and you can turn a profit for your investment. One bit of information that you would like to ascertain is how many televisions each family will have so you will know how many cable hook-ups to install in each household. Short of conducting a survey, how will you determine a solution to this problem?

Probability Distributions from Data

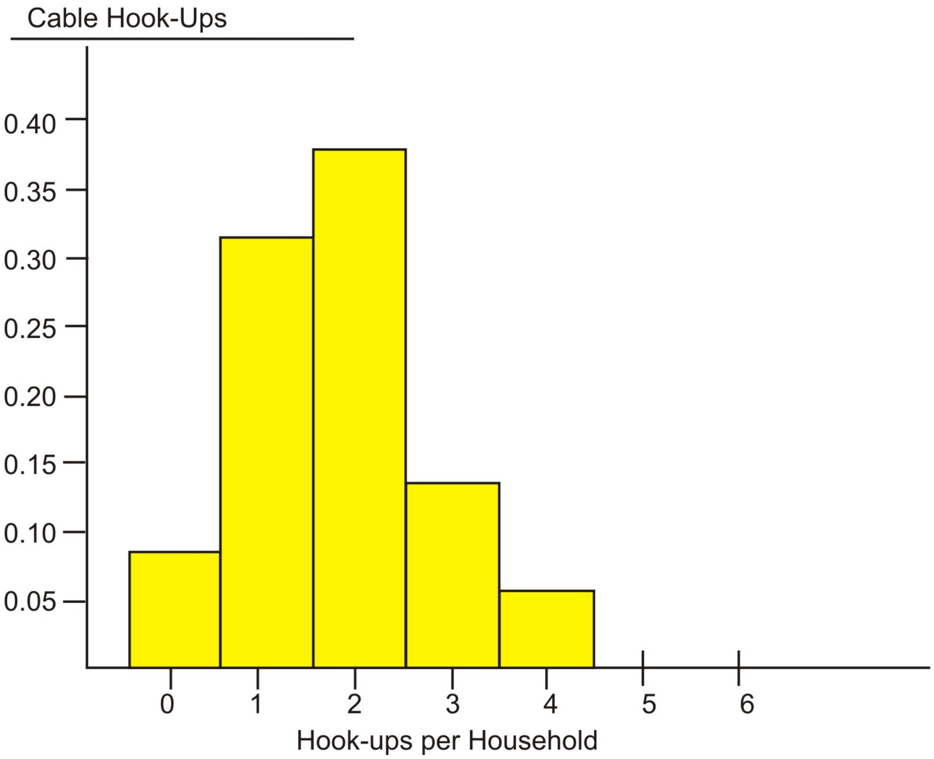
To begin, you should contact the local cable provider in the city and ask them to provide you with a record of the distribution of cable hook-ups per household that are currently on their books. Suppose they release this data:

Table 7.5:

Hook-ups per Household	Proportion of Households
0	0.092

Table 7.5: (continued)

Hook-ups per Household	Proportion of Households
1	0.328
2	0.380
3	0.142
4	0.058



From the above table and histogram, you now have some estimates with which to work. The ′4′ represents '4 or more' but there are only 0.058 or 5.8% of the houses that fall in this category. As well, if you add $0.380 + 0.142 + 0.058$, you will see that a little more than 50%(0.580) of the households will have 2 or more cable hook-ups. The above distribution is skewed toward the larger numbers and has a mean of approximately 1.746. By using probability from samples with known distributions, you now have some data for the unknown population.

The data above can also be shown in another way to display numbers that represent the possible number of hook-ups in each household. This display is done by using a list of random digits to select a household at random from this distribution. The numbers must be three-digit numbers because the percentages all have three decimal places.

Table 7.6:

Number of Hook-ups per Household	Proportions of Households	Random Numbers Representing this Category
0	0.092	001 – 092
1	0.328	093 – 329
2	0.380	330 – 382
3	0.142	383 – 524
4	0.058	525 – 582

A three-digit random number does not represent an individual household, but with other three-digit numbers in its category, it represents the many households in the category.

From this display, you can see that 92 households out of 1000 require no cable hook-up while 328 out of 1000 require one cable hook-up. This way of displaying the data allows you to see actual numbers for each household.

Probability distribution is the set of values that a random variable takes on. Its value depends upon the result of a trial. The random variable, X, will represent the number of cable hook-ups in a randomly selected household. Therefore, $P(x = 2) \approx 0.380$, because approximately 38.0% of the randomly selected households require two cable hook-ups.

At this time, there are two ways that you can create probability distributions from data. Sometimes previously collected data, relative to the random variable that you are studying, can serve as a probability distribution. This was the case with the data received from the local cable company in Fort McMurray. 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 basic principles and assumptions by using the rules of theoretical probability. The following examples will lead to the understanding of these rules of theoretical probability.

Example:

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

			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 die are rolled simultaneously can now be used to construct other probability distributions. The first table will display the sum of the two die and the second will represent the larger of the two numbers.

Table 7.7:

Sum of Two Die, x	Probability, p
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36
Total	1

Table 7.8:

Larger Number, x	Probability, p
1	1/36
2	3/36
3	5/36
4	7/36
5	9/36
6	11/36

Table 7.8: (continued)

Larger Number, x	Probability, p
Total	1

When you roll the two die, what is the probability that the sum of the two die is 4? The probability that the sum of the two die is four is $\frac{3}{36}$.

What is the probability that the larger number is 4? The probability that the larger number is four 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 the large population with the disease.

Table 7.9:

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

Table 7.10:

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

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 applying this rule:

$$\begin{aligned}
P(\text{no for } 1^{st}) \cdot P(\text{no for } 2nd) &= (0.2)(0.2) = 0.04\\
P(\text{yes for } 1^{st}) \cdot P(\text{no for } 2nd) &= (0.8)(0.2) = 0.16\\
P(\text{no for } 1^{st}) \cdot P(\text{yes for } 2nd) &= (0.2)(0.8) = 0.16\\
P(\text{yes for } 1^{st}) \cdot P(\text{yes for } 2nd) &= (0.8)(0.8) = 0.64
\end{aligned}$$

If X represents the number of mine workers 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 of X is given in the table below:

Table 7.11:

x	Probability of x
0	0.04
1	$0.16 + 0.16 = 0.32$
2	0.64

These probabilities are added because the outcomes are disjoint.

Example:

The Quebec Junior Major Hockey League has five teams from the Maritime Provinces. These teams are Cape Breton Screaming Eagles, Halifax Mooseheads, PEI Rockets, Moncton Wildcats and Saint John Sea Dogs. Each team has its own hometown arena and each arena has a seating capacity that is listed below:

Table 7.12:

Team	Seating Capacity (Thousands)
Screaming Eagles	5
Mooseheads	10
Rockets	4
Wildcats	7
Sea Dogs	6

A schedule can now be drawn up for the teams to play pre-season exhibition games. One game will be played in each home arena and the possible capacity attendance will also be calculated. In addition, the probability of the total possible attendance being at least $12,000$ people will also be calculated.

The number of possible combinations of two teams from these five is 10. $({}_5C_2)$. The following table shows the possible attendance for each of the pre-season, exhibition games.

Table 7.13:

Teams	Combined Attendance Capacity for Both Games (Thousands)
Eagles/Mooseheads	$5 + 10 = 15$
Eagles/Rockets	$5 + 4 = 9$
Eagles/Wildcats	$5 + 7 = 12$
Eagles/Sea Dogs	$5 + 6 = 11$
Mooseheads/Rockets	$10 + 4 = 14$
Mooseheads/Wildcats	$10 + 7 = 17$
Mooseheads/Sea Dogs	$10 + 6 = 16$
Rockets/Wildcats	$4 + 7 = 11$
Rockets/Sea Dog	$4 + 6 = 10$
Sea Dogs/Wildcats	$6 + 7 = 13$

The last calculation is to determine the probability distribution of the capacity attendance.

Table 7.14:

Capacity Attendance, x	Probability, p
9	0.1
10	0.1
11	0.2
12	0.1
13	0.1
14	0.1
15	0.1
16	0.1
17	0.1

The probability that the capacity attendance will be at least $12,000$ is $0.6(0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1)$

Expected Values and Standard Deviation

Returning to the original problem of the number of cable hook-ups for the 500 single-family homes, take another look at figures one and two. From these displays, you can find the mean number of hook-ups per household. You expect 9.2% of households to have no hook-up, 32.8% to have one hook-up, 38.0% to have two hook-ups, 14.2% to have three hook-ups and 5.8% to have four hook-ups. To calculate the mean number of hook-ups per household, use the previous figure and add another column.

Table 7.15:

Hook-ups per Household, x	Proportion of Households, p	Contribution to Mean, $x \cdot p$
0	0.092	0
1	0.328	0.328
2	0.380	0.760
3	0.142	0.426
4	0.058	0.232
	Sum ⟶	1.746

The mean of a probability distribution for the random variable X is denoted by μ_x or $E(X)$ which represents *expected value.* Since you now know the expected number of cable hook-ups for each household, you can also calculate how much each household will differ from this mean. In other words, you can calculate the expected standard deviation. To do this, simply determine the expected value of the square of the deviations from the mean. As you recall from chapter 1, this value is called the variance of the probability distribution, and gives a representation of how far an actual value will in general stray from this mean.

$$\begin{aligned}\sigma^2{}_x =&(0-1.746)^2(0.092)+(1-1.746)^2(0.328)+(2-1.746)^2(0.380)\\&+(3-1.746)^2(0.142)+(4-1.746)^2(0.058)\\\approx&1.0054\end{aligned}$$

The standard deviation (σ_x) is $\sqrt{1.0054} \approx 1.002$. This indicates that each household will have 1.746 cable hook-ups and differ from this mean by an average of about 1 hook-up. These calculations yield the following formulas for calculating the expected value and the standard deviation (and its square, the variance) for a probability distribution.

$$E(X)=\mu_x=\sum x_i p_i \qquad \text{and} \qquad \begin{aligned}Var(X)=\sigma^2{}_x=\sum(x_i-\mu_x)^2p_i\\ \sigma_x=\sqrt{\text{var}(X)}\end{aligned}$$

where p_i is the probability of the random variable X produced when x takes on a specific value x_i.

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 7.16:

Winnings, x	Probability, p
−$2.00	0.30
$0.00	0.40
$6.00	0.20
$20.00	0.10

Solution:

Now use the table to calculate the expected value and the variance of this distribution.

$$\begin{aligned}
&\mu_x = \sum x_i p_i \\
&\mu_x = (-2 \cdot 0.30) + (0 \cdot 0.40) + (6 \cdot 0.20) + (20 \cdot 0.10) \\
&\mu_x = 2.6
\end{aligned}$$

The player can expect to win $2.60 playing this game.

The variance of this distribution is:

$$\begin{aligned}
&{\sigma^2}_x = \sum x_i - {\mu_x}^2 p_i \\
&{\sigma^2}_x = (-2 - 2.6)^2(0.30) + (0 - 2.6)^2(0.40) + (6 - 2.6)^2(0.20) + (20 - 2.6)^2(0.10) \\
&{\sigma^2}_x \approx 41.64
\end{aligned}$$

So the standard deviation, σ_x, is approximately $\sqrt{41.64} \approx \$6.46$

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 7.17:

Number of Meals Purchased Within 12 Months, x	Probability, p
0	0.04
$[1-6)$	0.30
$[6-11)$	0.29
$[11-21)$	0.17
$[21-51)$	0.15
> 50	0.05
Total	1.00

The mean for each interval is in the center of each interval, so you must begin by estimating a mean for each interval. For the first interval of $[1-6)$, six is not included in this interval so a value of 3 would be the center. This same procedure will be used to estimate the mean of all the intervals. Therefore the expected value is:

Solution:

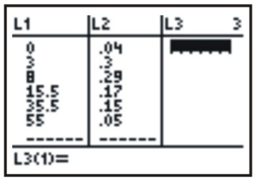
$$\begin{aligned}\mu_x &= \sum x_i p_i \\ \mu_x &= 0(0.04) + 3(0.30) + 8(0.29) + 15.5(0.17) + 35.5(0.15) + 55(0.05) \\ \mu_x &= 13.93\end{aligned}$$

And

$$\begin{aligned}{\sigma^2}_x &= \sum (x_i - \mu_x)^2 p_i \\ &= (0-13.93)^2(0.04) + (3-13.93)^2(0.30) \\ &\quad + (8-13.93)^2(0.29) + (15.5-13.93)^2(0.17) \\ &\quad + (35.5-13.93)^2(0.15) + (55-13.93)^2(0.05) \\ &\approx 208.3451 \text{ and } \sigma_x \approx 14.43\end{aligned}$$

The expected number of fast food meals purchased by a student at the local university is 13.93. This number should not be rounded since the mean 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 long tails toward the larger numbers.

Stat enter

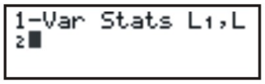


```
L1      L2      L3     3
0       .04
3       .3
8       .29
15.5    .17
35.5    .15
55      .05
------  ------
L3(1)=
```

2nd Mode Clear

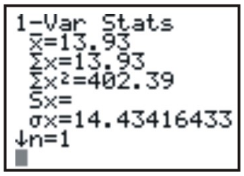
(returns to home screen)

Stat → Calc enter



```
1-Var Stats L1,L
2
```

enter



```
1-Var Stats
 x̄=13.93
 Σx=13.93
 Σx²=402.39
 Sx=
 σx=14.43416433
↓n=1
```

Notice that $\bar{x} = 13.93$ and $\sigma_x = 14.43$.

Linear Transformations of X on Mean of x and Standard Deviation of x

If you add the same number to all values of a data set, the shape or standard deviation of the data remains the same but the value is added to the mean. This is referred to as recentering the data set. Likewise, if you rescale the data – multiply all 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. The standard deviation must be multiplied by the absolute value of the number. If you multiply the mean and the standard deviation by a constant d and then add a constant c, then the mean and the standard deviation of the transformed values are expressed as:

$$\mu_{c+dx} = c + d\mu_x$$
$$\sigma_{c+dx} = |d|\sigma_x$$

The implications of these can be better understood if you return to example 1.

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?

Solution:

Recall that the expected value was \$2.60 and the standard deviation was \$6.46. 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(\$6.46), or \$19.38. Here $c = 0$ and $d = 3$. Another method of calculating the expected value would be to create a new table for the tripled prize:

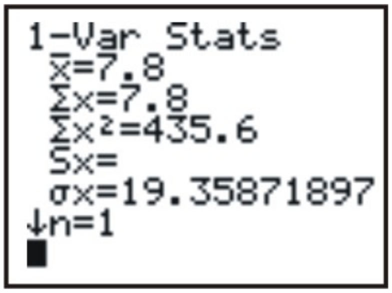
Table 7.18:

Winnings, x	Probability, p
−\$2.00	0.30
\$0.00	0.40
\$6.00	0.20
\$20.00	0.10

New Table

Table 7.19:

Original Winnings, x	New Winnings, $3x$	Probability, p
−\$2.00	−\$6.00	0.30
\$0.00	\$0.00	0.40
\$6.00	\$18.00	0.20
\$20.00	\$60.00	0.10



```
1-Var Stats
 x̄=7.8
 Σx=7.8
 Σx²=435.6
 Sx=
 σx=19.35871897
↓n=1
█
```

The calculations can be done using the formulas or by using the graphing calculator.

Using the graphing calculator:

Notice that the same results are obtained.

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

Solution:

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: y

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

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 $X + Y + Z$ is:

$$\sigma^2{}_{X+Y+Z} = \sigma^2{}_X + \sigma^2{}_Y + \sigma^2{}_Z$$
$$\sigma^2{}_{X+Y+Z} = 6.46^2 + 6.46^2 + 6.46^2$$
$$\sigma^2{}_{X+Y+Z} \approx 125.1948 \quad \text{and} \quad \sigma \approx \sqrt{125.1948} \approx 11.19$$

The person playing the three games expects to win \$7.80 with a standard deviation of \$11.19. When the prize was tripled, there was a greater standard deviation (\$19.36) than when the person played three games (\$11.19).

The rules for addition and subtraction for random variables are:

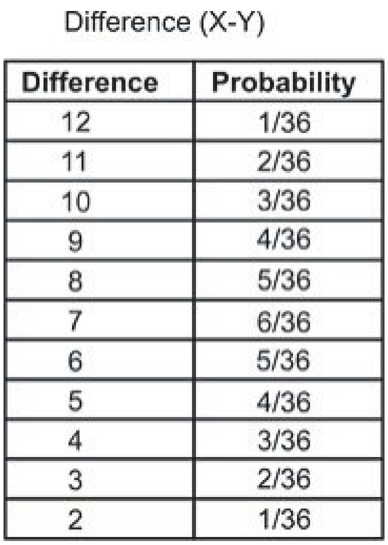
If X and Y are random variables then:

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

If X and Y are independent then:

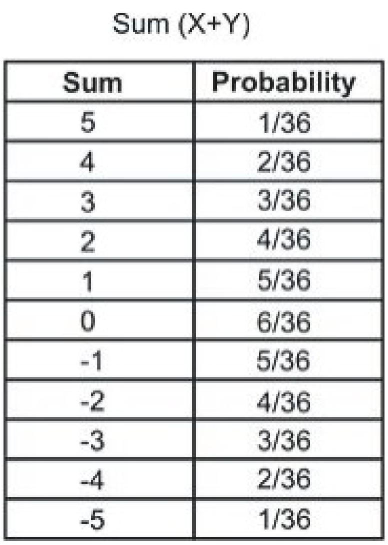
$$\sigma^2{}_{X+Y} = \sigma^2{}_X + \sigma^2{}_Y$$
$$\sigma^2{}_{X-Y} = \sigma^2{}_X + \sigma^2{}_Y$$

Variances are added for both the sum *and* difference of two independent random variables because the variation in each variable contributes to the variation in each case. Subtracting is the same as adding the opposite. Suppose you have two dice, one die (X) with the normal positive numbers 1 through 6, and another (Y) with the negative numbers -1 through -6. Then 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 (X) and then the second die (Y) 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

$$\begin{aligned} \mu_x &= \sum x_i p_i \\ \mu_X &= 3.5 \end{aligned} \qquad \begin{aligned} \mu_y &= \sum x_i p_i \\ \mu_Y &= -3.5 \end{aligned}$$

$$\begin{aligned} {\sigma^2}_x &\approx \sum (x_i - \mu_x)^2 p_i \\ {\sigma^2}_x &\approx 2.917 \end{aligned} \qquad \begin{aligned} {\sigma^2}_y &\approx \sum (x_i - \mu_y)^2 p_i \\ {\sigma^2}_y &\approx 2.917 \end{aligned}$$

$$\begin{aligned} \mu_{X+Y} &= \mu_X + \mu_Y \\ \mu_{X+Y} &= 3.5 + (-3.5) = 0 \\ {\sigma^2}_{X+Y} &= {\sigma^2}_X + {\sigma^2}_Y \\ {\sigma^2}_{X+Y} &\approx 2.917 + 2.917 = 5.834 \end{aligned} \qquad \begin{aligned} \mu_{X+Y} &= \mu_X - \mu_Y \\ \mu_{X-Y} &= 3.5 - (-3.5) = 7 \\ {\sigma^2}_{X-Y} &= {\sigma^2}_X + {\sigma^2}_Y \\ {\sigma^2}_{X-Y} &\approx 2.917 + 2.917 = 5.834 \end{aligned}$$

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

Example:

I earn \$25.00 an hour for tutoring but spend \$20.00 an hour for piano lessons. I save the difference between my earnings for tutoring and the cost of the piano lessons. The number of hours I spend on each activity in one week varies independently according to the probability distributions shown below. Determine my expected weekly savings and the standard deviation of these savings.

Table 7.20:

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

Table 7.21:

Hours of Tutoring, x	Probability, p
1	0.2
2	0.3
3	0.2
4	0.3

Solution:

X will represent the number of hours per week taking piano lessons and Y will represent the number of hours tutoring per week.

$$\begin{aligned} E(X) = \mu_x &= \sum x_i p_i & Var(X) = \sigma^2{}_x &= \sum (x_i - \mu_x)^2 p_i \\ \mu_x &= 0(0.3) + 1(0.3) + 2(0.4) & \sigma^2{}_x &= (0-1.1)^2(0.3) + (1-1.1)^2(0.3) + (2-1.1)^2(0.4) \\ \mu_x &= 1.1 & \sigma^2{}_x &= 0.69 \\ & & \sigma_x &= 0.831 \end{aligned}$$

$$\begin{aligned} E(Y) = \mu_y &= \sum y_i p_i \\ \mu_y &= 1(0.2) + 2(0.3) + 3(0.2) + 4(0.3) \\ \mu_y &= 2.6 \end{aligned}$$

$$\begin{aligned} Var(Y) = \sigma^2{}_y &= \sum (y_i - \mu_y)^2 p_i \\ \sigma^2{}_y &= (1-2.6)^2(0.2) + (2-2.6)^2(0.3) + (3-2.6)^2(0.2) + (4-2.6)^2(0.3) \\ \sigma^2{}_y &= 1.24 \\ \sigma_y &= 1.11 \end{aligned}$$

The expected number of hours spent on piano lessons is 1.1 with a standard deviation of 0.831 hours. Likewise, the expected number of hours I spend tutoring is 2.6 with a standard deviation of 1.11 hours.

I spend \$20 for each hour of piano lessons so my mean weekly cost for piano lessons is

$\mu_{20x} = 20 \cdot \mu_x = (20)(1.1) = \$22.00$ Linear Transformation Rule

I earn \$25 for each hour of tutoring, so my mean weekly earnings from tutoring are

$\mu_{25x} = 25 \cdot \mu_y = (25)(2.6) = \$65.00$ Linear Transformation Rule

My expected weekly savings are

$\mu_{25y} - \mu_{20x} = \$65.00 - \$22.00 = \$43.00$ Subtraction Rule

The standard deviation of the cost of my piano lessons is

$\sigma_{20x} = (20)(0.831) = \$16.62$ Linear Transformation Rule

The standard deviation of my earnings from tutoring is

$\sigma_{25y} = (25)(1.11) = \$27.75$ Linear Transformation Rule

The variance of my weekly savings is

$$\begin{aligned}
{\sigma^2}_{25y-20x} &= {\sigma^2}_{25y} + {\sigma^2}_{20x} = (27.75)^2 + (16.62)^2 = 1046.2896 \\
\sigma_{25y-20x} &\approx \$32.35
\end{aligned}$$

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 two columns in which all of the outcomes are listed in one column and all of the associated probability in the other. The expected value and the variance of a probability distribution can be calculated using the formulas:

$$\begin{aligned}
E(X) &= \mu_x = \sum x_i p_i \\
Var(X) &= {\sigma^2}_x = \sum (x_i - \mu_x)^2 p_i
\end{aligned}$$

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

$$\mu_{c+dx} = c + d\mu_x$$
$$\sigma_{c+dx} = \left|d\right| \sigma_x$$

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

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

If X and Y are independent, then the variance is computed by:

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

Points to Consider

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

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 that take the bus to school. (Hint: Begin by listing all of the possible outcomes).
2. The Safe Grad Committee at the high school is selling 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 of \$1500 dollars and two of \$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. A survey has revealed that 76% of drivers now refrain from using the cell phone while driving. Three drivers were randomly selected and a probability distribution table was constructed to record the outcomes. Let N represent those drives who never use the cell phone while driving and S represent those who seldom use the cell phone. Calculate the expected value and the variance using technology.

Review Answers

1. Outcomes and their probabilities are:

$$\begin{aligned}
\text{P (none take the bus)} &= (0.3)(0.3)(0.3) = 0.027\\
\text{p (only the first student takes the bus)} &= (0.7)(0.3)(0.3) = 0.063\\
\text{P (only the second student takes the bus)} &= (0.3)(0.7)(0.3) = 0.063\\
\text{P (only the third student takes the bus)} &= (0.3)(0.3)(0.7) = 0.063\\
\text{P (the first and second students take the bus)} &= (0.7)(0.7)(0.3) = 0.147\\
\text{P (the second and third students take the bus)} &= (0.3)(0.7)(0.7) = 0.147\\
\text{P (the first and third students take the bus)} &= (0.7)(0.3)(0.7) = 0.147\\
\text{P (all three students take the bus)} &= (0.7)(0.7)(0.7) = 0.343
\end{aligned}$$

Table 7.22:

Number of Students Who Ride the Bus, x	Probability, p
0	0.027
1	$(3)(0.063) = 0.189$
2	$(3)(0.147) = 0.441$
3	0.343

2.

$$\mu_x = 1200 \cdot \frac{1}{500} + 0 \cdot \frac{499}{500} = \$2.40 \qquad \text{The expected value of the ticket for the gift basket is \$2.40.}$$

$$\mu_x = 1500 \cdot \frac{1}{500} + 500 \cdot \frac{2}{500} + \cdot \frac{497}{500} = \$5.00 \qquad \text{The expected value of the ticket for three prize draws is \$5.00}$$

Table 7.23:

Outcome	Probability
NNN	$(0.76)(0.76)(0.76) = .438976$
NNS	$(0.76)(0.76)(0.24) = .138624$
NSN	$(0.76)(0.24)(0.76) = .138624$
SNN	$(0.24)(0.76)(0.76) = .138624$
NSS	$(0.76)(0.24)(0.24) = .043776$
SNS	$(0.24)(0.76)(0.24) = .043776$
SSN	$(0.24)(0.24)(0.76) = .043776$

Table 7.23: (continued)

Outcome	Probability
SSS	$(0.24)(0.24)(0.24) = .013824$

Table 7.24:

Number of Drivers Using a Cell Phone, x	Probability, p
0	0.438976
1	0.415872
2	.131328
3	.013824

3.

Vocabulary

Expected Value The mean of the probability distribution for the random variable X. The symbol for expected value is $E(X)$ or μ_X.

Probability Distribution The set of values that a random variable takes on, together with a means of determining the probability of each value.

Random Variable A variable that takes on numerical values as the result of a chance.

7.6 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

In a previous lesson you learned about the Central Limit Theorem. One of the attributes of this theorem was that the sampling distribution of sample mean will follow a normal distribution as long as the sample size is large. As the value of n increases, the sampling distribution is more and more likely to follow a normal distribution. You've also learned that when the standard deviation of a population is known, a z-score can be calculated and used with the normal distribution to evaluate probabilities with the sample mean. In real-life situations, the standard deviation of the entire population (σ), is rarely known. Also the sample size is not always large enough to emulate a normal distribution. In fact there are often times when the sample sizes are quite small. What do you do when either one or both of these events occur?

t-Statistic

People often make decisions from data by comparing the results from a sample to some hypothesized or predetermined parameter. These decisions are referred to as tests of significance or hypothesis tests since they are used to determine whether the predetermined parameter is acceptable or should be rejected. We know that if we flip a fair coin, the probability of getting heads is 0.5. In other words, heads and tails are equally likely. Therefore, when a coin is spun, it should land heads 50% of the time. Let's say that a coin of questionable fairness was spun 40 times it landed heads 12 times. For these spins the sample proportion of heads is $\hat{p} = \frac{12}{40} = 0.3$. If technology is used to determine a 95% confidence interval to support the standard that heads should land 50% of the time, the reasonably likely sample proportions are in the interval 0.34505 to 0.65495. The class with $\hat{p} = 0.3$, is not captured within this confidence interval. Therefore, the fairness of this coin should be questioned; or, in other words, value of 0.5 as a plausible value for the proportion of times this particular coin lands heads when it is spun should be rejected. This data has provided evidence against the standard.

The object is to test the significance of the difference between the sample and the parameter. If the difference is small (as defined by some predetermined amount), then the parameter is acceptable. The statement that the proposed parameter is true is called the null hypothesis. If the difference is large and can't reasonably be attributed to chance, then the parameter can be rejected.

When the sample size is large, reliable estimates of the mean and variance of the population from which the sample was drawn can be made. Up to this point, we have used the z-score to determine the number of standard deviations a given value lays above or below the mean.

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

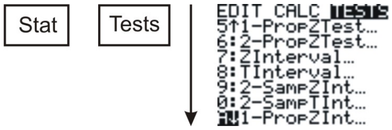
where $\bar{x}$ is the sample mean, μ_0 is the hypothesized mean stated in the null hypothesis $H_0 : \mu = \mu_0$ σ, is the population standard deviation and n is the sample size.

However the above formula cannot be used to determine how far a sample mean is from the hypothesized mean because the standard deviation of the population is not known. If the value of σ is unknown, s is substituted for σ and t for z. The t stands for the "test statistic," and it is given by the formula:

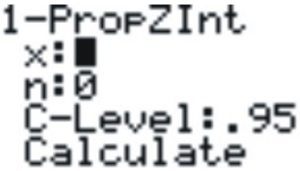
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean μ_0 is the population mean, s is the standard deviation of the sample and n is the sample size. The population mean μ is unknown but an estimate for this value is used. The t-test will be used to determine the difference between the sample mean and the hypothesized mean. The null hypothesis that is being tested is $H_0 : \mu = \mu_0$

So, suppose you want to see if a hypothesized mean passes a 95% level of confidence. The corresponding confidence interval can be determined by using the graphing calculator:



Press ENTER



```
1-PropZInt
 x:█
 n:0
 C-Level:.95
 Calculate
```

$x =$ the number of successes in the sample and

$n =$ the sample size

Press ENTER again. The confidence level will appear on the next screen. The value for t can now be compared with this interval to tell us whether or not the hypothesized mean can be accepted or rejected for this level of confidence.

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:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{3.21 - 3.16}{0.067/\sqrt{11}}$$

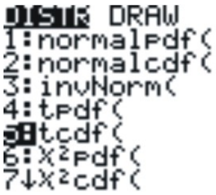
$$t \approx 2.48$$

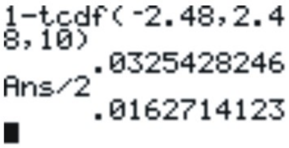
If the value of t from the sample fits right into the middle of the distribution of t constructed by assuming the null hypothesis is true, the null hypothesis is true. On the other hand, if the value of t from the sample is way out in the tail of the t -distribution, then there is evidence to reject 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 $df = 11 - 1$, or 10, that lies above $t = 2.48$ and below $t = -2.48$. This P-value can be calculated by using technology.

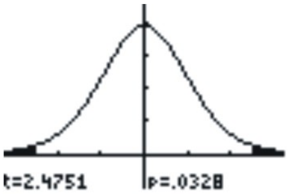
Press **2ND [DIST]** Use ↓ to select 5.tcdf (lower bound, upper bound, degrees of freedom)

This will be the total area under both tails. To calculate the area under one tail divide by 2.





```
1-tcdf(-2.48,2.4
8,10)
         .0325428246
Ans/2
         .0162714123
█
```

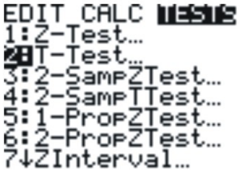


There is only a 0.016 chance of getting an absolute value of t as large as or even larger than the one from this sample ($2.48 \leq t \leq -2.48$). The small P-value tells us that the sample is inconsistent with the null hypothesis. The population mean differs from the estimated mean of 3.16.

When the P-value is close to zero, there is strong evidence against the null hypothesis. When the P-value is large, the result form the sample is consistent with the estimated or hypothesized mean and there is no evidence against the null hypothesis.

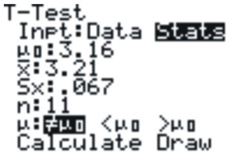
A visual picture of the P-value can be obtained by using the graphing calculator.

Stat Tests



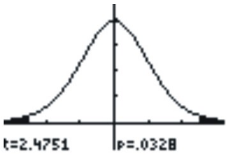
```
EDIT CALC TESTS
1:Z-Test…
2:T-Test…
3:2-SampZTest…
4:2-SampTTest…
5:1-PropZTest…
6:2-PropZTest…
7↓ZInterval…
```

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
```

This formula $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ is similar to that used in computing the z statistic with the unknown population standard deviation (σ) being substituted with the sample standard deviation.

There are numerous t-distributions and all are determined by a property of a set of data called the number of degrees of freedom. The degrees of freedom refer to the number of independent observations in a set of data. When estimating a mean score from a single sample, the number of independent observations is equal to the sample size minus one. In a single sample, there are n observations but only one parameter that needs to be estimated (the mean). This means that there are $n - 1$ degrees of freedom for estimating variability. In other words $df = n - 1$, where n is the sample size. The distribution of the t-statistic from samples of size 7 would be described by a t-distribution having $7 - 1$ or 6 degrees of freedom. Likewise, a t-distribution with 12 degrees of freedom would be used with a sample size of 13.

The t-score produced by this formula is associated with a unique cumulative probability which represents the chance of finding a sample mean less than or equal to $\bar{x}$, using a random sample of size n. The symbol t_α is used to represent the t-score that has a cumulative probability of $(1 - \alpha)$. If you needed the t-score to have a cumulative probability of 0.95, then α would be equal to $(1 - 0.95)$ or simply 0.05. This means that the t-score would be written as $t_{0.05}$. This value depends on the number of degrees of freedom and this value can be determined by using the table of values:

Table 7.25:

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413

Table 7.25: (continued)

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

From the table it can be determined that $t_{0.05}$ for 2 degrees of freedom is 2.92 while for 20 degrees of freedom the value is 1.72.

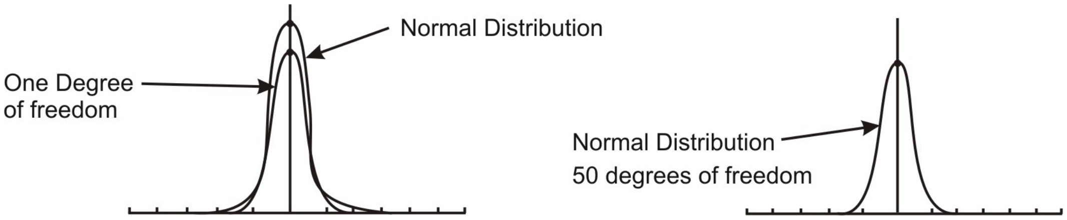
Since the t-distribution is symmetric about a mean of zero, the following statement is true.

$$t_\alpha = -t_{1-\alpha} \qquad \text{and} \qquad t_{1-\alpha} = -t_\alpha$$

Therefore, if $t_{0.05} = 2.92$ then by applying the above statement $t_{0.95} = -2.92$

A t-distribution is mound shaped, with mean 0 and a spread that depends on the degrees of freedom. The greater the degrees of freedom, the smaller the spread. As the number of degrees of freedom increases, the t-distribution approaches a normal distribution. 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.



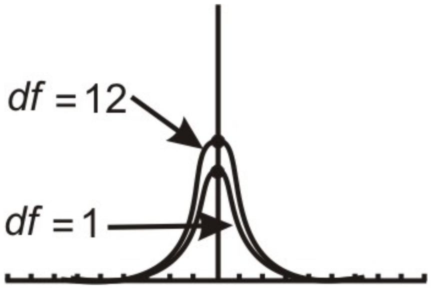
Notice in the first distribution graph the spread of the first (inner curve) is small but in the second one the both distributions are basically overlapping, so are roughly normal. This is due to the increase in the degrees of freedom.

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) ↓ 4. **tpdf(**

You are now on the $Y =$ screen.

$Y = \text{tpdf}(X, 1)$ [**Graph**]



Repeat the steps to plot more than one t-distribution on the same screen.

Notice the difference in the two distributions.

The one with $12 = df$ 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:

1. The population distribution is normal.
2. The sampling distribution is symmetric and the sample size is ≤ 15.
3. The sampling distribution is moderately skewed and the sample size is $16 \leq n \leq 30$.
4. The sample size is greater than 30, without outliers.

The t-distribution also has some unique properties. These properties are:

1. The mean of the distribution equals zero.

2. The population standard deviation is unknown.

3. 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.

$$\text{Variance} = \frac{\text{df}}{\text{df} - 2} \text{ and df} > 2$$

4. 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.

5. 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.

6. 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?

Solution:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \text{The degrees of freedom are } (n-1) = 15 - 1. \text{ This means 14 degrees of freedom.}$$

$$t = \frac{290 - 300}{50/\sqrt{15}}$$

$$t = \frac{-10}{12.9099}$$

$$t = -.7745993$$

Using the graphing calculator or a table of values, the cumulative probability is 0.286, which means that if the true life span of a battery were 300 hours, there is a 28.6% 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} $\boxed{\text{Vars}}$ (Dist) ↓ 4. **tpdf(**

You are now on the $Y =$ screen.

$$Y = \text{tpdf}(-.7745993, 14) = [0.286]$$

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?

Solution:

Begin the problem by determining the mean of the sample and the sample standard deviation.

This can be done using the graphing calculator. $\bar{x} \approx 70.143$ and $s \approx 1.676$.

$$\begin{aligned}
t &= \frac{\bar{x} - \mu}{s/\sqrt{n}} \\
t &= \frac{70.143 - 72}{1.676/\sqrt{7}} \\
t &\approx -2.9315
\end{aligned}$$

Example:

In the example before last the test statistic for testing that the mean weight of the cheese wasn't 72 was computed. Find and interpret the P-value.

Solution:

The test statistic computed in the example before last was -2.9315. Using technology, the $P-$ value is 0.0262. If the mean weight of cheese is 72 ounces, the probability that the volume of 7 random measurements would give a value of t greater than 2.9315 or less than -2.9315 is about 0.0262.

Example:

In the previous example, the P-value for testing that the mean weight of cheese wasn't 72 ounces was determined.

a) State the hypotheses.

b) Would the null hypothesis be rejected at the 10% level? The 5% level? The 1% level?

Solution:

a) H_0 : The mean weight of cheese, μ is 72 ounces.

$$H_\alpha : \mu \neq 72$$

b) Because the P-value of 0.0262 is less than both 10% and 5%, the null hypothesis would be rejected at these levels. However, the P-value is greater than 1% 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 mean. 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?

Review Questions

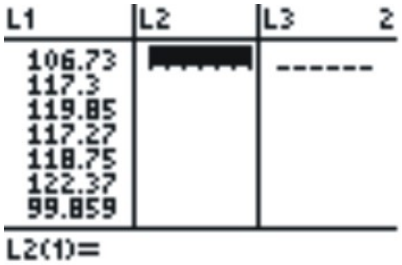
1. You intend to use simulation to construct an approximate t-distribution with 8 degrees of freedom by taking random samples from a population with bowling scores that are normally distributed with mean, $\mu 110$ and standard deviation, $\sigma = 20$.
   (a) Explain how you will do one run of this simulation.
   (b) Produce four values of t using this simulation.

Review Answers

1. (a) 8 degrees of freedom mean that the sample size is $(8 + 1)$ or 9. The graphing calculator will be used in order to randomly select scores from a normally distributed population.

   $\boxed{\text{Math}} \longrightarrow \boxed{\text{PRB}} \downarrow$ 6.randNorm(110,20,9) $\boxed{\text{STO}}$ 2^{nd} 1

This command will select 9 scores from the population and store the values in List One.

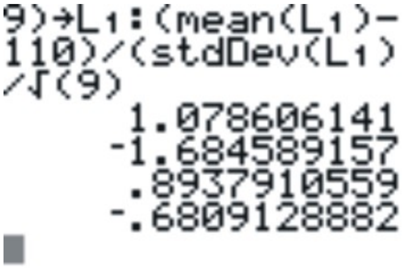


```
L1     |L2     |L3      2
106.73 |██████ |------
117.3  |       |
119.85 |       |
117.27 |       |
118.75 |       |
122.37 |       |
99.859 |       |
L2(1)=
```

These are values selected by the calculator.

(b) To calculate $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$, enter on the calculator

randNorm$(110, 20, 9)L_1$: (mean$(L_1) - 100)/($stdDev$(L_1)/\sqrt{9}$ **Press enter** 4 **times to generate the** t **values.**



```
9)→L1:(mean(L1)-
110)/(stdDev(L1)
/√(9)
     1.078606141
    -1.684589157
     .8937910559
    -.6809128882
█
```

The mean and stdDev functions are easily found in the catalog of the calculator ($2^{nd}0$)

Vocabulary

Alternative Hypothesis The set of values that an investigator believes may contain the plausible values of a population parameter in a significance test.

Null Hypothesis The standard value of a parameter that is assumed to be true in a significance test until possibly refuted by the data in favor of an alternative hypothesis.

P-**Value** The probability of seeing a result from a random sample that is as extreme as or more extreme than the result obtained from the random sample if the null hypothesis is true.

t-**Distribution** The distribution of $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ when the data are a random sample from a normally distributed population.

Image Sources

Chapter 8

Hypothesis Testing

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 single- 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

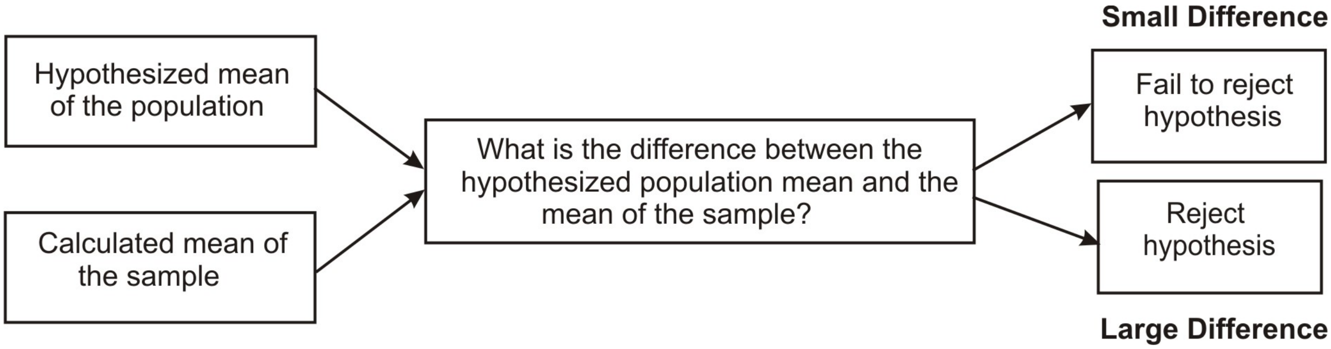
In this chapter we will explore **hypothesis testing,** which involves making educated guesses about a population based on a sample drawn from the population. Most times, hypothesis testing involves making guesses about the difference between the hypothesized value of the mean of an overall population and that of the sample. This is 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 an educated guess 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 the different types of hypothesis testing, 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

As mentioned in the introduction, hypothesis testing involves testing the difference between a hypothesized value of the mean of an overall population and the mean calculated from a sample. 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.


Small Difference
Hypothesized mean of the population
Fail to reject hypothesis
What is the difference between the hypothesized population mean and the mean of the sample?
Calculated mean of the sample
Reject hypothesis
Large Difference

In statistics, the hypothesis to be tested is called the **null hypothesis** and given the symbol H_0. The null hypothesis states that there is no relationship or no difference between an accepted population mean and a sample mean. So finding a significant result means refuting the null hypothesis, showing that the true population mean is likely to be closer to the sample mean. We would calculate the mean of the sample and generalize these findings to the overall population. For example, if we were to test the hypothesis that the seniors had a mean SAT score of $1,100$, our null hypothesis would be that the SAT score would be equal to $1,100$ or:

$$H_0 : \mu = 1100$$

where:

$H_0 =$ symbol for null hypothesis

$\mu =$ population mean

$1,100 =$ value to be tested

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 a couple of 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:

$$\begin{aligned} H_0 &: \mu = 14 \\ H_a &: \mu \neq 14 \end{aligned}$$

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.

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?

Solution:

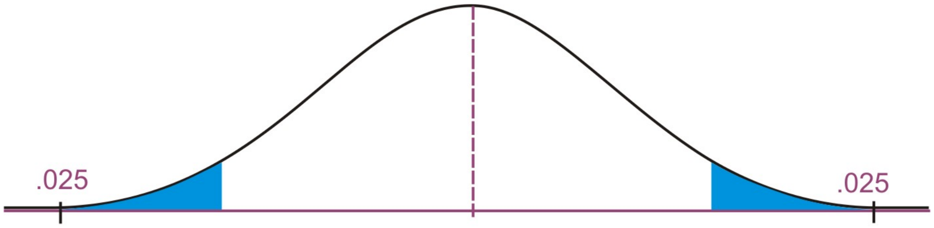
$$\begin{aligned} H_0 &: \mu = 3.2 \\ H_a &: \mu \neq 3.2 \end{aligned}$$

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.

Deciding Whether to Reject the Null Hypothesis: Single 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. Statisticians first choose a **level of significance** or **alpha** (α) **level**, which is an event probability below which discrepancies from the null hypothesis are deemed significant. The most frequently used levels of significance are 0.05 and 0.01. In other words, these levels mean that when we make the decision to reject the null hypothesis, we are correct 95 or 99 percent of the time. The areas outside of these levels of significance are called the **critical regions**. 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

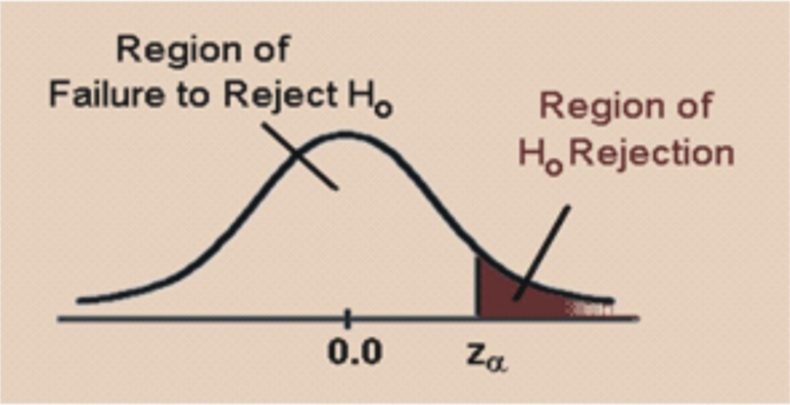
Therefore, if the mean from sample taken from the population falls within these critical regions, we would conclude that there was too much of a difference and we would reject the null hypothesis. However, if the mean from the sample falls in the middle of the distribution (in between the critical regions) 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:

$$H_0 : \mu = 1100$$
$$H_a : \mu \neq 1100$$

In this scenario, our null hypothesis states that the mean SAT scores would be equal to $1,100$ while the alternate hypothesis states that the SAT scores would be greater than $1,100$. 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).


Region of Failure to Reject H_o
Region of H_o Rejection
0.0
z_α

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:

Use the z-distribution table to determine the critical value for a single-tailed hypothesis test with a 0.05 significance level.

Solution:

Using the z-distribution table, we find that a significance level of 0.05 corresponds with a critical value of 1.645.

Calculating the Test Statistic

Before evaluating our hypotheses by determining the critical region and calculating the test statistic, we need to first:

1. Confirm that the distribution is normal.
2. Determine the hypothesized mean (μ) of the distribution.
3. If we don't have the population variance, we will need to calculate the standard deviation of the sample so that we can calculate the **standard error of the mean** (σ_X).

Remember that since we have a random sample from the population, we do not expect the sample mean to be *exactly* equal to the hypothesized value of the population mean. Therefore, the question really is: "How different can the observed sample mean be from the hypothesized mean before rejecting the null hypothesis?" Or, in other words, "If the null hypothesis is true, is it likely that we will obtain such an observed sample mean?" We use our critical values taken from the z-distribution to determine those cutoffs.

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. As we learned in previous lessons, the z-score is calculated by using the formula:

$$z = \frac{(\bar{X} - \mu)}{\sigma_X}$$

where:

$z =$ standardized score

$\bar{X} =$ sample mean

$\mu =$ hypothesized population mean

σ_X = standard error?. If we do not have the population variance, we can estimate the deviation of the samples from the true population mean by dividing the standard deviation by the square root of the number of observations $\left(\frac{\sigma}{\sqrt{n}}\right)$.

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.

Let's calculate the test statistic for several different scenarios.

Example:

College A has an average SAT score of 1,500. From a random sample of 125 freshman psychology students we find the average SAT score to be 1,450 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?

Solution:

Let's first develop our null and alternative hypotheses:

$$H_0 : \mu = 1500$$
$$H_a : \mu \neq 1500$$

Our standard z-score for the sample of freshman psychology students would be:

$$z = \frac{\bar{X} - \mu}{\sigma_x} = \frac{1450 - 1500}{100/\sqrt{125}} \approx -5.59$$

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 his plants and finds the average number of pods to be 147. He wonders whether or not this is a statistically significant increase. What is his hypotheses and the test statistic?

Solution:

First, we develop our null and alternative hypotheses:

$$H_0 : \mu = 145$$

Let's calculate the test statistic for several different scenarios.

Example:

$$H_a : \mu > 145$$

This alternative hypothesis is > since we are only concerned with the pod *gain* which translates to above the mean.

Next, we calculate the standard z-score for the sample of pea plants.

$$z = \frac{\bar{X} - \mu}{\sigma_X} = \frac{147 - 145}{100/\sqrt{144}} = 0.24$$

In the following lessons, we will use these standard z-scores and the critical regions to evaluate the null and the alternative hypotheses.

Testing the P-Value of an Event

We can also evaluate a hypothesis by testing the probability, or the P-value, of an event occurring. When we assume that we have normal distributions, we can determine approximately where on the normal distribution that the sample mean will fall. When we know where it falls, we can determine the **probability** of obtaining a sample value either greater or smaller than the mean by using the z-score.

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). But the farmer is really hoping that some plants have a more dramatic yield increase. What is the probability of a plant having a much higher yield of over 155 pea pods?

To find this probability, first find the z-score for the hypothesized sample mean using the formula that we learned in the section above. Therefore, a z-score for a sample of plants with 155 pods would be:

$$z = \frac{\bar{X} - \mu}{\sigma_X} = \frac{155 - 145}{100/\sqrt{144}} = 1.20$$

Using the z-score distribution, we find that the area beyond a z-score of 1.20 is equal to .1151. This means that there is .1151 or 11.5% chance that a pea plant will produce over 155 pods.

Type I and Type II Errors

When we decide to reject or not reject the null hypothesis, we have four possible scenarios:

1. A true hypothesis is rejected.
2. A true hypothesis is not rejected.
3. A false hypothesis is not rejected.

4. A false hypothesis is rejected.

If a hypothesis is true and we do not reject it (Option 2) or if a false hypothesis is rejected (Option 4), we have made the correct decision. But if we reject a true hypothesis (Option 1) or a false hypothesis is not rejected (Option 3) we have made an error. Overall, one type of error is not necessarily more serious than the other. Which type is more serious depends on the specific research situation, but ideally both types of errors should be minimized during the analysis.

Table 8.1: **The Four Possible Outcomes in Hypothesis Testing**

Decision Made	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct Decision
Do not Reject Null Hypothesis	Correct Decision	Type II Error

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.

Calculating the probability of making a **Type II** error (?) is not as straightforward as calculating a Type I error. The probability of making a Type II error can only be determined when values have been specified for both the alternative hypothesis and the null hypothesis. Once the value for the alternative hypothesis has been specified, it is possible to determine the probability of making a correct decision (1- ?). This quantity, 1- ?, is called the **power of the test** and is discussed in the next section.

As mentioned, our goal 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 (?) decreases and vice versa. While ? is under our direct control, ? is not. We will look at this relationship a bit more in depth in the next section.

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 (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 which is specified much in the same way as we specify the value in the null hypothesis. For example, suppose that a doctor is concerned about making a Type II error only if the active ingredient in the new medication is less 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.

$$
\begin{aligned}
H_0 &: \mu = 250 \\
H_a &: \mu = 253
\end{aligned}
$$

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. In other words, we want to determine the power of our test for detecting this difference. In this example, we may choose a certain dosage if there were medical repercussions above that level.

We want to find the area under the curve that is associated with making a Type II error. In the example above, this means that we need to find the power that the test has for detecting this difference. 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. Say that we have an alpha level of .05 – we would then have a critical value of 1.64 for the single-tailed test which would have a value of:

$$
\begin{aligned}
z &= \frac{\bar{X} - \mu}{\sigma_X} \\
1.64 &= \frac{\bar{X} - 250}{50/\sqrt{200}} \\
\bar{X} &= 1.64\left(\frac{50}{\sqrt{200}}\right) + 250 \approx 255.8
\end{aligned}
$$

Now, with a new mean set at the alternative hypothesis ($H_a : \mu = 253$) we want to find the value of the critical score (255.8) when centered around this score. Therefore, we can figure that:

$$z = \frac{\bar{X} - \mu}{\sigma_X} = \frac{255.8 - 253}{3.54} \approx 0.79$$

Using the standard z distribution we find that the area to the right of a z-score of .79 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. The power of this test is about 0.215. In other words, this test of the null hypothesis is not very powerful and has only a 0.215 probability of detecting the real difference between the 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.

Lesson Summary

1. Hypothesis testing involves making educated guesses about a population based on a sample drawn from the population. We generate null and alternative hypotheses based on the mean of the population to test these guesses.
2. We establish critical regions based on level of significance or alpha (α) levels. If the value of the test statistic falls in these critical regions, we are able to reject it.
3. 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.
4. 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.
5. 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 making a Type II error.

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 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 $1,020$. From a random sample of 144 Sacramento High School students we find the average SAT score to be $1,100$ 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.
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. Please 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 \$25,100. ($H_a : \mu = \$25,100$). For an $\alpha = .05$ and a sample of 144, determine the power of a one-tailed test.

Review Answers

1. Reject, Fail to Reject
2. $H_0 : \mu = 19,\ H_a : \mu \neq 19$
3. Fail to Reject, Reject
4. $Z = 2.325$
5. $H_0 : \mu = 1020, H_a : \mu \neq 1020, Z = 6.67$
6. Probability

7. Area beyond a z-score of $0.20 = .4207$. Therefore, there is a probability of 42.07% that a teenager will have savings greater than \$520.
8. Type I error, Type II error
9. Power of the Test
10. 0.44

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.
- Test a hypothesis about a population proportion using confidence intervals.

Introduction

For most hypotheses that we will study, we use the general formula for calculating the test statistic:

$$\text{Test Statistic} = \frac{\text{Observed Sample Mean-Population Mean}}{\text{Standard Error}}$$

or, the familiar

$$z = \frac{\bar{X} - \mu_0}{\sigma_x}$$

This formula helps us determine the magnitude of the difference between the observed sample mean and the hypothesized population mean. However, many times in statistics we study or make inferences about **proportions** of a population. For example, when we look at election results we often look at the proportion of people that vote and who this proportion of voters will 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 Barak 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

As mentioned, we perform hypothesis tests about population proportions (often called percentages) quite often. We could perform these tests in the following examples:

- 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 percent of graduating seniors will attend a $4-$year college.
- H_0: 42 percent of voters will vote for John McCain.
- H_0: 26 percent of people will choose Diet Pepsi over Diet Coke.

While we can use similar methods to test these hypotheses, we do need to take several different factors into account. Because it is impractical to measure every member of the population, we follow a series of steps:

1. Hypothesize a value for the population proportion (P) like we did above.
2. Randomly select a sample.
3. Use the sample proportion (p) to test the stated hypothesis.

Essentially, the sampling distribution of this sample proportion is used the same way that we use the sample mean distribution. So how do we account for the different sampling distribution of p? 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). However, we should remember that when the sample size is relatively large, we can use the normal distribution to approximate the binomial distribution.

In order to calculate the standard deviation of the sample distribution, we need to calculate something called the **standard error of the proportion** which is defined as:

$$s_p = \sqrt{\frac{PQ}{n}}$$

where:

$P =$ the hypothesized value of the proportion

$Q =$ proportion *not* possessing the characteristic

$n =$ sample size

Let's take a look at an example on how we would calculate the standard error of the proportion.

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 standard error of the proportion?

Solution:

Since we want to test the proportion of graduating seniors and we think that proportion is around 60 percent, our hypotheses are:

$$\begin{aligned} H_0 &: P = 0.60 \\ H_a &: P \neq 0.60 \end{aligned}$$

And the standard error would be:

$$s_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.60 \times 0.40}{400}} = 0.0245$$

Therefore, the sampling distribution of p for this example has a mean equal to 0.60 (the hypothesized value of P) and a standard deviation of 0.0245. With this information, we can easily evaluate hypotheses using a standard formula.

Testing a Proportion Hypothesis Using the P-Value

Similar to testing hypotheses dealing with population means, we use a similar set of steps when testing proportion hypotheses.

1. Determine and state the null and alternative hypotheses.
2. Set the criterion for rejecting the null hypothesis.
3. Calculate the test statistic.
4. Interpret the results and decide whether to reject or fail to reject the null hypothesis.

To test a proportion hypothesis, we use the formula for calculating the test statistic for a mean, but modify it accordingly. Therefore, our formula for the test statistic of a proportion hypothesis is:

$$z = \frac{p - P}{s_p}$$

where:

$p =$ the sample proportion

$P =$ the hypothesized population proportion

$s_p =$ the standard error of the proportion

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 vote for the bill or not?

Solution:

First, we develop our null and alternative hypotheses.

$$H_0 : P = 0.70$$
$$H_a : P > 0.70$$

Next, we should set the criterion for rejecting the null hypothesis. We will use a probability (?) level of 0.05 and since we are interested only in the probability that the percentage of constituents is *greater* than 0.70, we will use a single-tailed test. Looking at the standard z-table, we find that the **critical value** for a single-tailed test at an alpha level of 0.05 is equal to 1.64.

To calculate the test statistic, we first find the standard error of the proportion.

$$S_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.70 \times 0.30}{300}} \approx 0.0265$$

After finding the standard error, we can calculate the standard z-score needed to evaluate our hypothesis.

$$z = \frac{p - P}{s_p} = \frac{0.74 - 0.70}{0.0265} \approx 1.51$$

Since our critical value is 1.64 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 0.70 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?

Solution:

First, we develop our null and alternative hypotheses.

$$\begin{aligned} H_0 &: P = 0.30 \\ H_a &: P \neq 0.30 \end{aligned}$$

Next, we should set the criterion for rejecting the null hypothesis. The 0.05 alpha level is used and for an ? = 0.05 the critical values of the test statistic are 1.96 standard deviations above or below the mean.

To calculate the test statistic, we first find the standard error of the proportion.

$$S_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.30 \times 0.70}{400}} \approx 0.0229$$

After finding the standard error, we can calculate the standard z-score needed to evaluate our hypothesis.

$$Z = \frac{p - P}{s_p} = \frac{0.25 - 0.30}{0.0229} \approx -2.18$$

Since our critical value is 1.96 and our test statistic is -2.18, 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.

Confidence Intervals for Hypotheses about Population Proportions

When making a decision, we like to be able to determine how confident we are about a decision. For example, when a congressman is deciding whether or not to vote for a bill, he would like to be able to say something to the effect of "I am 99% confident that 70 percent of my constituents will support this decision." With statistical analysis, we can construct something called the **confidence interval** that specifies the level of confidence that we have in our results.

The confidence interval is a range of values that we are confident, but not certain, contains the population parameter that we are studying (most often this parameter is the mean).

We interpret the results of the confidence intervals by calculating:

- The level of confidence (i.e. $-95\%, 99\%$, etc.)
- The interval (i.e. -40.4 to 45.6 or 102 to 108, etc.)

If we are estimating the confidence interval for a population mean, then we use the sample mean for the statistic. However, if we are estimating for a population proportion, we use the sample population proportion.

The confidence interval always includes the population parameter. Therefore, when we construct a confidence interval we can conclude that that interval also contains the sample statistic. A confidence interval statement would look something like:

- We are 95 percent confident that the interval from 34.2 to 39.1 contains the mean

- $(2.10, < \mu < 2.90)$ – We are 90 percent confident that this interval contains the population proportion

We can *not* say that the probability is 95 percent that the interval contains the mean since either the interval contains the mean or it does not. Therefore, when we talk of our confidence level we say that we '$X\%$ certain" that the specific interval contains the mean.

Example:

In our example about the congressman voting for the bill on gay marriage, the congressman decides that he wants an estimate of the proportion of voters in the population that are likely to vote for a bill. Construct a confidence interval for this population proportion.

Solution:

As a reminder, our sample proportion was 0.746 and our standard error of the proportion was 0.0265. To correspond with the $? = .05$, 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 $z = +1.96$. The confidence interval for this proportion would be:

$$CI_{95} :$$
$$p \pm 1.96(\text{standard error})$$
$$0.746 \pm (1.96)(0.0265)$$

So $0.694 < p < 0.798$

With respect to the population proportion, we are 95% confident that the interval from 0.69 to .077 contains the population proportion. This means that we are 95% confident that the average proportion of voters who will support the bill is between 69 and 77%.

Lesson Summary

1. 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.

2. 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 $s_p = \sqrt{\frac{PQ}{n}}$

3. The formula for calculating the test statistic for a population proportion is

$$z = \frac{p - P}{s_p}$$

where:

$p =$ the sample proportion

$P =$ the hypothesized population proportion

$s_p =$ the standard error of the proportion

4. We can construct something called the confidence interval that specifies the level of confidence that we have in our results. The confidence interval is a range of values that we are confident, but not certain, contains the population parameter that we are studying.

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. True or false: A confidence interval states the probability that the interval contains the mean. For example, a confidence interval of 95% would say that "This interval contains the mean 95% of the time."

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.

4. What are the null and alternative hypotheses for this problem?
5. What is the observed value of the sample proportion?
6. What is the standard error of the proportion?
7. What is the test statistic for this scenario?
8. What decision would you make about the null hypothesis if you had an alpha level of .01?
9. The state senator decided that she is still wants an estimate of the proportion of voters in the population who are likely to vote for the bill. Construct a 99% confidence interval around this proportion.
10. Please write a statement describing the results of the confidence interval.

Review Answers

1. The magnitude of the difference between the observed sample mean and the hypothesized population mean.
2. True
3. False

We *can not* say that the probability is 95 percent that the interval contains the mean since either the interval contains the mean or it does not. Therefore, when we talk of our confidence level we say that we are 'X% certain" that the specific interval contains the mean.

4. $H_0 : P = 0.60, H_a : P > 0.60$
5. $p = 495/750 = 0.66$
6. 0.0179
7. $z = 3.35$
8. Since the test statistic of 3.35 is exceeds the critical value of 2.33 (one-tailed z-test at .01), we reject the null hypothesis and conclude that the probability is less than 0.01 that a sample proportion of 0.66 would appear due to sampling error if in fact the population proportion was equal to 0.60.

9. $CI = (0.614, 0.706)$
10. We are 99% confident that the interval $(0.614, < p < 0.706)$ contains the proportion mean. In other words, this confidence interval shows perhaps as many as 70 percent of the voters favor the bill, but it is very unlikely that less than 61 percent favor the bill.

8.3 Testing a Mean Hypothesis

Learning Objectives

- Calculate the sample test statistic to evaluate a hypothesis about a population mean based on large samples.
- Differentiate the difference in hypothesis testing for situations with small populations and use the Student's t-distribution accordingly.
- Understand the results of the hypothesis test and how the terms 'statistically significant' and 'not statistically significant' apply to the results.

Introduction

In the previous sections, we have covered:

- the reasoning behind hypothesis testing.
- how to conduct single and two-tailed hypothesis tests.
- the potential errors associated with hypothesis testing.
- how to test hypotheses associated with population proportions.

In this section we will take a closer look at some examples that will give us a bit of practice in conducting these tests and what these results really mean. In addition, we will also look at how the terms **statistically significant** and **not statistically significant** apply to these results.

Also, it is important to look at what happens when we have a small sample size. All of the hypotheses that we have examined thus far have assumed that we have normal distributions. But what happens when we have a small sample size and are unsure if our distribution is normal or not? We use something called the Student's t-distribution to take small sample size into account.

Evaluating Hypotheses for Population Means using Large Samples

When testing a hypothesis for a normal distribution, we follow a series of four basic steps:

1. State the null and alternative hypotheses.
2. Set the criterion (critical values) for rejecting the null hypothesis.
3. Compute the test statistic.
4. Decide about the null hypothesis and interpret our results.

In Step 4, we can make one of two decisions regarding the null hypothesis.

- If the test statistic falls in the regions above or below the critical values (meaning that it is far from the mean), we can reject the null hypothesis.
- If the test statistics falls in the area between the critical values (meaning that it is close to the mean) we fail to reject the null hypothesis.

When 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. If we reject the null hypothesis, we are also saying that the probability that the observed sample mean will have occurred by chance is less than the ? level of .05, .01 or whatever we decide.

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. This decision is based on the properties of sampling and the fact that there is not a large difference is reason to not reject the null hypothesis. Essentially, we are willing to attribute this difference to sampling error.

Let's perform a hypothesis test for the scenarios we examined in the first lesson.

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. Is the sample of freshman psychology students representative of the overall population?

Solution:

Let's first develop our null and alternative hypotheses:

$$H_0 : \mu = 1500$$
$$H_a : \mu \neq 1500$$

At a 0.05 significance level, our critical values would be 1.96 standard deviations above and below the mean.

Next, we calculate the standard $z-$score for the sample of freshman psychology students.

$$z = \frac{X - \mu}{\sigma_x} = \frac{1500 - 1450}{100\sqrt{125}} \approx 5.59$$

Since the calculated z-score of 5.59 falls in the critical region (as defined by a 0.05 significance level or anything with a z-score of above 1.96) we reject the null hypothesis. Therefore, we can conclude that the probability of obtaining a sample mean equal to 1450 if the mean of the population is 1500 is very small and the sample of freshman psychology students is not representative of the overall population. Furthermore, the probability of this difference occurring by chance is less than 0.05.

Example:

The school nurse was wondering if the average height of 7th graders has been increasing. Over the last 5 years, the average height of a 7th grader was 145 cm with a standard deviation of 20 cm. The school nurse takes a random sample of 200 students and finds that the average height this year is 147 cm. Conduct a single-tailed hypothesis test using a 0.05 significance level to evaluate the null and alternative hypotheses.

Solution:

First, we develop our null and alternative hypotheses:

$$H_0 : \mu = 145$$
$$H_a : \mu \neq 145$$

At a 0.05 single-tailed significance level, our critical value for a single-tailed test would be 1.64 standard deviations above the mean.

Next, we calculate the standard z-score for the sample of 7th graders.

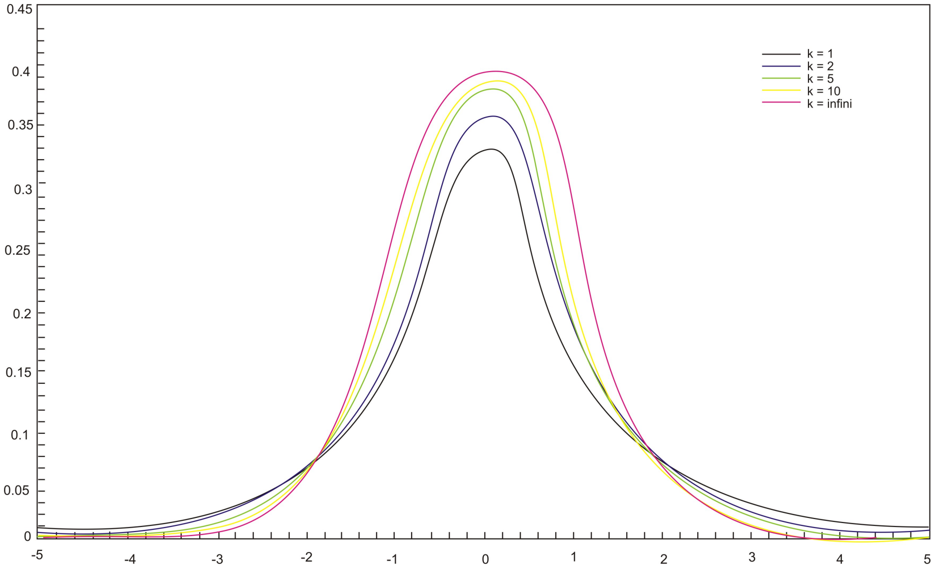
$$z = \frac{X - \mu}{\sigma_X} = \frac{147 - 145}{20\sqrt{200}} \approx 1.41$$

Since the calculated z-score of 1.41 does not fall in the critical region (as defined by a 0.05 significance level or anything with a z-score of above 1.67) we fail to reject the null hypothesis. We can 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.

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 excerpt 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. Below is an excerpt from the Student's t-table for one-sided critical values.

Table 8.3:

DF	Probability of Exceeding the Critical Value					
	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313

Table 8.3: (continued)

DF	Probability of Exceeding the Critical Value					
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143

As the number of observations gets larger, the t-distribution approaches the shape of the normal distribution. In general, once the sample size is large enough - usually about 120 - we would use the normal distribution or the z-table instead.

In calculating the t-test statistic, we use the formula:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

where:

$t =$ test statistic

$\bar{X} =$ sample mean

$\mu =$ hypothesized population mean

$s_{\bar{x}} =$ estimated standard error

To estimate the standard error ($s_{\bar{x}}$, we use the formula $s/\sqrt{n}$ where s is the standard deviation of the sample and n is the sample size.

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 and that the standard deviation of the sample is 0.54. After an initiative to help improve the GPA of student athletes, the athletic director samples 20 football players and finds that their GPA is 3.18. Is there a significant improvement? Use a .05 significance level.

Solution:

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 (α) 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 standard errors above and below the mean.

In calculating the test statistic, we use the formula:

$t = \frac{\bar{X}-\mu}{s_{\bar{x}}} = \frac{3.18-3.10}{0.54/\sqrt{20}} \approx 0.66$

This means that the observed sample mean (3.18) of football players is 0.66 standard errors above the hypothesized value of 3.10. Because $t = 0.66$ does not exceed 2.093 (the standard critical value), the null hypothesis is not rejected.

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.

How to Interpret the Results of a Hypothesis Test

In the previous section, we discussed how to interpret the results of a hypothesis test. As a reminder, when 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.

But what is meant by **statistical significance**? Technically, the difference between the hypothesized population mean and the sample mean is said to be *statistically significant* when the probability that the difference occurred by chance is less than the significance (α) level. Therefore, when the calculated test statistic (whether it is the z- or the t-score) falls in the area beyond the critical score, we say that the difference between the sample mean and the hypothesized population mean is **statistically significant.** When the calculated test statistic falls in the area between the critical scores we say that the difference between the sample mean and the hypothesized population mean is **not statistically significant.**

Lesson Summary

1. When testing a hypothesis for the mean of a distribution, we follow a series of four basic steps:

- State the null and alternative hypotheses.
- Set the criterion (critical values) for rejecting the null hypothesis.
- Compute the test statistic.
- Decide about the null hypothesis and interpret our results.

2. When 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.

3. 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.

4. We use the t-distribution in hypothesis testing the same way that we use the normal distribution. However, the t-distribution is used when the sample size is small (typically less than 120) and the population standard deviation is unknown.

5. When calculating the t-statistic, we use the formula:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

where:

$t =$ test statistic

$\bar{X} =$ sample mean

$\mu =$ hypothesized population mean

$s_{\bar{x}} =$ estimated standard error, which is computed by $\frac{s}{\sqrt{n}}$

6. The difference between the hypothesized population mean and the sample mean is said to be statistically significant when the probability that the difference occurred by chance is less than the significance (α) level.

Review Questions

1. In hypothesis testing, when we work with large samples (typically samples over 120), we use the _____ distribution. When working with small samples (typically samples under 120), we use the _____ distribution.
2. 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.

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.

3. What would the null and alternative hypotheses be for this scenario?
4. What would the standard error be for this particular scenario?
5. Describe in your own words how you would set the critical regions and what they would be at an alpha level of .05.
6. Test the null hypothesis and explain your decision
7. 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?
8. Using the appropriate t-distribution, test the same null hypothesis with a sample of 30.
9. With a sample size of 30, do you need to have a **larger** or **smaller** difference between then hypothesized population mean and the sample mean to obtain statistical significance? Explain your answer.
10. 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 $n = 444$

Review Answers

1. z, t
2. True
3. $H_0 : \mu = 2.75, H_a : \mu \neq 2.75$
4. 0.406
5. When setting the critical regions for this hypothesis, it is important to consider the repercussions of the decision. Since there does not appear to be major financial or health repercussions of this decision, a more conservative alpha level need not be chosen.

With an alpha level of .05 and a sample size of 256, we find the area under the curve associated in the z-distribution and set the critical regions accordingly. With this alpha level and sample size, the critical regions are set at 1.96 standard scores above and below the mean.

6. With a calculated test statistic of 2.463, we reject the null hypothesis since it falls beyond the critical values established with an alpha level of .05. This means that the probability that the observed sample mean would have occurred by chance if the null hypothesis is true is less than 5%.
7. Yes, because the sample size is below 120, in most cases the t -distribution would be the appropriate distribution to use and what you have is s not t.
8. The critical values for this scenario using the t-distribution are 2.045 standard scores above and below the mean. With a calculated t-test statistic of 0.8425, we do not reject the null hypothesis. Therefore, we can conclude that the probability that the observed sample mean could have occurred by chance if the null hypothesis was true is greater than 5%.
9. You would need a larger difference because the standard error of the mean would be greater with a sample size of 30 than with a sample size of 256.
10. (a) one-tailed test
    (b) .05 level of significance
    (c) $n = 144$

8.4 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 proportions, large samples 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 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 between two fixed populations by testing 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.), selecting a random sample and then assigning the half of the subjects to one group and half to another.

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:

- A group may be measured twice such as in a pretest-posttest situation (scores on a test before and after the lesson).
- In a matched sample where each observation is matched with an observation in the other 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:

$$H_0 : \delta = \mu_1 - \mu_2 = 0$$

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 $(s_{\bar{X}_1} - s_{\bar{X}_2})$. Usually, with one sample this calculation is pretty easy since it is based on either standard deviation of the sample or the population variance. However, when calculating this statistic for two samples, it is a bit more difficult. To calculate this statistic we use the formula:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

Where n_1 *and* n_2 the sizes of the two samples

$s^2 =$ the pooled sample variance, which is computed as shown below

The pooled estimate of variance is found by adding the sums of the squared deviations (s) around the sample means and then dividing the total by the sum of the degrees of freedom in the two samples.

Therefore, we can find this estimate by using the formula:

$$s^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

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:

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

Let's calculate this estimate using a sample set of data.

Example:

Say that we have two independent samples of student reading scores. The data are as follows:

Table 8.4:

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 8.5:

Descriptive Statistic	Sample 1	Sample 2
Number (n)	5	6
Sum of Observations (X)	35	78
Mean of Observations $(\bar{X})$	7	13
Sum of Squared Deviations $(\sum_{i=1}^{n}(X_i - \bar{X})^2)$	20.0	40.0

Using the formula for the pooled estimate of variance, we find that

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{20.0 + 40.0}{5 + 6 - 2} \approx 6.67$$

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:

1. State the null and alternative hypotheses.
2. Set the criterion (critical values) for rejecting the null hypothesis.
3. Compute the test statistic.
4. Decide about the null hypothesis and interpret our results.

When stating the null hypothesis, we are assuming that there is no difference between the means of the two independent samples. Therefore, our null hypothesis in this case would be:

$$H_0 : \mu_1 = \mu_2 \qquad \text{or} \qquad H_0 : \mu_1 - \mu_2 = 0$$

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:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1} - s_{\bar{X}_2}}$$

where:

$\bar{X}_1 - \bar{X}_2 =$ the difference between the sample means

$\mu_1 - \mu_2 =$ the difference between the hypothesized population means

$S_{\bar{X}_1 - \bar{X}_2} =$ standard error of the difference between sample means

Let's take a look at an example using these formulas.

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 .05 significance level.

Solution:

First, we would generate our hypotheses based on the two samples.

$$H_0 : \mu_1 = \mu_2$$
$$H_0 : \mu_1 \neq \mu_2$$

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 is so low, we use the t-distribution. If our samples were above 120, we would generally use the z-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 8.6:

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 8.7:

Descriptive Statistic	Sample 1	Sample 2
Number (n)	9	8
Sum of Observations (X)	445	551
Mean of Observations $(\bar{X})$	49.44	68.875
Sum of Standard Deviations $(\sum(X-X)^2)$	862.22	$1,058.88$

Therefore:

$$s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{892.22 + 1058.88}{9 + 8 - 2} \approx 128.07$$

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:

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_{\bar{X}_1-\bar{X}_2}} = \frac{(49.44-68.66)-(0)}{5.50} \approx -3.53$$

Since the difference of -19.22 is 3.53 standard errors below the hypothesized difference of the population mean (zero) and exceeds the critical value of 2.13 standard errors below the mean, we *reject the null hypothesis* and conclude that there *is a significant difference* in the achievement of the students assigned to the 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 120 we generally use the z-distribution).

When solving for the test statistic in large samples, we use the formula:

$$z = \frac{(p_1-p_2)-(P_1-P_2)}{s_{p_1-p_2}}$$

where:

p_1 and p_2 = the observed sample proportions

P_1 and P_2 = the hypothesized population proportions

s_{p1-p2} = 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 (s_{p1-p2}). To calculate this statistic, we use the formula:

$$s_{p1-p2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where:

$$\begin{aligned} p &= \frac{f_1 + f_2}{n_1 + n_2} \\ q &= 1 - p \\ f_1 &= \text{frequency of success in the first sample} \\ f_2 &= \text{frequency of success in the second sample} \end{aligned}$$

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 8.8:

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 .05 level of significance.

Solution:

First, we establish the null and alternative hypotheses:

$$\begin{aligned}H_0 &: P_1 = P_2\\ H_a &: P_1 \neq P_2\end{aligned}$$

Since we have a large sample size ($n > 120$) it is probably best to use the z-distribution. At a .05 level of significance, our critical values are 1.96 standard scores above and below the mean. To solve for the test statistic, we must first solve for the standard error of the difference between proportions.

$$\begin{aligned}p &= \frac{f_1 + f_2}{n_1 + n_2} = \frac{122 + 84}{200 + 150} = \frac{206}{350} = 0.59\\ s_{p_1 - p_2} &= \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.59)(0.41)\left(\frac{1}{200} + \frac{1}{150}\right)} = 0.053\end{aligned}$$

Therefore, the test statistic is:

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{s_{p_1 - p_2}} = \frac{(0.61 - 0.56) - (0)}{0.053} = 0.94$$

Since the test statistic ($z = 0.94$) does not exceed the critical value (1.96), the null hypothesis is *not* rejected. Therefore, we can conclude that the difference in the probabilities (0.61 and 0.56) 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:

1. State the null and alternative hypotheses.
2. Set the criterion (critical values) for rejecting the null hypothesis.
3. Compute the test statistic.
4. Decide about the null hypothesis and 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, $\bar{d} = \bar{X}_1 - \bar{X}_2$ where X equals the mean of the sample.

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:

$$s_d^2 = \frac{\sum(d - \bar{d})^2}{n - 1}$$

(or when simplified)

$$s_d = \sqrt{\frac{\sum(d^2) - \frac{(\sum d)^2}{n}}{n - 1}}$$

where:

$s_d^2 =$ sample variance

$d =$ difference between corresponding pairs within the sample

$\bar{d} =$ the difference between the means of the two samples

$n =$ the number in the sample

$s_d =$ 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:

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}}$$

This may seem a bit confusing, but let's take a look at an example to help clarify.

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:

$$H_0 : \delta = \mu_1 - \mu_2 = 0$$
$$H_0 : \delta = \mu_1 - \mu_2 \neq 0$$

The results for the pre- and post-tests are below:

Table 8.9:

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 standard scores above and below the mean. Since our test statistic of 2.57 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

1. 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.

2. 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:

$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, where $s^2 = \frac{ss_1 + ss_2}{n_1 + n_2 - 2}$

3. We carry out the test 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:

$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$, where $s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 (\frac{1}{n_1} + \frac{1}{n_2})}$

4. 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:

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{s_{p_1 - p_2}}$$

5. 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:

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}}$$

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.

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 8.10:

	Drug A	Drug B
X	75.6	72.8
n	18	24
s^2	12.25	10.24
s	3.5	3.2

3. Does this scenario involve dependent or independent samples? Explain.
4. What would the hypotheses be for this scenario?
5. Compute the pooled estimate for population variance.
6. Calculate the estimated standard error for this scenario.
7. What is the test statistic and at an alpha level of .05 what conclusions would you make about the null hypothesis?

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-smoking scale. The scores are as follows:

Table 8.11:

Husbands	Wives
16	15

Table 8.11: (continued)

Husbands	Wives
20	18
10	13
15	10
8	12
19	16
14	11
15	12

8. What would be the hypotheses for this scenario?
9. Calculate the estimated standard deviation for this scenario.
10. Compute the standard error of the difference for these samples.
11. What is the test statistic and at an alpha level of .05 what conclusions would you make about the null hypothesis?

Review Answers

1. Answers are at the reviewers discretion.
2. True
3. This scenario involves independent samples since we assume that the scores of one sample do not affect the other.
4. $H_0 : \mu_1 = \mu_2, H_a : \mu_1 \neq \mu_2$
5. $s^2 = 11.09$
6. $s_{\bar{X}_1 - \bar{X}_1} = 1.04$
7. The calculate test statistic is 2.69, which exceeds the critical value of $t = 2.021$ scores above or below the mean. Therefore, we would reject the null hypothesis and conclude that it is highly unlikely that the difference between the means of the two samples occurred by chance.
8. $H_0 : \delta = \mu_1 - \mu_2 = 0, H_a : \delta = \mu_1 - \mu_2 \neq 0$
9. $s_d = 3.15$
10. $s_{\bar{d}} = 1.11$
11. The calculated test statistic is 1.13 and with critical values set at $t = 2.365$ scores above or below the mean, we fail to reject the null hypothesis. Therefore, we can conclude that the attitudes towards drinking for married couples are dependent or related to each other.

Image Sources

Chapter 9

Regression and Correlation

9.1 Scatterplots and Linear Correlation

Learning Objectives

- Understand the concept of bivariate data, correlation and the use of scatterplots to display bivariate data.
- Understand when the terms "positive," "negative" "strong," and "perfect" apply to correlation between two variables in a scatterplot graph.
- Calculate the linear correlation coefficient and coefficient of determination using technology tools to assist in the calculations.
- Understand properties and common errors of correlation.

Introduction

So far we have learned how to describe the distribution of a single variable and how to perform hypothesis tests that determine if samples are representative of a population. But what if we notice that two variables seem to be related to one another and we want to determine the nature of the relationship. For example, we may notice that scores for two variables – such as verbal SAT score and GPA – are related and that students that have high scores on one appear to have high scores on another (see table below).

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

Table 9.1: (continued)

Student	*SAT Score*	*GPA*
3	715	3.9
4	405	2.3
5	680	3.9
6	490	2.5
7	565	3.5

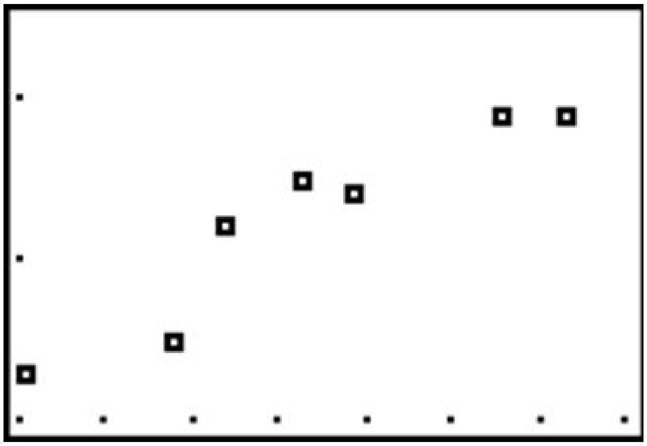
These types of studies are quite common and we can use the concept of **correlation** to describe the relationship between variables.

Bivariate Data, Correlation Between Values and the Use of Scatterplots

Correlation measures the relationship between **bivariate data**. In general, bivariate data are data sets with two observations that are assigned to the same subject. 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?

As mentioned, correlation measures the relationship between two variables. 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.

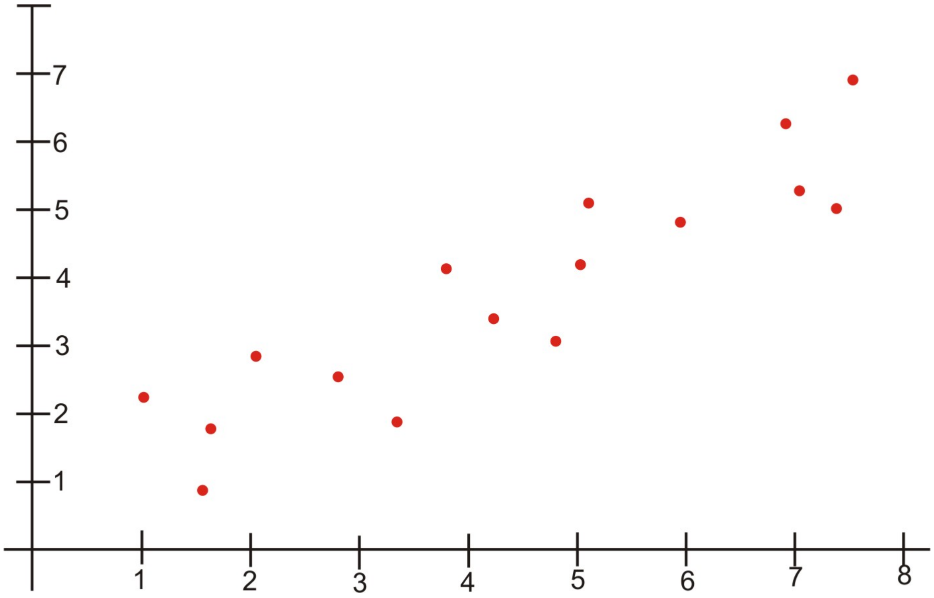
Scatterplots display these bivariate data sets and provide a visual representation of the relationship between variables. In a scatterplot, each point represents a paired measurement of two variables for a specific subject. Each subject is represented by one point on the scatterplot which corresponds to the intersection of imaginary lines drawn through the two observations in the bivariate data set. Therefore, each point represents a paired measurement (see below).



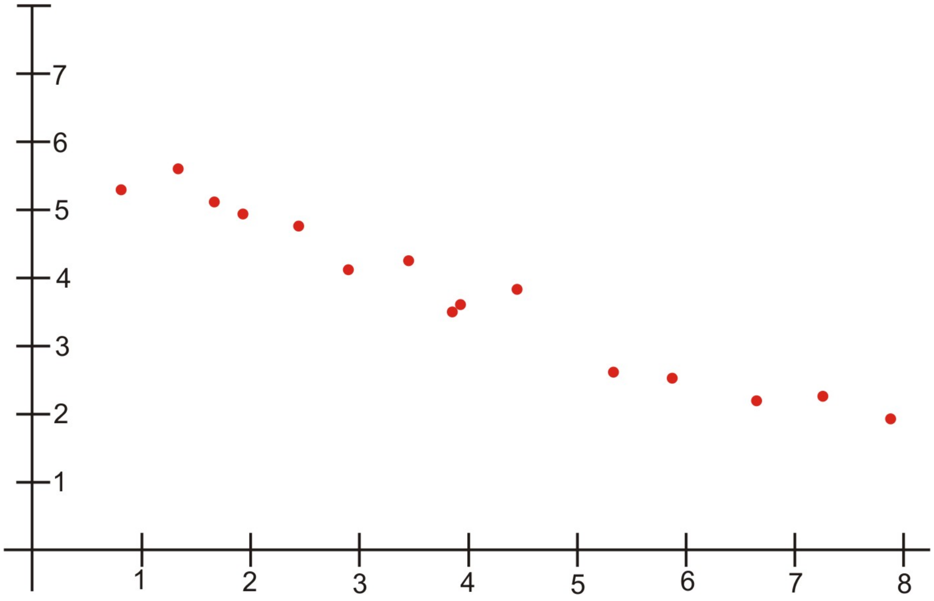
Correlation Patterns in Scatterplot Graphs

Simply examining a scatterplot graph allows us to obtain some idea about the relationship between two variables. Typical patterns include:

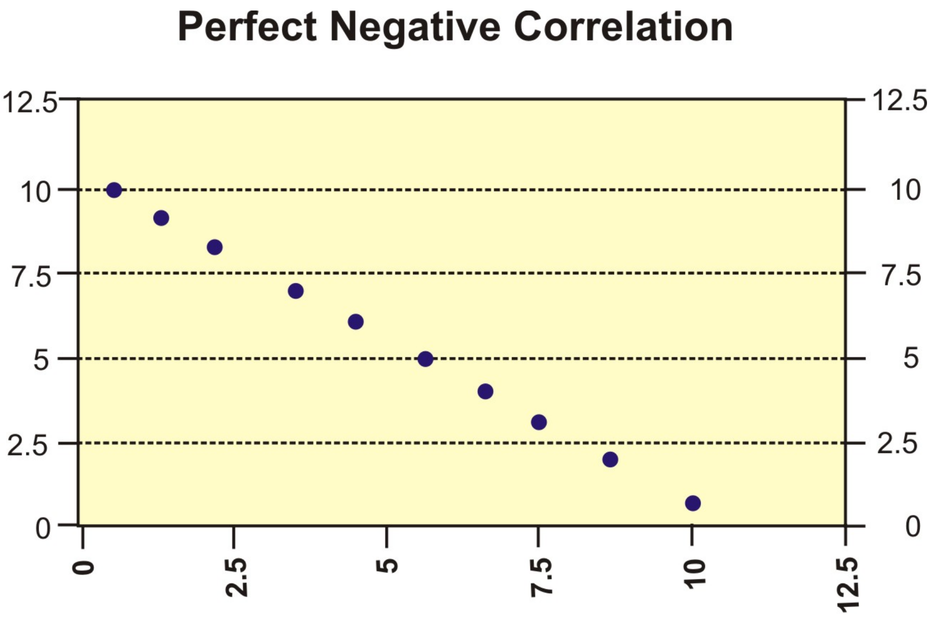
- **A positive correlation** - When the points on a scatterplot graph produce a lower-left-to-upper-right pattern (see below), we say that there is a *positive correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be high as well and vice versa.



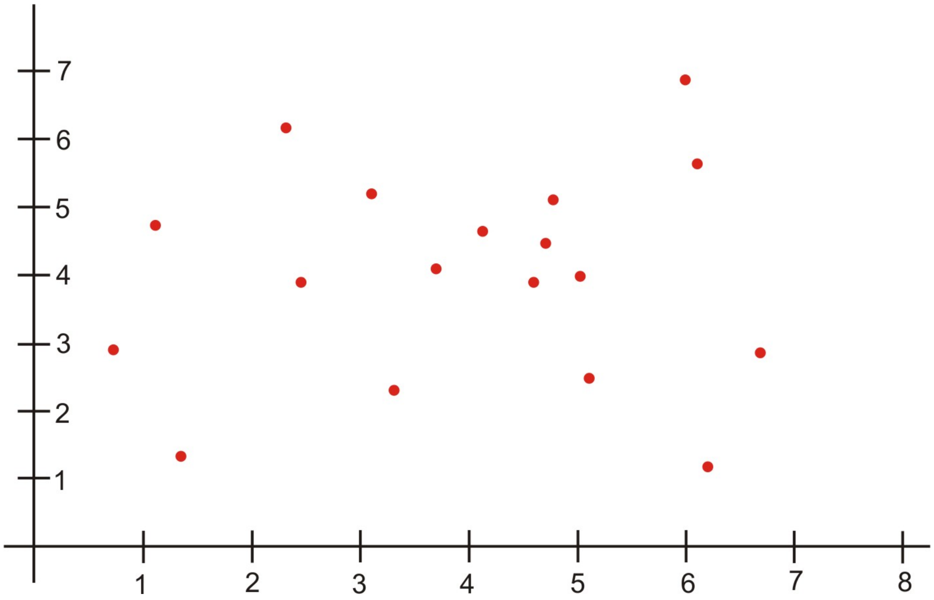
- **A negative correlation** – When the points on a scatterplot graph produce a upper-left-to-lower-right pattern (see below), we say that there is a *negative correlation* between the two variables. This pattern means that when the score of one observation is high, we expect the score of the other observation to be low and vice versa.



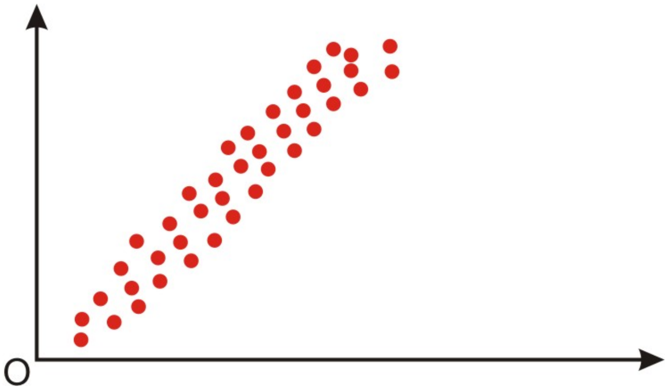
- **A perfect correlation** – If there is a perfect correlation between the two variables, all of the points in the scatterplot will lie on a straight line (see below).



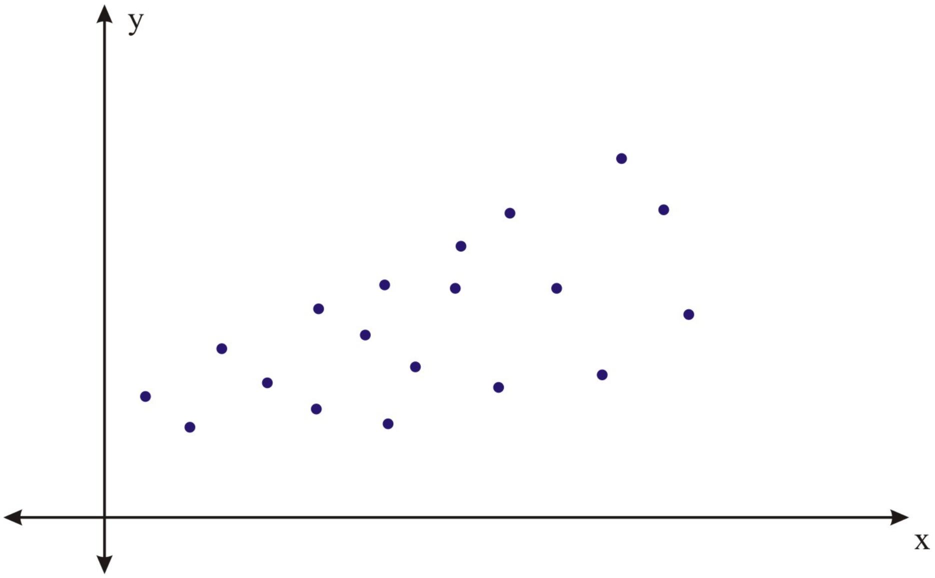
- **Zero correlation** – A scatterplot in which the points do not have a linear trend (either positive or negative) is called a zero or a near-zero correlation (see below).



When examining scatterplots, we also want to look at the **magnitude** of the relationship. If we drew an imaginary oval around all of the points of 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 of two variables, we use a statistic something called the **correlation coefficient** to give us a more precise measurement of the relationship between two variables.The correlation coefficient is an index that describes the relationship between two variables and can take on values between -1.0 and $+1.0$. We can tell a lot from a correlation coefficient including:

- A positive correlation coefficient $(0.10, 0.56, \text{etc.})$ indicates a positive correlation.
- A negative correlation coefficient $(-0.32, -0.82, \text{etc.})$ indicates 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 not mush of a 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.

The most often used correlation coefficient is the **Pearson product-moment correlation coefficient**, or the linear correlation, which 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 them to have a score on Y that is above the mean as well. Pearson developed his correlation coefficient by computing the sum of **cross products** which is multiplying the two scores (X and Y) for each subject and then adding these cross products across the individuals. Then, he divided this sum by the number of subjects minus one. In short, this coefficient is the mean of the cross products of scores.

Because Pearson was measuring the difference between two variables, he used standard scores (z-scores, t-scores, etc.) when determining the coefficient. Therefore, the formula for this coefficient is:

$$r_{xy} = \frac{\sum z_x z_y}{n-1}$$

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.

The equivalent formula that uses the raw scores rather than the standard scores is called the **raw score formula**, which is:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

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 these two variables?

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:

Table 9.3: **Values of**

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

Applying the formula to these data we find:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} = \frac{7 * 13262 - 3970 * 22.7}{\sqrt{[7 * 2321400 - 3970^2][7 * 76.01 - 22.7^2]}}$$
$$= \frac{2715}{2864.22} \approx 0.95$$

The correlation coefficient not only provides a measure of the relationship between the variables, but also gives us an idea about how much of the total variance of one variable can be associated with the variance of another. 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.

The calculation of this variance is called the **coefficient of determination** and is calculated by squaring the correlation coefficient (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. We can think about this concept by examining a series of overlapping circles. The varying degrees of overlap in the circles reflect the proportion of the variance in Y that can be associated with the variance in X. We will study this concept more in depth in later sections.

The Properties and Common Errors of Correlation

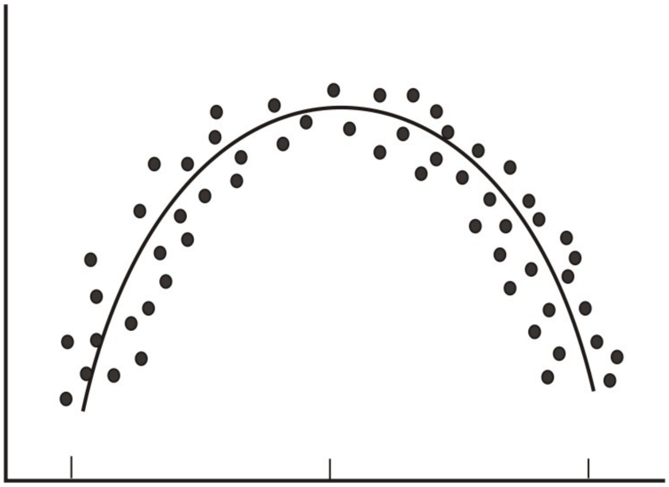
Again, correlation indicates the linear relationship between two variables – it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other things may be causing the two correlated variables to relate as they do. Therefore, it is important to remember that we are interpreting the variables and the variance as not causal, but instead as relational.

When examining correlation, there are three things that could affect our results:

- Linearity
- Homogeneity of the group
- Sample size

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 their performance up to a point (we sometimes call this 'good stress'). However, at that point the increase in the anxiety may cause their 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.



Another error we could encounter when calculating the correlation coefficient is **homogeneity of the group**. When a group is homogeneous or possessing 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 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.

Finally, we should consider sample size. One may assume that the number of observations used in the calculation of the coefficient may influence the magnitude of the coefficient itself. However, this *is not* the case. While the number in the sample size does not affect the coefficient, it may affect the accuracy of the relationship. The larger the sample, the more accurate of a predictor the correlation coefficient will be on the relationship between the two variables.

Lesson Summary

1. **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.
2. When examining **scatterplot graphs**, we can determine if correlations are **positive**, **negative**, **perfect** or **zero**. A correlation is **strong** when the points in the scatterplot are close together.
3. 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 $+1.0$.
4. 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:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

.

5. When calculating correlation, there are several things that could affect our computation including **curvilinear relationships**, **homogeneity** of the group and the size of the group.

Review Questions

1. Please give 2 scenarios or research questions where you would use bivariate data sets.
2. In the space below, please draw and label four scatterplot graphs showing (a) a positive correlation, (b) a negative correlation, (c) a perfect correlation and (4) zero correlation.
3. In the space below, please draw and label two scatterplot graphs showing (a) a weak correlation and (b) a strong correlation.

4. What does the correlation coefficient measure?

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

5. Draw a scatterplot for these data. What type of relationship does this correlation have?
6. Use the raw score formula to compute the Pearson correlation coefficient.

A teacher gives two quizzes to his class of 10 students. The following are the scores of the 10 students.

Table 9.5: **Quiz results for ten students.**

Student	Quiz 1	Quiz 2
1	15	20
2	12	15
3	10	12
4	14	18
5	10	10
6	8	13
7	6	12
8	15	10
9	16	18
10	13	15

7. Compute the Pearson correlation coefficient (r) between the scores on the two quizzes.
8. Find the percentage of the variance (r^2) in the scores of Quiz 2 associated with the variance in the scores of Quiz 1.
9. Interpret both r and r^2 in words.
10. What are the three factors that we should be aware of that affect the size and accuracy of the Pearson correlation coefficient?

Review Answers

1. Various answers are possible. Answers could include scores between two tests, effectiveness of two medications, behavior patterns, etc.
2. Various answers are possible.
3. Various answers are possible.
4. The correlation coefficient measures the nature and the magnitude of the linear relationship between two variables.
5. The scatterplot should show the 5 points plotted in a line. This is a perfect correlation.
6. $r = 1.00$
7. $r = 0.568$
8. $r^2 = 0.323$
9. The correlation between the two quizzes is positive and is moderately strong. Only a small proportion of the variance is shared by the two variables (32.3%)
10. Curvilinear relationships, homogeneity of the group and small group size.

9.2 Least-Squares Regression

Learning Objectives

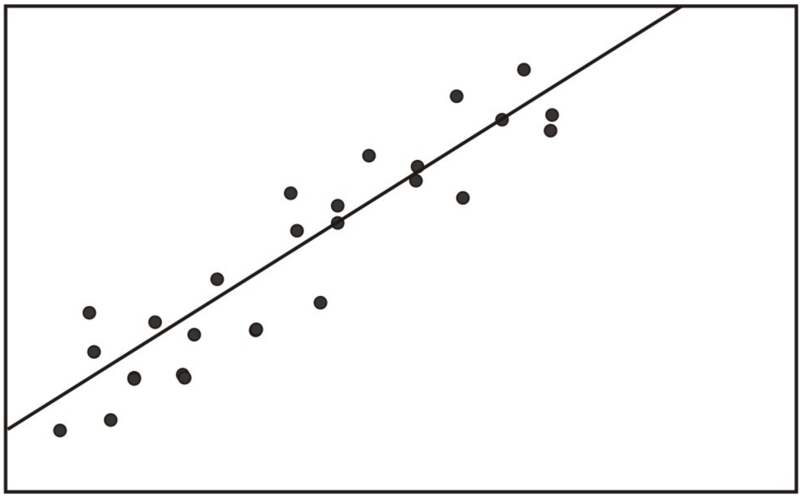
- Calculate and graph a regression line.
- Predict values using bivariate data plotted on a scatterplot.
- Understand outliers and influential points.
- Perform transformations to achieve linearity.
- Calculate residuals and understand the least-squares property and its relation to the regression equation.
- 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 this graph tended to follow a straight line. If we could draw this straight line it, in theory, would represent the change in one variable associated with the other. This line is called the **least squares** or the **linear**

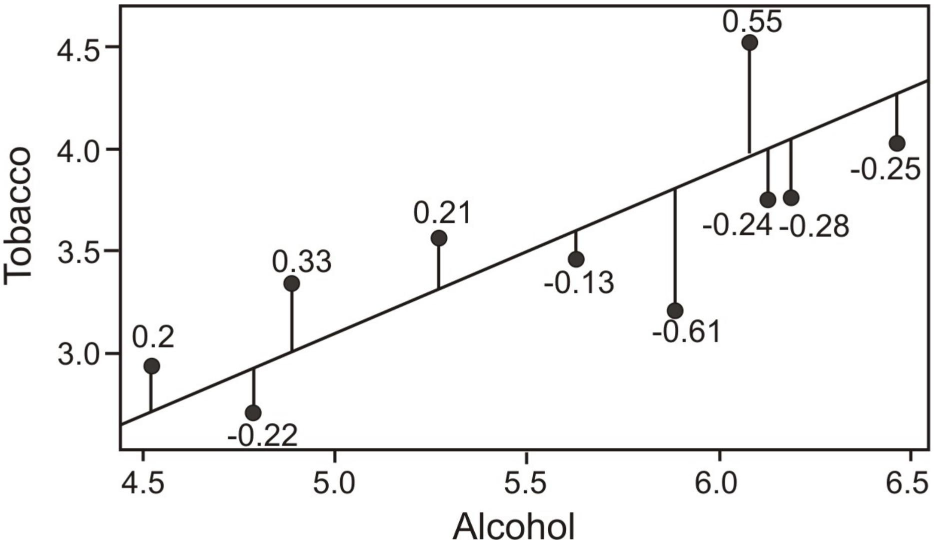
regression line (see figure below).



Calculating and Graphing the Regression Line

Linear regression involves using existing data to calculate a line that best fits the 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** or the **criterion variable**). To calculate this line, we analyze the patterns between two variables and use a series of calculations to determine the different parts of 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 from each of the data points to the line itself. In the example below, you can see the calculated distance from each of the observations to the regression line, or **residual values**. This method of fitting the data line so that there is minimal difference between the observation 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 equivalent to the *slope-intercept form* of the equation for a straight line:

$$Y = bX + a$$

where:

$Y =$ the score that we are trying to predict

$b =$ the slope of the line

$a =$ the Y intercept (value of Y when $X = 0$)

While the linear regression equation is equivalent to the slope intercept form $y = mx + b$ (swapping b for m and a for b), the form above is often used in statistical regression.

To calculate the line itself, we need to find the values for b (the **regression coefficient**) and a (**the regression constant**). The regression coefficient is a very important calculation and explains the nature of the relationship between the two variables. Essentially, the regression coefficient tells us that a certain change in the predictor variables is associated with a 1% change in the outcome or the criterion variable. For example, if we had a regression coefficient of 10.76, we would say that a "10.76% change in X is associated with a 1% change in Y." To calculate this regression coefficient we can use the 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:

$r =$ correlation between variables X and Y

$s_y =$ standard deviation of the Y scores

$s_x =$ 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 squared regression line (also known as the regression line or the line of best fit) for the example measuring the verbal SAT score and GPA that was used in the previous section.

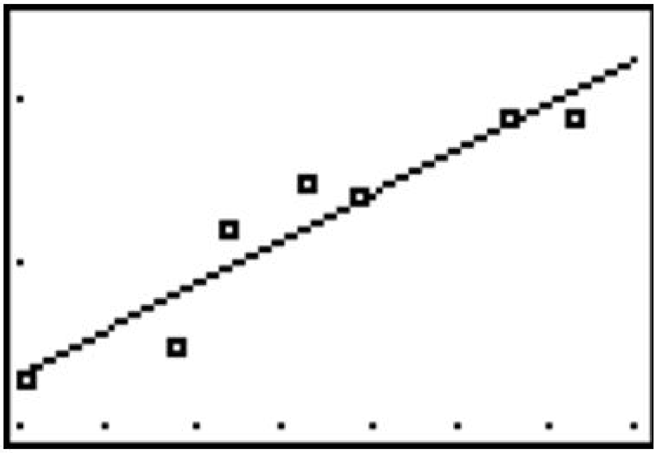
Table 9.6: **SAT and GPA data including intermediate computations for computing a linear regression.**

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, we first calculate the regression coefficient and the regression constant:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{7 \cdot 13,262 - 3,970 \cdot 22.7}{7 \cdot 2,321,400 - 3,970^2} = \frac{2715}{488900} = 0.0056$$
$$a = \frac{\sum Y - b\sum X}{n} \approx 0.097$$

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 several pairs of coordinates. Let's say that we wanted to plot this example on a scatterplot. We would choose two hypothetical values for X (say, 400 and 500) and then solve for Y in order to identify the coordinates $(400, 2.1214)$ and $(500, 2.6761)$. From these pairs of coordinates, we can draw the regression line on the scatterplot.



Predicting Values Using Scatterplot Data

One of the uses of the 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 a students' GPA from their SAT score by plugging in the desired values into our regression equation ($Y = .0056X - 0.07$).

For example, say that we wanted to predict the GPA for two students, one of which had an SAT score of 500 and the other of which 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 (500 and 600) into the equation and solve for Y (see below).

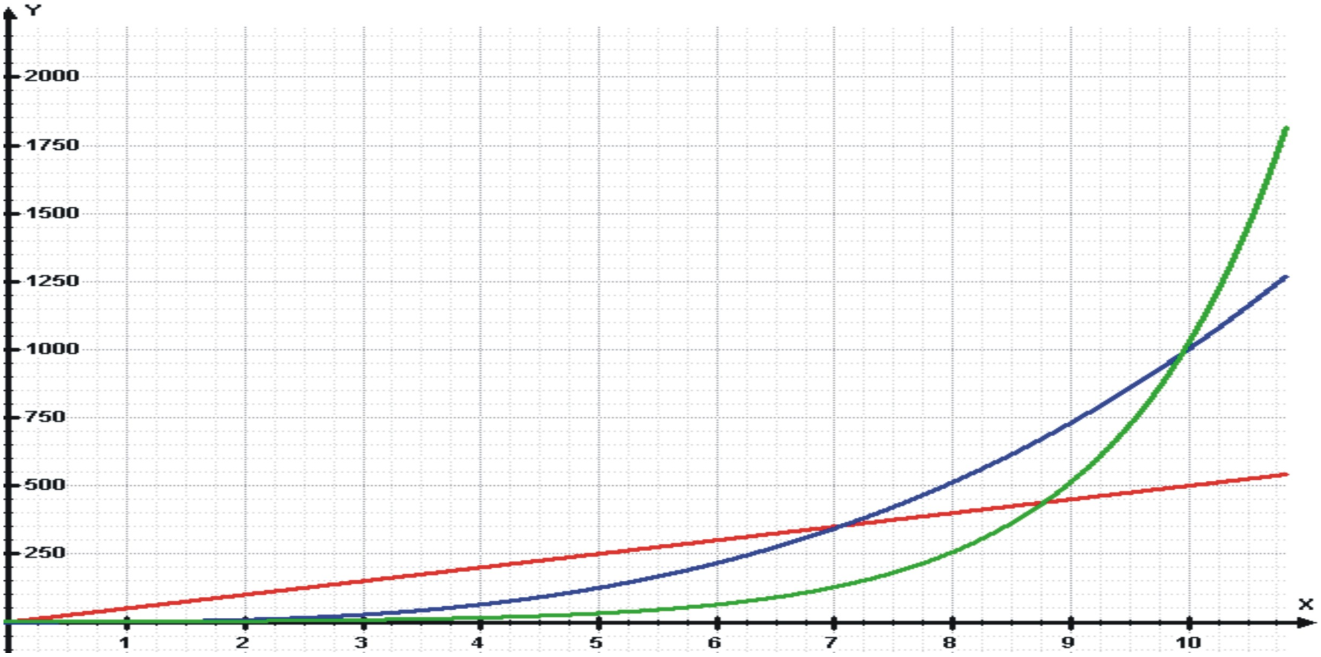
Table 9.7: **GPA/SAT data including predicted GPA values from the linear regression.**

Student	*SAT Score* (X)	*GPA* (Y)	*Predicted GPA* ($\hat{Y}$)
1	595	3.4	3.3
2	520	3.2	2.8
3	715	3.9	3.9
4	405	2.3	2.2
5	680	3.9	3.7
6	490	2.5	2.7
7	565	3.5	3.6
Hypothetical	600		3.4
Hypothetical	500		2.9

We are able to predict the values for Y for any value of X within a specified range.

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, we called these types of relationships **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) or exponential growth (doubling in size with each unit of growth). Below is an example of a growth curve describing the growth of complex societies.



Since this is not a linear relationship, one may think that we may not be able to fit a regression line. However, we can perform something called a **transformation** to achieve a linear relationship. We commonly use transformations in everyday life. For example, the Richter scale measuring for earthquake intensity, and the idea of describing pay raises in terms of percentages are both examples of making transformations on non-linear data.

Let's take a closer look at logarithms so that we can understand how they are used in nonlinear transformations. Notice that we can write the numbers $10, 100$ and $1,000$ as $10 = 10^1, 100 = 10^2, 1,000 = 10^3$, etc. We can also write the numbers $2, 4,$ and 8 as $2 = 2^1, 2 = 2^2, 2 = 2^3$, etc. All of these equations take the form: $x = c^a$ where a is the power to which the base (c) must be raised. We call a the logarithm because it is the power to which the base must be raised to yield the number. Applying this equation, we find that $\log_{10}10 = 1, \log_{10}100 = 2, \log_{10}1000 = 3$, etc. and $\log_2 2 = 1, \log_2 4 = 2, \log_2 8 = 3$, etc. Because of these rules, variables that are exponential or multiplicative (in other words, non-linear models) are linear in their logarithmic form.

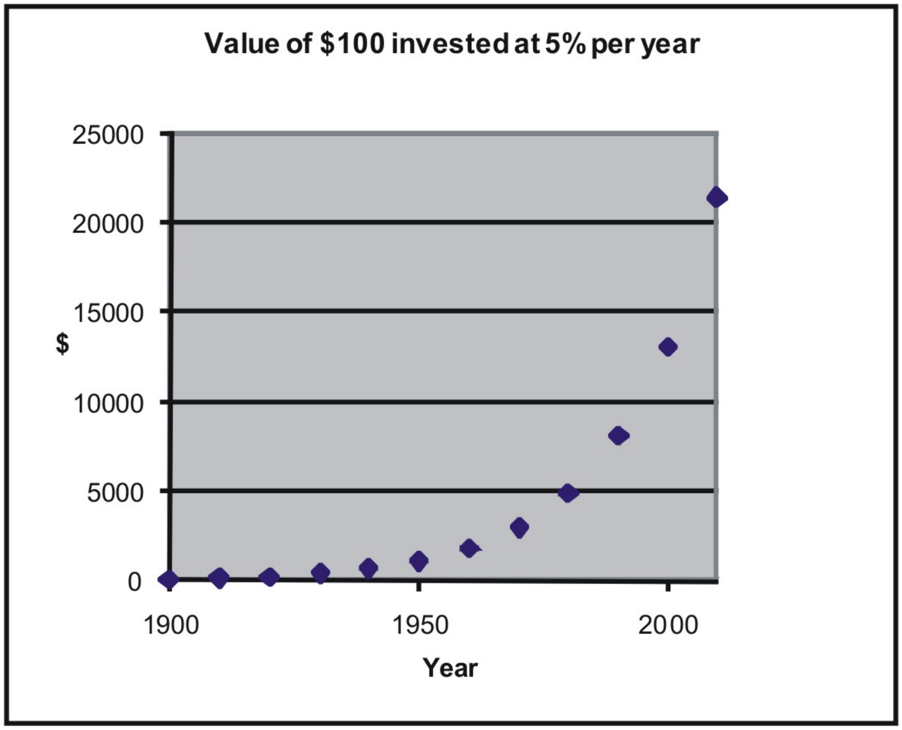
In order to transform data in the linear regression model, we apply logarithmic transformations to each point in the data set. This is most easily done using either the TI-83 calculator or a computer program such as Microsoft Excel, the Statistical Package for Social Sciences (SPSS) or Statistical Analysis Software (SAS). This transformation produces a linear correlation to which we can fit a linear regression line.

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 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**

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.

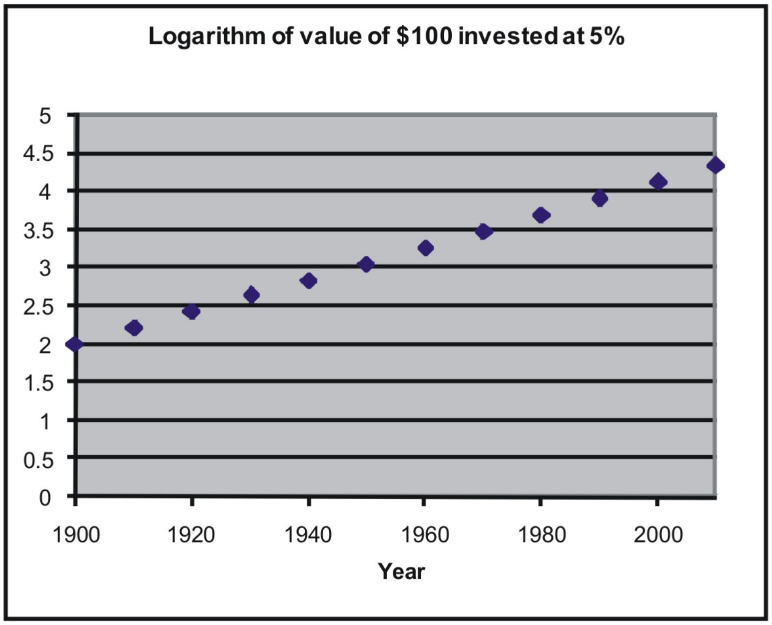


Say that we wanted to fit a linear regression line to these data. First, we would transform these data using logarithmic transformations.

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	4.11893
2010	21420	4.330823

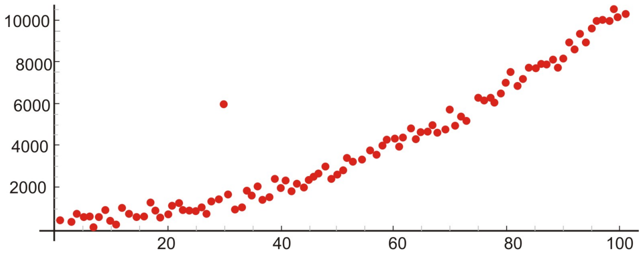
If we graphed these transformed data, we would see that we have a linear relationship.



Outliers and Influential Points

An **outlier** is an extreme observation that does not fit the general correlation or regression pattern (see figure below). By definition, an outlier is defined as an unusual observation;

therefore, the inclusion of this observation may affect the slope and the intercept of the regression line. When examining the scatterplot graph and calculating the regression equation, it is worth considering whether extreme observations should be included or not.

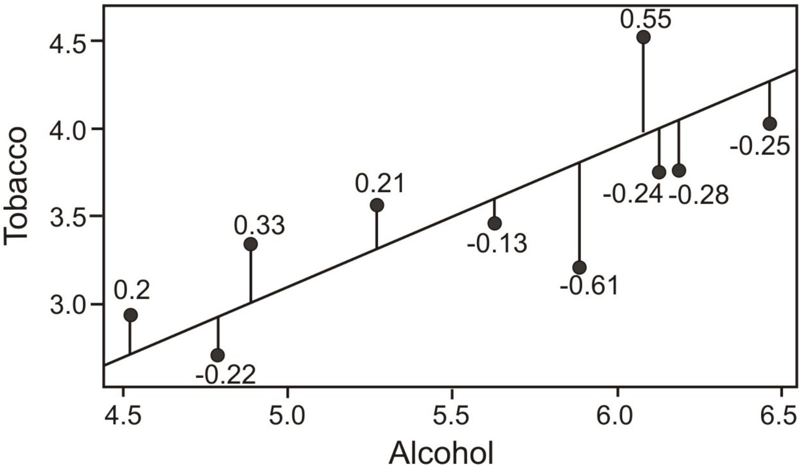


Let's use our example above to illustrate the effect of a single outlier. Say that we have a student that has a high GPA, but suffered from test anxiety the morning of the SAT verbal test and scored a 410. Using our original regression equation, we would expect the student to have a GPA of 2.2. But in reality, the student has a GPA equal to 3.9. The inclusion of this value would change the slope of the regression equation from -0.0056 to -0.0032 which is quite a large difference.

There is no set rule when trying to decide whether or not to include an outlier in regression analysis. This decision depends on the sample size, how extreme the outlier is and the normality of the distribution. As As a general rule of thumb, we should consider values that are 1.5 times the inter-quartile range below the first quartile or above the third quartile as outliers. **Extreme** outliers are values that are 3.0 times the inter-quartile range below the first quartile or above the third quartile.

Calculating Residuals and Understanding their Relation to the Regression Equation

As mentioned earlier in the lesson, the linear regression line is the line that best fits the given data. Ideally, we would like to minimize the distance of all data points to regression line. These distances are called the error (e) and 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 this calculated difference is actually the vertical distance between the observation and the predicted value on the regression line.



To find the residual values we subtract the predicted value from the actual value ($e = Y - \hat{Y}$). Theoretically, the sum of all residual values should be $'0'$ since we are finding the line of best fit with the predicted values as close as possible to the actual value. However, since we will have both positive and negative residuals, it does not make much sense to use this sum as an indicator since the residuals cancel each other out and total 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 scores from our sample above.

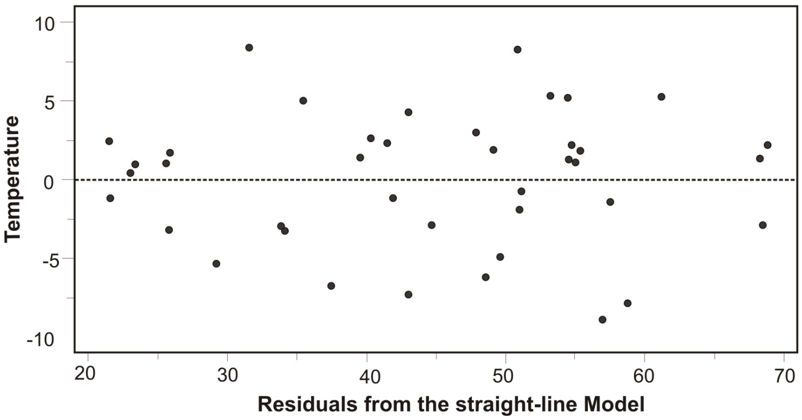
Solution:

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	.2	.04
3	715	3.9	4.1	$-.2$	.04
4	405	2.3	2.3	0	0
5	680	3.9	3.9	0	0
6	490	2.5	2.8	$-.3$	$-.09$
7	565	3.5	3.2	.3	.09
$\sum(Y - \hat{Y})^2$					.26

Plotting Residuals and Testing for Linearity

To test for linearity and when determining if we should drop extreme observations (or outliers) from the 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 too. Below is an example of what a residual scatterplot should look like if there are no outliers and a linear relationship.



If the plots of the residuals do not form this sort of pattern, we should exam them 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 vice versa. If the points are clustered close to the y-axis, we could have an x-value that is an outlier (see below). If this does occur, 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

1. **Prediction** is simply the process of estimating scores on one variable based on the scores of another variable. We use the **least-squares** (also known as the **linear**) **regression line** to predict the value of a variable.
2. 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 $(\ddot{Y})$.
3. When there is a **nonlinear** relationship, we are able to **transform** the data using **logarithmic and power transformations**. Since logarithms and power transformations are exponential in nature, this allows us to produce a linear relationship to which we can fit a regression line.
4. The difference between the actual and the predicted values is called the **residual**

value. We can calculate scatterplots of these residual values to examine **outliers** and test for **linearity**.

Review Questions

The 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
11	4	2
12	1	8
13	1	10
14	1	12
15	2	8

1. Please plot this data on a scatterplot (X axis - Exercise per week; Y axis – Social Events).
2. Does this appear to be a linear relationship? Why or why not?
3. What regression equation would you use to construct a linear regression model?
4. What is the regression coefficient in this linear regression model and what does this mean in words?
5. Calculate the regression equation for these data.
6. Draw the regression line on the scatterplot.
7. What is the predicted memory test score of a student that exercises 3 times per week?
8. Do you think that a data transformation is necessary in order to build an accurate linear regression model? Why or why not?
9. Please calculate the residuals for each of the observations and plot these residuals on a scatterplot.

10. Examine this scatterplot of the residuals. Is a transformation of the data necessary? Why or why not?

Review Answers

1. Answer to the discretion of the teacher.
2. Yes. When plotted, the data appear to be negatively correlated and in a linear pattern.
3. $Y = bX + a$
4. -2.951. This regression coefficient means that every -2.951 percent change in memory test score is associated with a one percent change in exercise per week.
5. $\hat{Y} = -2.951X + 13.65$
6. Answer to the discretion of the teacher
7. If a student exercised 3 times per week, we would expect that they would have a memory test score of 4.8.
8. No. A data transformation is not necessary because the relationship between the two variables is linear.
9. See Table Below.

Table 9.12:

Student	Exercise Per Week	Memory Test Score	Predicted Value	Residual Score
1	0	15	13.7	1.4
2	2	3	7.7	−4.7
3	2	12	7.7	4.3
4	1	11	10.7	0.3
5	3	5	4.8	0.2
6	1	8	10.7	−2.7
7	2	15	7.7	7.3
8	0	13	13.7	−0.7
9	3	2	4.8	−2.8
10	3	4	4.8	−0.8
11	4	2	1.8	0.2
12	1	8	10.7	−2.7
13	1	10	10.7	−0.7
14	1	12	10.7	1.3
15	2	8	7.7	0.3

10. Upon first glance, a transformation of the data is not necessary since the residual values are relatively evenly distributed on the scatterplot. However, it is worth considering dropping several one or two of the outliers (namely observation 7) if they are over three

standard deviations from the mean.

9.3 Inferences about Regression

Learning Objectives

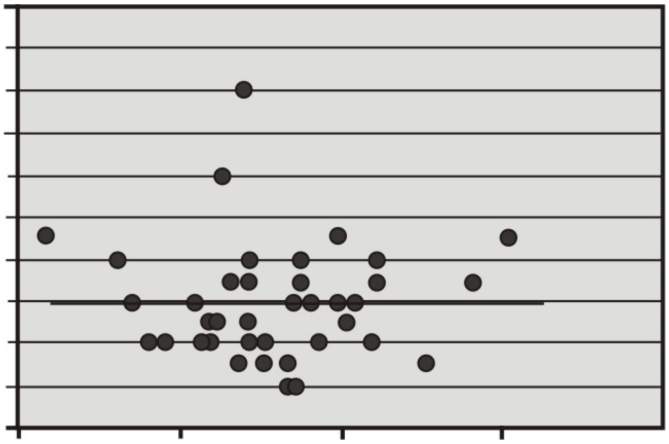
- Make inferences about the regression models including hypothesis testing for linear relationships.
- Make inferences about regression and predicted values including the construction of confidence intervals.
- Check regression assumptions.

Introduction

In the previous section, we learned about the least-squares or the linear regression model. The linear regression model uses the concept of correlation to help us predict a variable based on our knowledge of scores on another variable. As we learned in the previous section, this concept is used quite frequently in statistical analysis to predict variables such as IQ, test performance, etc. 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 two variables (X and Y), then it would be near impossible to fit a meaningful regression line to the points in the scatterplot graph. If there was no correlation and our correlation (r) value was 0, we would always come up with the same predicted value which would be the mean of all the predicted variables (Y). The figure below shows an example of what a regression line fit to variables with no relationship ($r = 0$) would look like. As you can see for any value of X, we always get the same predicted value.



Using this knowledge, we can determine that if there is no relationship between Y and X constructing a regression line or model doesn't help us very much because the predicted score would always be the same. In other words, a regression model would be highly inaccurate. Therefore, when we estimate a linear regression model, we want to ensure that the regression coefficient in the population (β) 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.

$$H_0 : (\beta) = 0$$
$$H_a : (\beta) \neq 0$$

We perform this hypothesis test similar to the previous conducted hypothesis test and need to next establish the critical values for the hypothesis test. We use the t-distribution with $n - 2$ *degrees of freedom* to set such values. The general formula used to calculate the test statistic for testing this null hypothesis is:

$$t = \frac{\text{observed value} - \text{hypothesized or predicted value}}{\text{Standard Error of the statistic}} = \frac{b - \beta}{s_b}$$

To calculate the test statistic for this regression coefficient, we also need to estimate the sampling distributions of the regression coefficients. This statistic about this distribution that we will use is the **standard error of the regression coefficient** (s_b) and is defined as:

$$S_b = \left(\frac{s_{y*x}}{\sqrt{SS_x}} \right)$$

where:

s_{y*x} = the standard error of estimate

SS_x = the sum of squares for the predictor variable (X)

Example:

Let's say that the 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 of $Y = .635X + 1.22$ and the standard error of estimate $s_{Y*x} = .56$. The summary statistics are as follows:

Summarystatisticsfortwofootballfitnesstests.

$n = 24$	$\sum XY = 591.50$
$\sum X = 118$	$\sum Y = 104.3$
$\bar{X} = 4.92$	$\bar{Y} = 4.35$
$\sum X^2 = 704$	$\sum Y^2 = 510.01$
$SS_x = 123.83$	$SS_y = 56.74$

Using a $\alpha = .05$, test the null hypothesis that, in the population, the regression coefficient is zero $(H_0 : \beta = 0)$.

Solution:

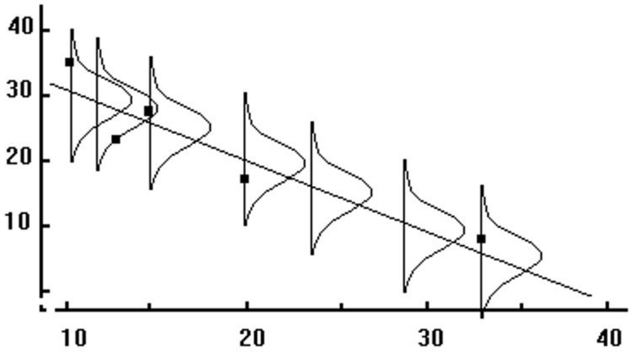
We use the t-distribution for this test statistic and find that the critical values in the t-distribution at 22 degrees of freedom $(n - 2)$ are 2.074 standard scores above and below the mean. Therefore,

$$S_b = \left(\frac{s_{y*x}}{\sqrt{SS_x}} \right) = \left(\frac{.56}{\sqrt{123.83}} \right) = 0.05$$
$$t = \frac{b - \beta}{s_b} = \frac{0.635 - 0}{0.05} = 12.70$$

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 was 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, the regression line simply 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 predicted variable (X).



If we assume that these distributions are normal, we are able to make inferences about each of the predicted scores. One example of making inferences about the predicted scores is identifying probability levels associated with predicted scores. Using this concept, we are able to ask questions such as "If the predictor variable (X value) equals 4.0, what percentage of the distribution of Y scores will be lower than 3?"

The reason that we would ask questions like this depends on the scenario. Say, for example, that we want to know the percentage of students with a 4 on their short physical fitness test that have predicted scores higher than 5. 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. Remember the general formula for calculating the standard score:

$$\text{Test Statistic} = \frac{\text{Observed Statistic} - \text{Population Mean}}{\text{Standard error}}$$

Applying this formula to the regression distribution, we find that the corresponding formula would be:

$$z = \frac{Y - \hat{Y}}{s_{XY}}$$

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 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 they would have a score of 5 or greater on the long physical fitness test?

Solution:

From the regression equation $Y = .635X + 1.22$, we find that the predicted score for $X = 5$ is $Y = 4.40$. Consider the conditional distribution of Y scores for $X = 5$. 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 and find that:

$$z = \frac{Y - \hat{Y}}{s_{Y*X}} = \frac{5 - 4.40}{0.56} = 1.07$$

Using the z-distribution table, we find that the area to the right of a z score of 1.07 is .1423. Therefore, we can conclude that the proportion of predicted scores of 5 or greater given a predicted score of 5 is .1423 or 14.23%.

Confidence 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 value was equal to X, what are the likely values for Y?" This 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 standard error of a predicted score is calculated by using the formula:

$$s_{\hat{Y}} = s_{Y*X}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_x}}$$

The general formula for the confidence interval for predicted scores is found by using the following formula:

$$CI = \hat{Y} \pm (t_{cv} s_Y)$$

where:

$\hat{Y} =$ the predicted score

$t_{cv} =$ critical value of t for $df(n - 2)$

$s_Y =$ standard error of the predicted score

Example:

Develop a 95% confidence interval for the predicted scores from a student that scores a 4 on the short physical fitness exam ($X = 4$).

Solution:

We calculate the standard error of the predicted value using the formula:

$$s_{\hat{Y}} = s_{Y*X}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_x}} = 0.56\sqrt{1 + \frac{1}{24} + \frac{(4 - 4.92)^2}{123.83}} = 0.57$$

Using the general formula for the confidence interval, we find that

$$\begin{aligned} CI &= \hat{Y} \pm (t_{cv} s_Y) \\ CI_{95} &= 3.76 \pm (2.074)(0.57) \\ CI_{95} &= 3.76 \pm 1.18 \\ CI_{95} &= (2.58, 4.94) \\ 2.58 &< CI_{95} < 4.94) \end{aligned}$$

Therefore, we can say that we are 95% confident that given a students' short physical fitness test score (X) of 4, the interval from 2.58 to 4.94 will contain the students' score for the longer physical fitness test.

Regression Assumptions

We make several assumptions under a linear regression model including:

1. *At each value of X, there is a distribution of Y.* These distributions have a mean centered around the predicted value and a standard error that is calculated using the sum of squares.
2. *The best regression model is a straight line.* Using a regression model to predict scores only works if the regression line is a straight line. 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.
3. *Homoscedasticity.* The standard deviations, or the variances, of each of these distributions for each of the predicted values is equal.
4. *Independence of observation.* For each give value of X, the values of Y are independent of each other.

Lesson Summary

1. When we estimate a linear regression model, we want to ensure that the regression coefficient in 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**.
2. For each 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 predicted variable (X). We can use these distributions and the concept of standardized scores to make predictions about probability.
3. 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.
4. We make several assumptions when dealing with a linear regression model including:

- At each value of X, there is a distribution of Y
- The regression model is a straight line
- **Homoscedasticity**
- **Independence** of observations

Review Questions

The 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 they have (including college) and their yearly income (recorded in the thousands). The summary data for this survey are as follows:

$$n = 120 \quad r = 0.67 \quad \sum X = 1,782 \quad \sum Y = 1,854 \quad s_x = 3.6 \quad s_Y = 4.2 \quad SS_x = 1542$$

1. What is the predictor variable? What is your reasoning behind this decision?
2. Do you think that these two variables (income and level of formal education) are correlated? Is so, please describe the nature of their relationship.
3. What would be the regression equation for predicting income (Y) from the level of education (X)?
4. Using this regression equation, predict the income for a person with 2 years of college (13.5 years of formal education).
5. Test the null hypothesis that in the population, the regression coefficient for this scenario is zero.
   (a) First develop the null and alternative hypotheses.
   (b) Set the critical values at $\alpha = .05$.
   (c) Compute the test statistic.
   (d) Make a decision regarding the null hypothesis.
6. For those parents with 15 years of formal education, what is the percentage that will have an annual income greater than $18,500$?
7. For those parents with 12 years of formal education, what is the percentage that will have an annual income greater than $18,500$?
8. Develop a 95% confidence interval for a predicted annual income when a parent indicates that they have a college degree (i.e. - 16 years of formal education).
9. 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?

Review Answers

1. The predictor variable is the number of years of formal education. The reasoning behind this decision is that we are trying to determine and predict the financial benefits of further education (as measured by annual salary) by using the number of years of formal education (the predictor, or the X, variable.
2. Yes. With an r-value of 0.67, these two variables appear to be moderately to strongly correlated. The nature of the relationship is a relatively strong, positive correlation.
3. $Y = 0.782X + 3.842$
4. For $X = 13.5, Y = 14.39$ or $\$14,390$
5. (a) $H_0 : \beta = 0, H_a : \beta \neq 0$

(b) The critical values are set at $t = \pm 1.98$

(c) $S_b = \left(\frac{s_{y*x}}{\sqrt{SS_x}}\right) = \left(\frac{3.12}{\sqrt{1542}}\right) = .08, t = \frac{b-\beta}{s_b} = \frac{0.792-0}{.08} = 9.9$

(d) Since the calculated test statistic of 9.9 exceeds the critical value of 1.98, we decide to reject the null hypothesis and can conclude that the if the null hypothesis was true, we would observe a regression coefficient of 0.792 by chance less than 5% of the time.

6. For $X = 15$, $\hat{Y} = 15.57$. Therefore, 18.50 has a z-value of 0.93:

$$z = \frac{Y - \ddot{Y}}{s_{Y*X}} = \frac{18.5 - 15.57}{3.12} = 0.93$$

The z-value of 0.936 has a corresponding p-value of .1677. This means that with 15 years of formal education, an estimated 16.77% of the parents will have an income greater than $\$18,500$

7. For $X = 12$, $\hat{Y} = 13.2$. Therefore, 18.50 has a z -value of 0.93:

$$z = \frac{Y - \ddot{Y}}{s_{Y*X}} = \frac{18.5 - 13.25}{3.12} = 1.68$$

The z-value of 0.936 has a corresponding p-value of .0465. This means that with 15 years of formal education, an estimated 4.65% of the parents will have an income greater than $\$18,500$

8. $s_{\hat{Y}} = s_{Y*X}\sqrt{1 + \frac{1}{n} + \frac{(X-\hat{X})^2}{SS_x}} = 3.12\sqrt{1 + \frac{1}{120} + \frac{(16-14.85)^2}{1542}} = 3.14$

Using the general formula for the confidence interval ($CI = \hat{Y} \pm (t_{cv} s_Y)$), we find that

$$CI_{95} = 16.35 \pm (1.98)(3.14) = 16.35 \pm 6.22$$
$$CI_95 = (10.13, 22.57)$$

9. Answer is to the discretion of the teacher.

9.4 Multiple Regression

Learning Objectives

- Understand the multiple regression equation and the coefficients of determination for correlation of three or more variables.
- Calculate the 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 has on a university 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, university ranking) using multiple predictor variables (class size and number of faculty members).

Another common use of the multiple regression model 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, etc. Analysts use multiple regression to estimate the selling price in relation to all of these different types of variables.

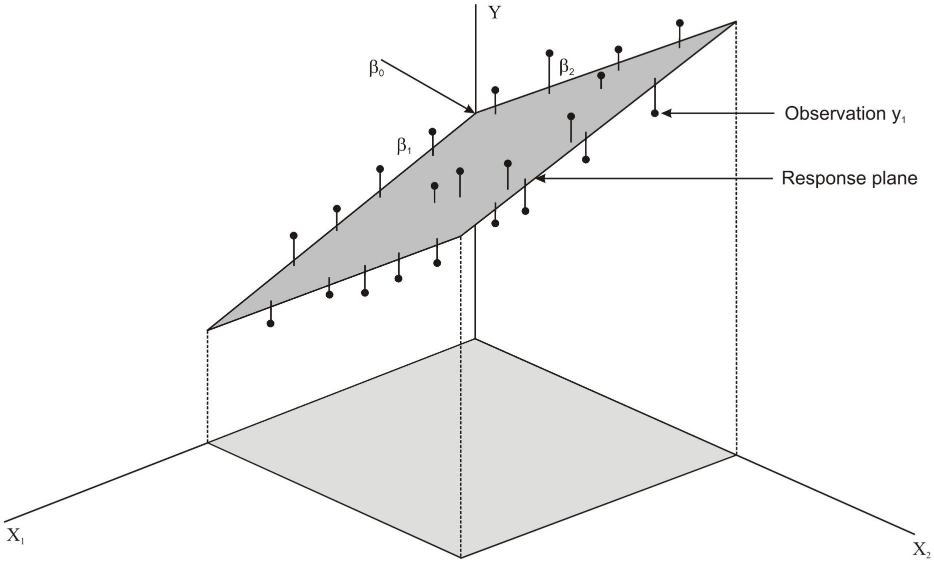
In this section, we will examine the components of the multiple regression equation, calculate the equation using technological tools and use this equation to test for significance to evaluate a hypothesis.

Understanding the Multiple Regression Equation

If we were to try to draw a multiple regression model, it would be a bit more difficult than drawing the 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:

$$\ddot{Y} = b_1X_1 + b_2X_2 + a$$

Since there are three variables, each would have three scores and therefore these scores would 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 that we have, we still use the **least squares** method to try to reduce the distance between the actual and predicted values.



When predicting values using multiple regression, we can also use the standard score form of the formula:

$$z_{\hat{Y}} = \beta_1 z_1 \beta_2 z_2 + \text{etc} \ \ldots$$

where:

$z_{\hat{Y}}$ = the predicted or criterion variable

β = the regression coefficient

z = the predictor variable

To solve for the regression and constant coefficients, we first need to determine the multiple correlation coefficient (r) and **coefficient of determination**, also known as the proportion of shared variance (R^2). In a linear regression model, we measured R^2 by adding the sum of the distances from the actual to the points predicted by the regression line. So what does R^2 look like in a multiple regression model? Let's take a look at the figure above. Essentially, like the linear regression model, the theory behind the computation of the multiple regression equation is to minimize the sum of the squared deviations from the observation to the regression plane.

In most situations, we use the **computer to calculate the multiple regression equation** and determine the coefficients in this equation. We can also do multiple regression on a TI83/84 calculator (this program can be downloaded from http://www.wku.edu/~david.neal/manual/ti83.html). However, it is helpful to explain the calculations that go into the 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 each of the predictor (X) variables:

$$\beta_1 = \frac{r_{Y1} - (r_{Y2})(r_{12})}{1 - r_{12}^2}$$
$$\beta_2 = \frac{r_{Y2} - (r_{Y1})(r_{12})}{1 - r_{12}^2}$$

where:

$\beta_1 =$ the correlation coefficient

$r_{Y1} =$ correlation between the criterion variables (Y) and the first predictor variable (X_1)

$r_{Y2} =$ correlation between the criterion variables (Y) and the second predictor variable (X_2)

$r_{12} =$ correlation between the two predictor variables

After solving for the beta coefficients, we can compute for the b coefficients 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:

$s_Y =$ the standard deviation of the criterion variable (Y)

$S_1 =$ the standard deviation of the particular predictor variable (1 for the first predictor variable and so forth)

After solving for the regression coefficients, we can finally solve for the regression constant by using the formula:

$$a = \bar{Y} - \sum_{i=1}^{k} b_i \bar{X}_i$$

Again, since these formulas and calculations are extremely tedious to complete by hand, we use the computer or TI-83 calculator to solve for the coefficients in the multiple regression equation.

Calculating the Multiple Regression Equation using Technological Tools

As mentioned, there are a variety of technological tools to calculate the coefficients in the multiple regression equation. When using the 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) software. 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 the format, they all provide us with the information needed to build a multiple regression model, 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 models.

Example:

Let's say that 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 the temperature. He measures the average water consumption, temperature and practice time for seven practices and records the following data:

Table 9.13:

Temperature (F)	Practice Time (Hrs)	$H2O$ 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.

Here is the procedure for performing a multiple regression in Excel using this set of data.

1. Copy and paste the table into an empty Excel worksheet
2. Select Data Analysis from the Tools menu and choose "Regression" from the list that

appears

3. Place the cursor in the "Input Y range" field and select the third column.
4. Place the cursor in the "Input X range" field and select the first and second columns
5. Place the cursor in the "Output Range" and click somewhere in a blank cell below and to the left of the table.
6. Click "Labels" so that the names of the predictor variables will be displayed in the table
7. Click OK and the results shown below will be displayed.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.996822
R Square	0.993654
Adjusted R Square	0.990481
Standard Error	1.244877
Observations	7

Table 9.14: **ANOVA**

	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 9.15:

	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

Remember, we can also use the TI-83/84 calculator to perform multiple regression analysis. The program for this analysis can be downloaded at http://www.wku.edu/~david.neal/

manual/ti83.html.

In this excerpt, we have a number of summary statistics that give us information about the model. As you can see from the print out above, we have information for each variable on the regression coefficient (β), the standard error of the regression coefficient $se(\beta)$ and the R^2 value.

Using this information, we can take all of the regression coefficients and put them together to make our model. In this example, our regression equation would be $\hat{Y} = -121.66 + 1.51X + 12.53Z$. Each of these 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 ounce of water consumed if we hold the practice time constant. Similarly, we find that with every 10 minute increase in practice time, we predict players to consume an additional 15 ounces of water if we hold the temperature constant.

With an R^2 of 0.99, 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. Notice that the adjusted R^2 is only slightly different from the unadjusted R^2. This is due to the relatively small number of observations and the small number of predicted variables. With an R^2 of 0.99 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 model, we are looking at whether or not the variables explain most of the variation (R^2) and if this R^2 value is statistically significant. We can use technological tools to conduct a hypothesis test testing the significance of this R^2 value and in constructing 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 ($H_0 = R_{\text{pop}} = 0$). Under this scenario, the predicted or fitted values would all be very close to the mean and the deviations ($\hat{Y} - \bar{Y}$) or the sum of squares would be very small (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 to the effect of "The probability that $R^2 = XX$ would have occurred by chance if the null hypothesis were true is less than .05 (or .10, .01, etc.)." 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 value. We see that we have a very high R^2 value 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 $4.03E - 05(0.0000403)$. This means that the probability that 0.99 of the variance would have occurred by chance if the null hypothesis were true (i.e., none of the variance explained) is 0.0000403. In other words, it is *highly unlikely* that this large level of explained variance was by chance.

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.

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 $1.55E - 05$ and that practice time has a t-statistic of 6.48 and a corresponding p-value of 0.002918. Depending on the situation, we can set our critical values at $0.10, 0.05$, 0.01, etc. For this situation, we will use a p-value of .05. Since both variables have t-values that exceed the critical value, 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 lesson 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 the predictor variables in multiple regression, the confidence interval is based on t-tests and is the range around the observed sample regression coefficient, within which we can be 95% (or any other predetermined level) confident 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

1. In **multiple linear regression**, scores for one variable are predicted using multiple predictor variables. The regression equation we use is

$$Y = b_1X_1 + b_2X_2 + \text{etc.}$$

2. 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.

3. These programs calculate the multiple regression coefficients, combined R^2 value and confidence interval for the regression coefficients.

Supplemental Links

- Manuals by a professor at Western Kentucky University for use in statistics, plus TI-83/4 programs for multiple regression that are available for download.
  - www.wku.edu/~david.neal/web1.html
- Texas Instrument Website that includes supplemental activities and practice problems using the TI-83 calculator
  - education.ti.com/educationportal/activityexchange/activity_list.do

Review Questions

The 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 9.16: **ANOVA**

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2342.7228	780.9076	13.621	.0000
Residual	46	2637.2772	57.3321		
Total	49	4980.0000			

Table 9.17:

	Coefficients	*Standard Error*	*t Stat*	*P−value*
Intercept	10.7592	7.6268		
Test 1	0.0506	.1720	.2941	.7700
Test 2	.5560	.1431	3.885	.0003
Test 3	.2128	.1782	1.194	.2387

1. How many predictor variables are there in this scenario? What are the names of these predictor variables?
2. What does the regression coefficient for Test 2 tell us?
3. What is the regression model for this analysis?
4. What is the R^2 value and what does it indicate?
5. Determine whether the multiple R is statistically significant.
6. Which of the predictor variables are statistically significant? What is the reasoning behind this decision?
7. Given this information, would you include all three predictor variables in the multiple regression model? Why or why not?

Review Answers

1. There are 3 predictor values – Test 1, Test 2 and Test 3.

2. The regression coefficient of 0.5560 tells us that every 0.5560 percent change in Test 2 is associated with a 1.000 percent change in the final exam when everything else is held constant.
3. From the data given, the regression equation is $Y = 0.0506 \quad X_1 + 0.5560 \quad X_2 + 0.2128 \quad X_3 + 10.7592$.
4. The R^2 value is 0.4707 and indicates that 47% of the variance in the final exam can be attributed to the variance of the combined predictor variables.
5. Using the print out, we see that the F statistic is 13.621 and has a corresponding p value of 0.000. This means that the probability that the observed R value would have occurred by chance if it was not significant is very small (slightly greater than 0.000)
6. Test 2. Upon closer examination, we find that only the Test 2 predictor variable is significantly significant since the t value of 3.885 exceeds the critical value (as evidenced by the low p value of .003).
7. No. It is not necessary to include Test 1 and Test 3 in the multiple regression model since these two variables do not have a significant test statistic that exceeds the critical value.

Image Sources

Chapter 10

Chi-Square

10.1 The Goodness-of-Fit Test

Learning Objectives

- Understand the difference between the Chi-Square distribution and the 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 an 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. For example:

- 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; and
- We use **regression analysis** to examine the relationship between two or more continuous variables.

But what test do we run if we are trying to examine patterns between distinct categories such as gender, political candidates, locations or preferences? To analyze patterns like these we use the **Chi-Square test**.

The Chi-Square test is a statistical test used to examine patterns in distinct or categorical variables, which we learned about in the earlier chapter entitled *Planning and Conducting an Experiment or Study.* This test is used in:

1. Estimating how closely a sample matches the expected distribution (also known as the **Goodness-of-Fit test**) and
2. Estimating if two random variables are independent of one another (also known as the **Test of Independence** - see Chapter 9).

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 Goodness-of-Fit test is used to compare the **observed values** of a categorical variable with the **expected values** of that same variable. For example, we would use this test to analyze surveys that contained categorical variables (for example, gender, city of origin, or locations that people preferred to visit on vacation) to determine if there are in fact relationships between certain items.

Example: We would use the Chi-Square Goodness-of-Fit test to evaluate if there was a preference in the types of lunch that 11^{th} 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 to compare the observed and expected values.

Research Question: Do 11^{th} grade students prefer a certain type of lunch?

Using a sample of 11^{th} 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
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 as if school lunch preferences were 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 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. Or, if the difference between the two frequencies is big, then we expect the Chi-Square statistic to be large.

To calculate the Chi-Square statistic (X^2), we use the formula:

$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$ where:

$X^2 =$ Chi-Square statistical value

$O_i =$ observed frequency value for each event

$E_i =$ expected frequency value for each event

Once calculated, we take this Chi-Square value along with the degrees of freedom (this will be discussed later) and look up the Chi-Square value on a standard **Chi-Square distribution** table. The Chi-Square distribution allows us to determine the probability that a sample fits an expected pattern. In contrast, the t-distribution tests how likely it is that the means of two different samples will differ. Please see the table below for more details.

Table 10.2: **The Difference Between the Chi-Square and the Student's t-test when Using to Compare Two Sample Means**

Type of Distribution	Tells Us	Every Day Example	Data Needed to Determine Value
Chi-Square	The relationship between two or more categorical variables.	Analyzing survey data with categorical variables.	Observed and expected frequencies for categorical variables, degrees of freedom.
Student's t-Test	The differences between the means of two groups with respect to a continuous variable.	Determining if there is a difference in the mean of the SAT scores between schools.	The mean values for samples from two populations, degrees of freedom.

Features of the Goodness-of-Fit Test

As mentioned, the Goodness-of-Fit test is used to determine patterns of distinct or **categorical variables**. As we learned in Lesson 6, a categorical variable is one that is not continuous and has observations in separate categories. Examples of categorical variables include:

-gender (male or female)

-preferences (agreed, neutral or disagreed)

-behaviors (got sent to the office or didn't get sent to the office)

-physical traits (straight, wavy or curly hair)

Categorical variables *are not* the same as measurement or continuous variables. The following are normally *not* categorical variables:

– height
– weight
– test scores
– distance
– income

It is important to note that most of these continuous variables could in fact be converted to a categorical variable. For example, you could create a categorical variable with two values such as ¨Less that 10 miles¨ and ¨Greater than 10 miles.¨

In addition to categorical variables, a Goodness-of-Fit test also requires:

-data obtained through a **random sample**

-a calculation of the **Chi-Square statistic** using the formula explained in the last section

-the calculation of the **Degrees of Freedom**. For a Chi-Square test, the Degrees of Freedom are equal to the number of categories minus one or $df = c - 1$

Using our example about the preferences of types of school lunches, we calculate the $df = 3$.

$$\begin{aligned} \text{df} &= \# \text{ of categories} - 1 \\ 3 &= 4 - 1 \end{aligned}$$

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 and 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 the game may not be fair.

Evaluating Hypothesis 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 states "Do 11^{th} 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)***

Using an alpha level of .05, we look under the column for .05 and the row for Degrees of Freedom (remember the Degrees of Freedom = Number of categories −1 = 3). Using the standard Chi-Square distribution table, we see that the critical value for Chi-Square is 7.81. Therefore we would reject the null hypothesis if the Chi-Square statistic is greater than 7.81.

Reject(H_0) **if** $X_2 > 7.81$

Using the table from above, we can calculate the Chi-Square statistic with relative ease.

Table 10.3: **Frequency Which Student Select Type of School Lunch**

Type of Lunch	Observed Frequency	Expected Frequency	$(O-E)^2/E$
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

$$X^2 = \sum \frac{(0-E)^2}{E} = 0.64 + 0.64 + 4.84 + 4.84 = 10.96$$

Since our Chi-Square statistic of 10.96 is greater than 7.81, 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 11^{th} grade students prefer.

As review, we follow the following steps to formulate and evaluate hypothesis:

1. State the null and alternative hypothesis for the research question.
2. Select the significance level and use the Chi-Square distribution table to write a rule for rejecting the null hypothesis.
3. Calculate the value of the Chi-Square statistic.
4. Determine whether to reject or fail to reject the null hypothesis and write a summary statement based on the results.

Lesson Summary

1. We use the **Chi-Square test** to examine patterns between **categorical variables** such as gender, political candidates, locations or preferences.

2. 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.

3. 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.

4. To test the null hypothesis it is necessary to calculate the Chi-Square statistic. To calculate the Chi-Square statistic (x^2), we use the formula:

$$X^2 = \sum_i \frac{(0_i - E_i)^2}{E_i}$$

where:

X^2 = Chi-Square statistical value

O = observed frequency value

E = expected frequency value

5.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.

Supplemental Links

Distribution Tables (including the Student's t-distribution and Chi-Square distribution)

http://www.statsoft.com/textbook/stathome.html?sttable.html&#38;1

Review Questions

1. What is the name of the statistical test used analyze the patterns between two categorical variables?
   (a) the 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? The 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 10.4:

Type of Field Trip	Number Preferring
Sporting Event	53
Play	18
Science Museum	29

6. What is the observed frequency value for the Science Museum category?
7. What is the expected frequency value for the Sporting Event category?
8. What would be the null hypothesis for the situation above?
   (a) There is no preference between the types of field trips that students prefer
   (b) There is a preference between the types of field trips that students prefer
9. What would be the Chi-Square statistic for the research question above?
10. If the estimated Chi-Square level of significance was 5.99, would you reject or fail to reject the null hypothesis?

Review Answers

1. C

2. A
3. B
4. A
5. $X^2 = \sum \frac{(0-E)^2}{E}$
6. 29
7. 33.33
8. A
9. 20.0 (see table below)

Table 10.5:

Type of Field Trip	Observed Frequency	Expected Frequency	Chi-Square
Sporting Event	53	33.33	12.4
Play	18	33.33	7.0
Science Museum	29	33.33	0.6
Chi-Square Total			20.0

10. Reject the Null Hypothesis

10.2 Test of Independence

Learning Objectives

- Understand how to draw and calculate appropriate data from tables needed to run a Chi-Square test.
- Run a Test of Independence to determine whether two variables are independent or not.
- Use a 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 (1) estimate how closely an observed distribution matches an expected distribution **(Goodness-of-Fit test)** or (2) estimating 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, we can test if income and gender are independent, the Test of Independence cannot help us assess how one category might affect the other.

Drawing and Calculating Data from Tables

As mentioned in the previous lesson, tables help us frame our hypotheses and solve problems. Often, we use tables to list the variables and observation patterns that will help us to run the Chi-Square test. For example, we could use a table to record the answers to phone surveys or observed behavior 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. Specifically, we want to know if voting patterns are independent of gender. Hypothetical data for 76 females and 62 males is in the contingency table below.

Table 10.6: **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 Chi-Square Test of Independence is a comparison of the difference between the 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 cell of the table can be calculated using the formula:

$$\text{Expected Frequency} = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Number of Observations}}$$

In the table above, we calculated that the Row Totals are 76 (Females) and 62 (Males) while the Column Totals are 84 (Democrat) and 54 (Republican). Using this formula, we find the following expected frequency for each cell.

Expected Frequency for Female Democratic cell is $76 \times 84/138 = 46.24$

Expected Frequency for Female Republican cell is 76 × 54/138 = 29.74

Expected Frequency for Male Democratic cell is 62 × 84/138 = 37.74

Expected Frequency for Male Republican cell is 62 × 54/138 = 24.26

Using these calculated expected frequencies, we can modify the table above to look something like this:

Table 10.7:

	Democratic	Democratic	Republican	Republican	Total
	Observed	Expected	Observed	Expected	
Female	48	46.26	28	29.74	76
Male	36	37.74	26	24.26	62
Total	84		54		138

Using these figures above, we are able to calculate the Chi-Square statistic with relative ease.

The Chi-Square Test of Independence

As with the Goodness-of-Fit test described earlier, we use similar steps when running a Test-of-Independence. 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 with males. Therefore,

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)***

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:

$$\chi^2 = \sum_i \frac{(0_i - E_i)^2}{E_i}$$

where:

χ^2 = Chi-Square statistical value

O_i = observed frequency value for each event

E_i = expected frequency value for each event

Using this formula and the example above, we get the following expected frequencies and Chi-Square calculations.

Table 10.8:

	Democratic candidate	Democratic candidate	Democratic candidate	Republican candidate	Republican candidate	Republican Candidate
	Obs. Freq.	Exp. Freq.	$(O - E)^2/E$	Obs. Freq.	Exp. Freq.	$(O - E)^2/E$
Female	48	46.26	.07	28	29.74	.10
Male	36	37.74	.08	26	24.26	.12
Totals	84			54		

$$\text{and the Degrees of Freedom} = (C-1)(R-1)$$
$$\text{df} = (2-1)(2-1) = 1$$

Using the table and formula above, we see that the Chi-Square statistic is equal to the sum of all of these values for $(O - E)^2/E$.Therefore,

$$x^2 = .07 + .08 + .10 + .12 = 0.37$$

Using an alpha level of .05, we look under the column for .05 and the row for Degrees of Freedom ($df = 1$). Using the standard Chi-Square distribution table, we see that the critical value for Chi-Square is 3.84.Therefore we would reject the null hypothesis if the Chi-Square statistic is greater than 3.84.

Reject $H_0 : O$ if $X^2 > 3.84$

Since our calculated Chi-Square value of 0.37 is not greater than 3.84, we fail to reject the null hypothesis. Therefore, we can conclude that females are not significantly more likely to vote for democratic candidates than males. In other words, these two factors appear to be **independent** of one another.

Test of Homogeneity

The Chi-Square Goodness-of-Fit and Test of Independence are two ways to examine the relationships between categorical variables. But what test do we use if we are interested in testing whether or not the assignments of these categorical variables are random? We perform the **Test of Homogeneity,** which is computed the same way as the Test of Independence, to examine the randomness of a sample. In other words, the Test of Homogeneity tests whether samples from populations have the same proportion of observations with a common characteristic.

The Test of Homogeneity is used when we examine the probability that the assignment of one variable is equal to another. For example, we found in our last Test of Independence that the factors of gender and voting patterns were independent of one another. However, remember that our original question was if females were more likely to vote for Democratic candidates when compared to males. We would use the Test of Homogeneity to examine the probability that choosing a Democratic candidate was the same for females and males.

Another commonly used example of a Test of Homogeneity is comparing dice to see if they all work the same way. Let's use that example to conduct a sample Test of Homogeneity.

Example: A manager of a casino has two potentially 'loaded' ('loaded dice' are ones that are weighted on one side so that certain numbers have greater probabilities of showing up) that they want to examine. The manager rolls each of the dice exactly 20 times and comes up with the following results.

Table 10.9: **Number Rolled on the Potentially Loaded Dice**

	1	2	3	4	5	6	Totals
Dice 1	6	1	2	2	3	6	20
Dice 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 hypothesis based on a research question. In this case, our research question would look something like: "Is the probability of rolling a specific number the same for Dice 1 and Dice 2?" This would give us the following hypotheses:

Null Hypothesis $(H_0 : O) = E$ *(The probabilities are the same for both die)*

Alternative Hypothesis $(H_a : O) \neq E$ *(The probabilities differ for both die)*

Similar to the other test, we need to calculate the expected values for each cell and the total number of Degrees of Freedom. To get the expected frequency for each cell, we use the same formula as we used for the Test of Independence:

$$\text{Expected Frequency} = \frac{\text{(Row Total)(Column Total)}}{\text{Total Number of Observations}}$$

The following table has includes the Expected Frequency (in parenthesis) for each cell along with the Chi-Square statistic ($(O-E)^2/E$) in a separate column.

Number Rolled on the Potentially Loaded Dice

Table 10.10:

	1	X^2	2	X^2	3	X^2	4	X^2	5	X^2	6	X^2	X^2 Total
Dice 1	6(7.5)	0.3	1(1)	0	2(2.5)	.1	2(2.5)	.1	3(2)	.5	6(7)	.2	1.2
Dice 2	4(7.5)	1.6	1(1)	0	3(2.5)	.1	3(2.5)	.1	1(2)	.5	8(7)	.2	2.5
Totals	10		2		5		5		4		14		

$$\text{and the Degrees of Freedom} = (C-1)(R-1)$$
$$\text{df} = (6-1)(2-1) = 5$$

Using the same Chi-Square formula and the information from the table above, we find that:

$$X^2 = .1.2 + 2.5 = 3.7$$

Using an alpha level of .05,we look under the column for .05 and the row for Degrees of Freedom ($df = 5$). Using the standard Chi-Square distribution table, we see that the critical value for Chi-Square is 11.07. Therefore we would reject the null hypothesis if the Chi-Square statistic is greater than 11.07.

Reject$(H_0 : O)$ **if** $X^2 > 11.07$

Since our calculated Chi-Square value of 3.7 is not greater than 11.07, 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 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

1. The Chi-Square Test of Independence is used to assess if 2 factors are related. It is commonly used in social science research to examine behaviors, preferences, measurements, etc.

2. As with the Chi-Square Goodness-if-Fit test, tables help capture and display relevant information.

3. For each cell in the table constructed to run a chi-square test, we need to calculate the **expected frequency.** The formula used for this calculation is:

$$\text{Expected Frequency} = \frac{\text{(Row Total)(Column Total)}}{\text{Total Number of Observations}}$$

4. To calculate the Chi-Square statistic for the **Test of Independence,** we use the same formula as the Goodness-of-Fit test. If the calculated Chi-Square value is greater than the critical value, we reject the null hypothesis.

5. 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.

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 cell in a contingency table, you use the formula:

   (a)

$$\text{Expected Frequency} = \frac{\text{(Row Total)(Column Total)}}{\text{Total Number of Observations}}$$

   (b)

$$\text{Expected Frequency} = \frac{\text{(Total Observations)(Column Total)}}{\text{Row Total}}$$

   (c)

$$\text{Expected Frequency} = \frac{\text{(Total Observations)(Row Total)}}{\text{Column Total}}$$

Please use the table below to answer the following review questions.

Table 10.11: **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

4. What is the total number of females in the sample?
   (a) 450
   (b) 280
   (c) 612
   (d) 782
5. What is the total number of observations in this sample?
   (a) 782
   (b) 533
   (c) $1,062$
   (d) 612
6. What is the expected frequency for the number of males that did not study abroad?
   (a) 161
   (b) 208
   (c) 111
   (d) 129
7. How many Degrees of Freedom are in this example?
   (a) 1
   (b) 2
   (c) 3
   (d) 4
8. True or False: Our null hypothesis would be that females are as likely as males to study abroad.
9. What is the Chi-Square statistic for this example?
   (a) 1.60
   (b) 2.45
   (c) 3.32
   (d) 3.98
10. If the Chi-Square critical value at .05 and 1 degree of freedom is 3.81 and we have a calculated Chi-Square value of 2.22, we would:
    (a) reject the null hypothesis
    (b) fail to reject the null hypothesis

11. True or False: We use the Test of Homogeneity to evaluate the equality of several samples of certain variables.
12. The Test of Homogeneity is carried out the exact same way as:
    (a) The Goodness-of-Fit test
    (b) The Test of Independence

Review Answers

1. To examine if two variables are related.
2. True
3. A
4. D
5. C
6. A
7. A
8. True
9. A
10. B
11. True
12. B

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. But 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. Often times when we test variance we are assessing whether or not a **sample mean** differs from the population mean by more than we would expect due to chance. This test is somewhat similar to the test of z-scores where we measure the likelihood that a single observation came from a population but a bit different since we are using **samples** instead of **individual** observations.

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 single-sample Chi-Square test, the F-test and the *An*alysis *o*f *Va*riance (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 the testing of a single variance using the Chi-Square test in greater detail.

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. This testing 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.

To test a hypothesis about a single 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, which we learned how to calculate in previous lessons. For the purposes of this exercise, we will assume that we will be provided the standard deviation and the population variance.

Using these key pieces of information, we use the following formula to caluclate the Chi-Square value to test hypothesis surrounding single variance:

$$X^2 = \frac{df s^2}{\sigma^2}$$

where:

X^2 = Chi-Square statistical value

df = degrees of freedom = $N - 1$, where N = size of the sample

s^2 = sample variance

σ^2 = population variance

Similar to the $z-$test, we want to test a hypothesis that the sample comes from a population with a variance greater than the obseved 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 for height 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 the population that has a variance of 4.84 (σ^2 = the standard deviation of the overall population squared or 4.84). Therefore:

Null Hypothesis $H_0 : \sigma^2 \leq 4.84$ *(the variance of the sample is greater than or equal to that of the population)*

Alternative Hypothesis $H_a : \sigma^2 > 4.84$ *(the variance of the sample is less than that of the population)*

Using the sample of the 41 gymnasts, we compute the standard deviation and find it to be $1.2(s = 1.2)$. Using the information from above, we can calculate our Chi-Square value and find that:

$$= X^2 = \frac{df s^2}{\sigma^2} = (40 \cdot 1.2^2)/4.84 = 11.90$$

Therefore, since 11.90 is less than 55.76, we fail to reject the null hypothesis and therefore cannot conclude this sample female gymnasts has significantly higher variance in height when compared to 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.

Similar to constructing confidence intervals in other types of tests, we construct a confidence interval when testing a population variance to identify a range that we think will encompasses the variance. To construct a confidence interval for the population variance, we need three pieces of information: the number of observations in a sample, the variance of the sample, and the desired confidence interval. With the desired confidence interval (most often this is set at 90 or 95%), we can construct the upper and lower limits around the significance level.

To construct the upper limit of the confidence interval, we set the value equal to $\alpha/2$ (alpha is the Greek letter "a") where α = probability that the variance is *not in* the interval) and the lower limit to $(1 - (\alpha/2))$. Therefore, when constructing a 90% confidence interval ($\alpha = 0.1$)we would find that the two limits of the confidence interval would be at 0.05 ($\alpha/2$) and 0.95($1 - (\alpha/2)$). Similarly, a 98% confidence interval ($\alpha = 0.02$) would have limits set at 0.01 and 0.99. Using these limits and the number of degrees of freedom from the sample, we can use the standard Chi-Square distribution table to look up actual values to construct our confidence interval for population variance. Let's look at an example to help clarify.

Example: We randomly select 30 samples of Coca Cola and measure the amount of sugar in each sample. Using the formula that we learned earlier, we calculate that the variance of the sample is 5.20. What would be the population variance with a 90% confidence interval? In other words, if we were to repeatedly draw random samples 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:

$$X^2_{0.05} \leq \frac{df s^2}{\sigma^2} \leq X^2_{0.95}$$

Using our standard Chi-Square distribution table, we can look up the critical X^2 values for 0.05 and 0.95 at 29 Degrees of Freedom. Using our X^2 distribution table, we find that $X^2_{\{0.05\}}$ and that $X^2_{\{0.05\}} = 17.71$. Since we know the number of observations and the standard deviation for this sample, we can then solve for σ^2:

$$\begin{gathered}\frac{\text{dfs}^2}{42.56} \leq \sigma^2 \leq \frac{\text{dfs}^2}{17.71}\\ \frac{295.20}{42.56} \leq \sigma^2 \leq \frac{295.20}{17.71}\\ 3.54 \leq \sigma^2 \leq 8.51\end{gathered}$$

In other words, we are 90% confident that the population variance of this sample is between 3.54 and 8.51.

Lesson Summary

1. 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.

2. To test the variance using the Chi-Square statistic, we use the formula

$$X^2 = \frac{df s^2}{\sigma^2}$$

where:

X^2 = Chi-Square statistical value

df = Degrees of Freedom = $N - 1$, where N = size of the sample

s^2 = sample variance

σ^2 = 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.

3. We can construct a confidence interval which is a range of values that includes the population variance with a given degree of confidence. To find this interval, we use the formula.

$$X^2_{\frac{\alpha}{2}} \leq \frac{\mathrm{dfs}^2}{\sigma^2} \leq X^2_{1-\frac{\alpha}{2}}$$

For example, if $\sigma = 0.1$, the range is a 90% interval, from 0.05 to 0.95. We then say that the probability is 10% that the population variance is not in the resulting interval.

Review Questions

1. We use the Chi-Square distribution for the:
   (a) Goodness-of-Fit test
   (b) Test for Independence
   (c) Testing 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, our null hypothesis states that the two population means that we are testing are:
   (a) equal with respect to variance
   (b) are not equal
   (c) none of the above
4. In the formula for calculating the Chi-Square statistic for single variance, σ^2 =:
   (a) standard deviation
   (b) number of observations
   (c) hypothesized population variance
   (d) Chi-Square statistic

5. If we knew the number of observations in the 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 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 X^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 ($a = 0.1$) are 17.71 and 42.56, what is the range for σ^2?
10. What statement would you give surrounding this Confidence Interval?

Review Answers

1. D
2. False
3. A
4. C
5. D
6. The null hypothesis states that the sample comes from a population with a variance less than or equal to the population variance of 3.22 ($H_0 : O$) $\sigma^2 \leq 3.22$
7.

$$X^2 = \frac{\text{dfs}^2}{\sigma^2} = (29 \times 1.1^2)/3.22 = 10.90$$

8. Failure to Reject the Null Hypothesis since $3.22 \leq 10.90$. We cannot conclude that the sample comes from the larger population with respect to variance.
9. Between 0.82 and 1.98

$$\frac{\text{dfs}^2}{42.56} \leq \sigma^2 \leq \frac{\text{dfs}^2}{17.71}$$
$$\frac{295.20}{42.56} \leq \sigma^2 \leq \frac{295.20}{17.71}$$
$$0.82 \leq \sigma^2 \leq 1.98$$

10. We are 90% confident that the overall population variance (σ^2) is between 0.82 and 1.98

Image Sources

Chapter 11

Analysis of Variance and the F-Distribution

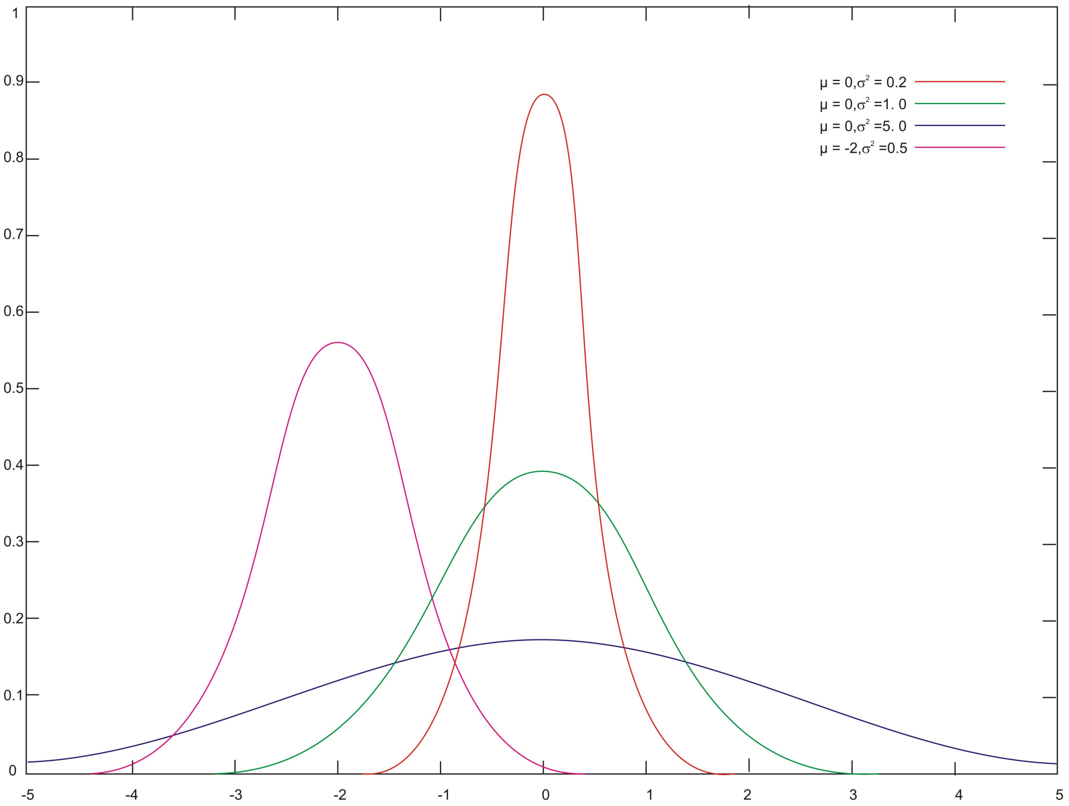
11.1 The F-Distribution and Testing Two Variances

Learning Objectives

- Understand the differences between the F- and the Student's t-distributions.
- 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 examining 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 difference variances.



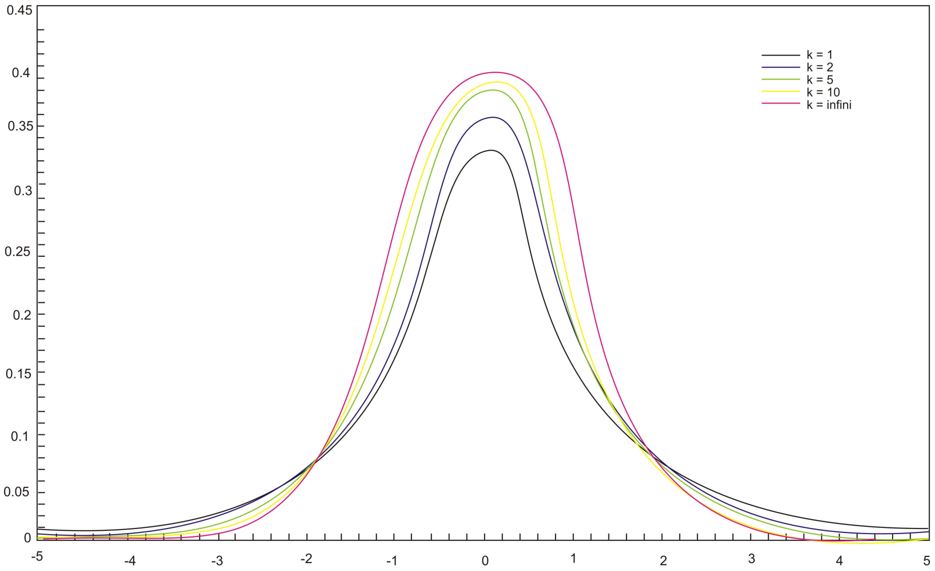
So why would we want to conduct a hypothesis test on variance? Let's consider an example. Say that a teacher wants to examine the effectiveness of two reading programs. She randomly assigns her students into two groups, uses the different reading programs with each group and gives her students an achievement 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*- and Student's *t*-distributions, calculate the test statistic and test hypotheses about multiple population variances. In addition, we will look a bit more closely at the limitations of this test.

Differences between the F- and Student's t-Distributions

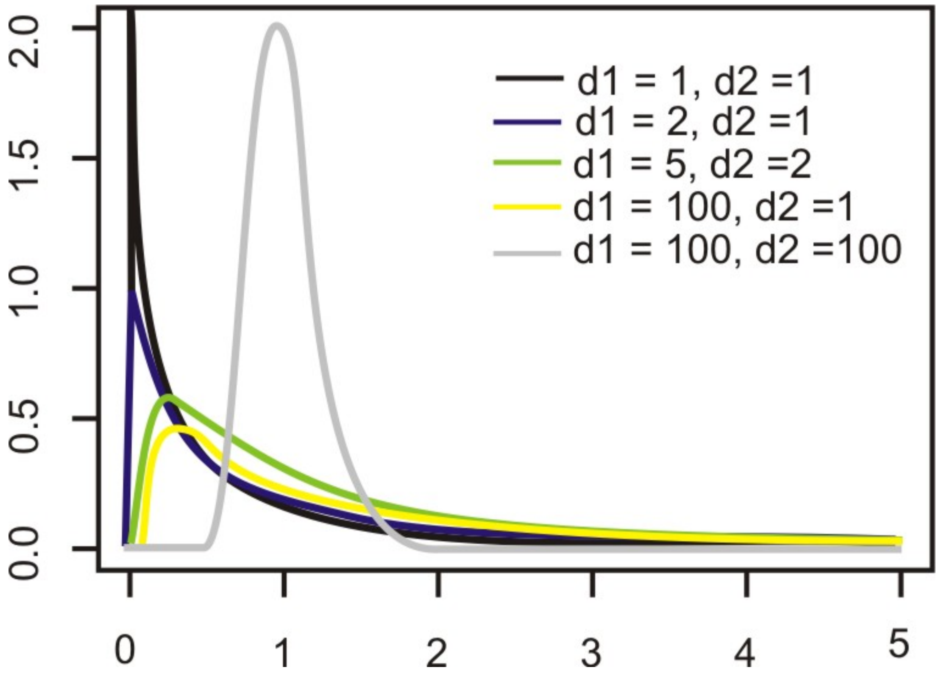
As review, we use the Student's *t*-distribution when we are conducting hypotheses tests where the variance of the population is unknown. Usually, the variance of the population is *not* known and it is necessary to estimate it by using the variance of the sample. Using the variance of a sample to estimate population variance can be inappropriate – especially if we have a small sample size. For estimating the population variance from a small sample we use a statistical tool called the **Student's *t*- distribution.**

The Student's t-distribution is a family of distributions that, like the normal distribution, are symmetrical, bell-shaped and centered on the mean. The shape of these distributions changes as the sample sizes changes (see below) and each t-distribution is associated with a unique number of Degrees of Freedom (number of observations in the sample minus one). As the number of observations (shown by k in the figure) increases, the difference between the t-distribution and the normal distribution (in pink) decreases.



The F-distribution is quite a bit different. When we test the hypothesis that two variances in the 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 $({\sigma_1}^2)/({\sigma_2}^2)$ equals 1.00), we call this test the F- **Max test.**

Since we are testing ratios, the F-distribution looks quite different from the Student's t-distribution (see below). Like the Student's t-distribution, the F-distribution is a family of distributions. The specific F-distribution for testing two population variances $H_0 : {\sigma_1}^2 = {\sigma_2}^2$ is based on two Degrees of Freedom (one for each of the populations). Unlike the normal and the t-distributions, the F-distributions are not symmetrical and span only non-negative numbers (unlike others that are symmetric and have both positive and negative values.) In addition, the shapes of the F-distribution vary drastically, especially when the degrees of freedom values are small. These characteristics make determining the critical values for the F-distribution more complicated than for the normal and Student's t-distributions.



F- Max Test: Calculating the Sample Test Statistic

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?

Solution:

$$F = \frac{{s_1}^2}{{s_2}^2} = \frac{105.96}{36.42} \approx 2.909$$

F-Max Test: Testing Hypotheses about Multiple Independent Population Variances

As mentioned, in certain situations we are interested in determining if there is a difference in the population variances between two independent samples. We can conduct a hypothesis test of no difference between the population variances with the null hypothesis of $H_0 : {\sigma_1}^2 = {\sigma_2}^2$. Therefore, 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 are in the right-hand tail (please see the supplemental links for an example of the table). We also need to use the degrees of freedom from **each** of the samples to determine the critical values.

Say, for example, that we are trying to determine the critical values for the scenario above and we set the level of significance at $.02(\alpha = .02)$. Because we have a two-tailed test, we assign .01 to the area of the right of the critical value. Using the F-table for $\alpha = .01$ (for example, see http://www.statsoft.com/textbook/sttable.html#f01) , we find the critical value at 2.20 ($df = 30$ and 40 for the numerator and denominator with a $\alpha = .01$ to the area to the right of the tail).

Once we set 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 and the Student's t-distributions.

Example:

Using our example above, 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.

Solution:

In the example above, we calculated an F ratio of 2.909 and found a critical value of 2.20.

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%(.02) of the time. We can conclude that the variance of the student achievement scores for the second sample is less than the variance for the students in the first sample. We can also see that the achievement test means are practically equal so the variance in student achievement scores may help the teacher in her selection of a program.

The Limits of Using the F-Distribution to Test Variance

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. 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

1. We use the F-Max test and the F-distribution when testing if two variances from independent samples are equal.
2. The F-distribution differs from the Student's t-distribution. Unlike the normal and the t-distributions, the F-distributions are not symmetrical and go from zero to infinity (∞) not from $-\infty$ to ∞ as the others do.
3. When testing the variances from independent samples, we calculate the F-ratio, which is the ratio of the variances of the independent samples.
4. 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.
5. 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.

Supplemental Links

- Distribution Tables http://www.statsoft.com/textbook/sttable.html

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- and the Student's t-distributions.
3. When we test the differences between the variance 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.

Suppose the 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
$n = 31$	$n = 21$
$s^2 = 42.30$	$s^2 = 18.80$

5. What are the null and alternative hypotheses for this scenario?
6. What is the critical value with a $\alpha = .10$?
7. Calculate the F-ratio.
8. Would you reject or fail to reject the null hypothesis? Explain your reasoning.
9. Interpret the results and what the guidance counselor can conclude from this hypothesis test.
10. 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.

Review Answers

1. Variance
2. Answers may vary but could include:
   (a) We use the t-distribution when testing the difference between the means of two independent samples and the F-distribution when testing the difference between the variances of two independent samples.
   (b) The t-distribution is based off of one degree of freedom and the F-distribution is based off of two.
   (c) F-distributions are not symmetrical, t-distributions are.
   (d) T-values range from $-\infty$ to ∞ while F-ratios range from zero to ∞
3. F-ratio
4. larger, smaller
5. $H_0 : {\sigma_1}^2 = {\sigma_2}^2$ or ${\sigma_1}^2/{\sigma_2}^2 = 1, H_a : {\sigma_1}^2 \neq {\sigma_2}^2$ or ${\sigma_1}^2/{\sigma_2}^2 \neq 1$
6. 2.04
7. 2.25
8. We would reject the null hypothesis because the calculated F ratio (2.25) exceeds the critical value (2.04).
9. We can conclude that the variance of the student achievement scores for the second sample is less than the variance for the students in the first sample. Since the achievement test means are practically equal, the variance in student achievement scores may help the guidance counselor in her selection of a preparatory program.
10. True

11.2 The One-Way ANOVA Test

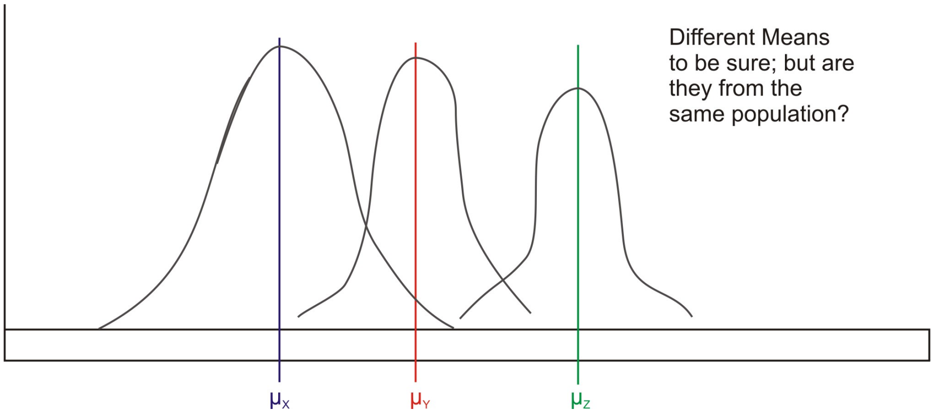
Learning Objectives

- Understand the shortcomings of comparing multiple means as pairs of hypotheses.
- Understand the steps of the ANOVA method and its advantages.
- Compare the means of three or more populations using the ANOVA method.
- Calculate the pooled standard deviation and confidence intervals as estimates of standard deviations of the populations.

Introduction

Previously, we have discussed analysis that allows us to test if the means and variances of two populations are equal. But let's say that 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 test that all of the sample means came from the same population. However, this would be tedious and has a major flaw which we will discuss later. Instead, we use something called the **An**alysis **of** **Va**riance (**ANOVA**) that allows us to test the hypothesis that **multiple** (K) population means and variance 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 the computer or TI-83/4 calculator to compute these easily and quickly. However, there is a major flaw with this reasoning.

When more than one t-test is run, each at its own level of significance ($\alpha = .10, .05, .01$, etc.) 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 tested 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 $\alpha = .05$, the probability of committing a Type I error is not .05 as one would initially expect. Instead, it would be 0.40 – extremely high!

The Steps of the ANOVA Method

In ANOVA, we are actually analyzing the **total variation** of the scores including (1) the variation of the scores within the groups and (2) 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, then we are saying that there is a small likelihood α that we would have obtained such an extreme F-statistic 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 ($\mu_1 = \mu_2 = \ldots = \mu_k$), 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 means in each group. To calculate this difference, we use the formula:

$$SS_B = \sum_{k=1}^{k} n_k(\bar{X}_k - \bar{X})^2$$

where:

$k =$ the group number

$n_k =$ the sample size in group k

$\bar{X}_k =$ the mean of group k

$\bar{X} =$ mean of all individual observations

$k =$ the number of groups

When simplified, the formula becomes:

$$SS_B = \sum_{k=1}^{k} \frac{T_k^2}{n_k} - \frac{T^2}{N}$$

where

$T_k =$ sum of the observations in group K

$T =$ sum of all observations.

Once we calculate this value, we divide by the number of degrees of freedom $(K - 1)$ to arrive at the MS_B.

$$MS_B = \frac{SS_B}{K - 1}$$

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 of 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 formula:

$$\frac{\sum(X_{i1} - \bar{X}_1)^2 + \sum(X_{i2} - \bar{X}_2)^2 + \ldots + \sum(X_{ik} - \bar{X}_k)^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1)}$$

Simplified, this formula states:

$$SS_W = \sum_{k=1}^{k} \sum_{i=1}^{n_k} X_{ik}^2 - \sum_{k=1}^{k} \frac{T_k^2}{n_k}$$

where

$T_k =$ 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 $(N - K)$.

$$MS_w = \frac{SS_w}{N - K}$$

3. Calculate the test statistic. The test statistic is as follows:

$$F = \frac{MS_B}{MS_W}$$

4. Find the critical value on the F- distribution. As mentioned above, $K - 1$ degrees of freedom are associated with MS_B and $N - K$ degrees of freedom are associated with MS_W. 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. Interpret 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 to using the ANOVA method is that it takes all types of variation into account so that we have an accurate analysis. In addition, we can use technological tools including computer programs (SAS, SPSS, Microsoft Excel) and the TI-83/4 calculator to easily conduct the calculations and test our hypothesis. We use these technological tools quite often when using the ANOVA method.

Let's take a look at an example to help clarify.

Example:

Let's go back to the example in the introduction with the teacher that 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 the five programs and 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		

Please (1) compare the means of these different groups by calculating the mean squares between groups and (2) use the standard deviations from our samples to calculate the mean squares within groups and estimate the pooled variance of a population.

Solution:

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	$= 1,452$
$\frac{\text{Sum of Obs. Squared}}{\text{Number of Obs}}$ $\left(\frac{T_k^2}{n_k}\right)$	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

$$SS_B = \sum_{k=1}^{k} \frac{T_k^2}{n_k} - \frac{T^2}{N}$$

$$\approx 1,364.57 - \frac{(194)^2}{31} \approx 150.5$$

Since there are four Degrees of Freedom for this calculation (the number of groups minus one), the mean squares between groups is

$$MS_B = \frac{SS_B}{K-1} \approx \frac{150.5}{4} \approx 37.6$$

Next we calculate the mean squares within groups (MS_W) which is also known as the estimation of the pooled variance of a population (σ^2).

To calculate the mean squares within groups, we use the formula

$$SS_W = \sum_{k=1}^{k} \sum_{i=1}^{n_k} X_{ik}^2 - \sum_{k=1}^{k} \frac{T_k^2}{n_k}$$

Using our summary statistics from above, we can calculate that the within groups mean square (MS_W) is equal to:

$$\begin{aligned} SS_W &= \sum_{k=1}^{k} \sum_{i=1}^{n_k} X_{ik}^2 - \sum_{k=1}^{k} \frac{T_k^2}{n_k} \\ &\approx 1,452 - 1,364.57 \\ &\approx 87.43 \end{aligned}$$

And so we have

$$MS_W = \frac{SS_W}{N-K} \approx \frac{87.43}{26} \approx 3.36$$

Therefore, our F-Ratio is

$$F = \frac{MS_B}{MS_W} \approx \frac{37.6}{3.36} \approx 11.18$$

We would then analyze this test statistic against our critical value (using the F-distribution table and a value of ($\alpha = .02$), we find our critical value equal to 4.14. Since our test statistic (11.18) exceeds our critical value (4.14), we reject the null hypothesis. Therefore, we can conclude that not all of the population means of the five programs are equal and that obtaining an F-ratio that extreme by chance is highly improbable.

Technology Note - Excel

Here is the procedure for performing a One-way ANOVA in Excel using this set of data.

1. Copy and paste the table into an empty Excel worksheet
2. Select Data Analysis from the Tools menu and choose "ANOVA: Single-factor" from the list that appears
3. Place the cursor is in the "Input Range" field and select the entire table.
4. Place the cursor in the "Output Range" 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.

Note: The TI-83/4 also offers a One-way ANOVA test.

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	5	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				

Lesson Summary

1. When testing multiple independent samples to determine if they come from the same populations, 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).

2. In ANOVA, we analyze the total variation of the scores including (1) the variation of the scores within the groups and (2) the variation between the group means and the total mean of all the groups (also known as the *grand mean*).
3. 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.
4. The total mean of squares within groups is also known as the estimate of the pooled variance of the population. 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 tested 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 means 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

   A school psychologist is interested 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**

Ms. Jones	Mr. Smith	Mrs. White
8	23	21
6	11	21
4	17	22
12	16	18
16	6	14

Table 11.3: (continued)

Ms. Jones	Mr. Smith	Mrs. White
17	14	21
12	15	9
10	19	11
11	10	
13		

5. State the null hypothesis.
6. Using the data above, please 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. $\left(\frac{T_k^2}{n_k}\right)$				=

7. What is the mean squares between groups (MS_B) value?
8. What is the mean squares within groups (MS_W) value?
9. What is the F-ratio of these two values?
10. Using a $\alpha = .05$, please use the F-distribution to set a critical value
11. What decision would you make regarding the null hypothesis? Why?

Review Answers

1. Analysis of Variance
2. Increase or increase exponentially
3. C
4. E, A, D, B, C
5. $H_0 : \mu_1 = \mu_2 = \mu_3$

Table 11.5:

	Ms. Jones	Mr. Smith	Mrs. White	Totals
Number (n_k)	10	9	8	$= 27$
Total (T_k)	109	131	137	$= 377$
Mean ($\bar{X}$)	10.9	14.6	17.1	$= 5,264$
Sum of Squared Obs. ($\sum_{i=1}^{n_k} X_{ik}^2$)	$1,339$	$2,113$	$2,529$	$= 5,981$
Sum of Obs. Squared/Number of Obs. $\left(\frac{T_k^2}{n_k}\right)$	$1,188$	$1,907$	$2,346$	$= 5,441$

7. 26.35
8. 4.03
9. 6.54
10. 3.40
11. The calculated test statistic exceeds the critical value so we would reject the null hypothesis. Therefore, we could conclude that not all the population means are equal.

11.3 The Two-Way ANOVA Test

Learning Objectives

- Understand the difference in situations that allow for one-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 variance 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 a physical endurance test in both males and females. The three different dosages of the medicine are (1) low, (2) medium and (3) high and the genders are (1) male and (2) female. Analyses with two independent variables, like the one just described, are called **two-way ANOVA tests**.

Table 11.6: **Mean Scores on a Physical Endurance Test for Varying Dosages and Genders**

	Dietary Supplement Dosage	Dietary Supplement Dosage	Dietary Supplement Dosage	
	Low	**Medium**	**High**	**Total**
Female	35.6	49.4	71.8	52.27
Male	55.2	92.2	110.0	85.8
Total	45.2	70.8	90.9	

There are several questions that can be answered by a study like this, for example:

- Does the medication improve physical endurance, as measured by the test?
- Do males and females respond in the same way to the medication?

While there are similar steps in performing one- and two-way ANOVA tests, there are some major differences. In the following sections we will explore the differences in situations that allow for the one- 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-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.

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.

1. **Efficiency**. With simultaneous analysis of two independent variables, the ANOVA is really carrying out two separate research studies at once.

2. **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, we are able to determine the effects of various reading programs, the effects of IQ and the possible interaction between the two.
3. **Interaction**. With two-way ANOVA 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:

1. Variation within the group (*'within-cell' variation*)
2. Variation in the dependent variable attributed to one independent variable (*variation among the row means*)
3. Variation in the dependent variable attributed to the other independent variable (*variation among the column means*)
4. Variation between the independent variables (*the interaction effect*)

The formulas that we use to calculate these types of variation 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 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, Microsoft Excel, or a calculator to compute these figures much more quickly and accurately than we can. 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/advanced/anova2.htm

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 (J) equals each other. In the example above, we would say that the mean for males equals the mean for females.

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$$

2. In the population, the means for the columns (K) equals 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 the there is no interaction between gender and amount of dosage.

$$H_0 : \text{all effects} = 0$$

Let's take a look at an example of a data set and how we can interpret the summary tables produced by technological tools to test our hypotheses.

Example:

Say that the 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 flexibility and recorded the following results. Each cell represents the score of each student:

Table 11.7:

		Length of Program	Length of Program	Length of Program
		1 Week	2 Weeks	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 11.8:

			Length of Program	Length of Program	Length of Program	
			1 *Week*	2 *Weeks*	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	23.2	27.3	34.6	28.4
		St. Dev.	4.58	2.93	7.6	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 deviations within each cell to get an idea of the variance within groups. This information is helpful, but it is necessary to calculate the test statistic to determine the effects and the interaction of the two independent variables.

Technology Note - Excel

Here is the procedure for performing a Two-way ANOVA in Excel using this set of data.

1. Copy and paste the above table into an empty Excel worksheet, *without* the labels, "Length of program" and "Gender." Or use this table:

Table 11.9:

	1 week	2 weeks	3 weeks
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

2. Select Data Analysis from the Tools menu and choose "ANOVA: Single-factor" from the list that appears

3. Place the cursor is in the "Input Range" field and select the entire table.

4. Place the cursor in the "Output Range" 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.

Note: The TI-83/4 requires a program to do a Two-way ANOVA test. See http://www.wku.edu/~david.neal/statistics/advanced/anova2.html.

Using technological tools, we can generate the following summary table:

Table 11.10:

Source	*SS*	*df*	*MS*	*F*	*Critical Value of F**
Rows (gender)	330.75	1	330.75	22.36	4.07
Columns (length)	1,065.5	2	532.75	36.02	3.22
Interaction	350.00	2	175.00	11.83	3.22
Within-cell	621.00	42	14.79		
Total	2,367.25				

* statistically significant at an $\alpha = .05$

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:

1. In the population, the mean for males differs from the mean of females.
2. In the population, the means for the three exercise programs differ.
3. For the interaction, 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 **completely 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 effect 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 the athletic director was studying the effect of various physical fitness programs on males and females, he would first categorize the randomly selected students into the homogeneous categories (males and females) before randomly assigning them to a 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

1. With two-way ANOVA we are not only able to study the effect of two independent variables but also the interaction between these variables.

2. There are several advantages to conducting a two-way ANOVA including efficiency, control of variables and the ability to study the interaction between variables.

3. 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*)
- Variation between the independent variables (*the interaction effect*)

4. It is more accurate and easier to use technological tools such as computer programs or 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 these variables.
2. We could conduct multiple t-tests between pair 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

A researcher is interested in determining the effects of different doses of a dietary supplement on a physical endurance test in both males and females. The three different doses of the medicine are (1) low, (2) medium and (3) high and the genders are (1) male and (2) 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, we generate the following summary ANOVA table

Table 11.11:

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			

$^*\alpha = .05$

4. What are the three hypotheses associated with the two-way ANOVA method?
5. What are the three null hypotheses?
6. What are the critical values for each of the three hypotheses? What do these tell us?
7. Would you reject the null hypotheses? Why or why not?
8. In your own words, describe what these results tell us about this experiment.

Review Answers

1. Interaction

2. d
3. d
4. $H_0 : \mu_M. = \mu_F.$, $H_0 : \mu_1. = \mu_2. = \mu_3.$, H_0 : all effects $= 0$
5. Answers may vary. They could include (1) $H_0 : \mu_1. = \mu_2. = \ldots = \mu_j.$, $H_0 : \mu_1. = \mu_2. = \ldots = \mu_k.$, H_0 : all effects $= 0$ or (2) written hypotheses that the means of the independent variable in the rows are equal to each other, the means of the independent variable in the rows columns are equal to each other and there is no interaction.
6. The three critical values are $4.07, 3.23$ and 3.23. These values are derived from the F-distribution. If the calculated F-statistic exceeds these values, we will reject the null hypothesis.
7. We would reject the first two null hypotheses and fail to reject the third null hypothesis.
8. We can conclude that not all means in the populations are equal with regard to gender and drug dosage. Because the F-ratio for the interaction effect (gender x drug dosage) was not statistically significant, the conclusion is that there is no difference in the performance of the male and female rats across the levels of drug dosage.

Image Sources

Chapter 12

Non-Parametric Statistics

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 a 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, the 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 populations were normally distributed and that there was homogeneity of variance within the population. 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**.

As mentioned, non-parametric tests are used when the assumptions of normality and homogeneity of variance are not met. 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 have more 'power,' they are not always applicable in statistical research. The following sections will examine situations in which we would use non-parametric methods and the advantages and disadvantages to 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? There are several *advantages of* using parametric tests (i.e., the t-test for independent samples, the correlation coefficient and the one way analysis of variance) including the fact that they are more robust and have greater **power**. Having more **power** means that they have a greater chance of rejecting the null hypothesis relative to the sample size.

However, one *disadvantage* of parametric tests is that they demand that the data meet stringent requirements such as normality and homogeneity. 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, it is also required that the standard deviations of the populations be equal as well. If either of these conditions are 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 test, its non-parametric counterpart and the purpose of the 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.** Technically, 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 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. In this example, we would match the two sets of data (pre-test and post-test), measure and record each of the observations and examine the differences between the two. Depending on the size of the sample, we would calculate either the z- or the t-test statistic.

With the sign test, we first must 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. Say that 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.

Table 12.3:

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	$-$

When performing the sign test, we use the t-distribution if the sample has less than 30 observations and we use the normal distribution if the sample has greater than 30 observations. Regardless of the distribution that we use, the formula for calculating the test statistic (either the t- or z-score) is the same.

$$t = \frac{|\#\text{ Positive Observations} - \#\text{ Negative Observations}| - 1}{\sqrt{n}}$$

This formula states that the standard score (the z or the t) is equal to the absolute value of the difference between positive differences within matched pairs and the negative differences within matched pairs minus one and divided by the square root of the number of observations. For our example above, we would have a calculated t-score of:

$$t = \frac{|2-6|-1}{\sqrt{8}} \approx 1.06$$

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 median 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- and post-tests, but nothing of significance.

$$H_0 : m = 0$$
$$H_a : m \neq 0$$

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 $\alpha = .05$ we would have critical values set at 2.37 standard scores above and below the mean. Since our standard score of 1.06 does not exceed the critical value of 2.37, we would fail to reject the null hypothesis and cannot conclude that there is a significant difference between the pre- and the post-test scores.

Using the Sign Test to Evaluate a Hypothesis about a Median of a Population

In addition to using the sign test to calculate standard scores and evaluate a hypothesis, we can also use it as a quick and dirty way to estimate the probability of obtaining a certain number of successes or positives if there was no difference between the observations in the matched data set. When we use the sign test to evaluate a hypothesis about a 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- and post-test data. Therefore, we can test these types of hypotheses using the sign test by either (1) conducting an exact test using the binomial distribution when working with small samples or (2) calculating a test statistic when working with larger samples as demonstrated in the section above.

When working with small samples, the sign test is actually the binomial test with the null hypothesis that the proportion of successes will equal 0.5. So how do these tests differ? While we use the same formula to calculate probabilities, the sign test is a specific type of test that has its own tables and formulas. These tools apply only to the case where the null hypothesis that the proportion of successes will equal 0.5 and not to the more general binomial test.

As a reminder, the formula for the binomial distribution is:

$$P(r) = \frac{N!}{r!(N-r)!}p^r(1-p)^{N-r}$$

where:

$P(r) =$ the probability of exactly r successes

$N =$ the number of observations

$p =$ the probability of success on one trial

Say that a physical education teacher is interested on 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 6 of the 8 students had increased strength after the program. But is this statistically significant? To answer this question we use the binomial formula:

$$P(r) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

Using this formula, we need to determine the probability of having either 7 or 8 successes.

$$\begin{aligned} P(7) &= \frac{8!}{7!(8-7)!} 0.5^7 (1-0.5)^{8-7} = (8)(00391) = 0.03125 \\ P(8) &= \frac{8!}{8!(8-8)!} 0.5^8 (1-0.5)^{8-8} = 0.00391 \end{aligned}$$

To determine the probability of having either 7 or 8 successes, we add the two probabilities together and get: $P(7)+P(8) = 0.03125+0.00391 = 0.0352$. This states 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 (7 out of 8) 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. As a reminder, we typically use the Chi-Square distribution to assess categorical data. However, because we use the sign test to assess the occurrence of a certain change (i.e. - a success, a 'positive,' etc.) we are not confined to using only nominal data when performing this test.

So when would using the sign test with categorical data be appropriate? 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 number 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 as whatever dichotomy we would 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 number of males and females that are accepted into four-year colleges differs 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, please calculate the z-score and use a $\alpha = .05$.

Solution:

To solve this question using the sign test, we would first establish our null and alternative hypotheses:

$$Ho : m = 0$$
$$Ha : m \neq 0$$

This null hypothesis states that the median number of males and females accepted into UC schools is equal.

Next, we use a $\alpha = .05$ to establish our critical values. Using the normal distribution chart, 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 formula:

$$z = |\# \text{ of positive obs.} - \# \text{ of negative obs.}| - 1/\sqrt{n}$$

However, instead of the number 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:

$$z = |\# \text{ of positive obs.} - \# \text{ of negative obs.}| - 1/\sqrt{n} = \frac{|134 - 66| - 1}{\sqrt{200}} \approx 4.74$$

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 dirty way to test if there is a difference between pre- 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 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 the Student's t-test. When we perform this test with small samples, the test is actually more sensitive than the 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- 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

1. We use non-parametric tests when the assumptions of normality and homogeneity of variance are not met.

2. 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 ranks test, the ranks-sum test, the Kruskal-Wallis test and the runs test.
3. The sign test examines the difference in the medians of matched data sets. When testing hypotheses using the sign test, we can either calculate the standard z-score when working with large samples or use the binomial formula when working with small samples.
4. We can also use the sign test to examine differences and evaluate hypotheses with categorical data sets.
5. 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. As review, we use nonparametric tests when analyzing data that are not normally distributed or homogeneous with respect to variance. While parametric tests are preferred since they have more 'power,' they are not always applicable in statistical research.

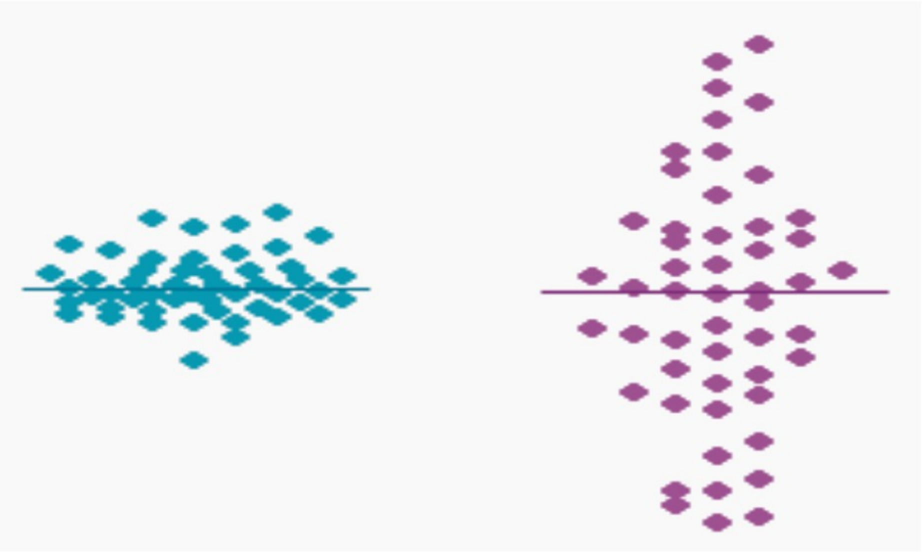
In the last section we explored two tests - the sign test and the sign rank test. We use these tests when analyzing matched data pairs or categorical data samples. In both of these tests, our null hypothesis states that there is no difference between the distributions of these variables. As mentioned, 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**) to assess whether two samples come from the same distribution. 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

As mentioned, the rank sum test tests the hypothesis that two independent samples are drawn from the same population. As a reminder, 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 very different distributions.



When performing the rank sum test, there are several different conditions that need to be met. These include:

- Although the population need not be normally distributed or have homogeneity of variance, the observations must be continuously distributed.
- That the samples drawn from the population are independent of one another.
- That the samples 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 median and the distribution of two independent samples, we establish two null hypotheses. Our null hypotheses state that the two medians and the distributions of the independent samples are equal. Symbolically, we could say that $Ho : m_1 = m_2$ and $\sigma_1 = \sigma_2$. The alternative hypotheses state that there is a difference in the median and the standard deviations of the samples.

Calculating the Mean and the Standard Deviation of Rank to Calculate a Z-Score

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 and the chi-square distributions. Similar to the t distribution, the U distribution approaches the normal distribution as the size of both samples grows. 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 score 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 = number of observations in sample 1

n_2 = number of observations in sample 2

R_1 = sum of the ranks assigned to sample 1

R_2 = 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 or 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:

$$\mu_U = \frac{n_1 n_2}{2}$$

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:

Say that we are interested in determining the attitudes on the current status of the economy from women that work outside the home and from women that do not work outside the home. We take a sample of 20 women that work outside the home (sample 1) and a sample of 20 women that do not work outside the home (sample 2) and administer a questionnaire that measures their attitude 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

Table 12.5: (continued)

Women Working Outside the Home	Women Working Outside the Home
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
$R = 408$	$R = 408$

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
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

Table 12.6: (continued)

Women Not Working Outside the Home	Women Not Working Outside the Home
$R = 412$	$R = 412$

Do these two groups of women have significantly different views on the issue?

Solution:

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 the U, the μ_U and the σ_U statistics. To calculate the U for each of the samples, we use the formulas:

$$U_1 = n_1n_2 + \frac{n_1[(n]_1+1)}{2} - R_1 = 20*20 + \frac{20(20+1)}{2} - 408 = 202$$
$$U_2 = n_1n_2 + \frac{n_2[(n]_2+1)}{2} - R_2 = 20*20 + \frac{20(20+1)}{2} - 412 = 198$$

Since we use the smaller of the two U statistics, we set $U = 198$. When calculating the other two figures, we find:

$$\mu_U = \frac{n_1n_2}{2} = \frac{20*20}{2} = 200$$

and

$$\sigma_u = \sqrt{\frac{[(n]_1(n_2)(n_1+n_2+1)}{12}} = \sqrt{\frac{(20)(20)(20+20+1)}{12}} = \sqrt{\frac{(400)(41)}{12}} = 36.97$$

When calculating the z-statistic we find,

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{198-200}{36.97} = -0.05$$

If we set the $\alpha = .05$, 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 variables 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 that the samples likely came from the same population.

Determining the Correlation between Two Variables Using the Rank Correlation Test

As we learned in Chapter 9, 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** (also known as simply the 'rank correlation' coefficient, ρ or 'rho') coefficient to measure the strength, magnitude and direction of the relationship between two variables. The test statistic from this test (ρ or 'rho') is the nonparametric alternative to the correlation coefficient and we use this test when the data do not meet the assumptions about normality. We also use the Spearman rank correlation test when one or both of the variables consist of ranks. The Spearman rank correlation coefficient is defined by the 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, etc. The lowest SAT score would get a rank of 1, the second lowest a rank of 2, etc. 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 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 the math department is interested in the correlation between scores on a final math exam and the math SAT score. She took a random sample of 15 students and recorded each students' final exam and math SAT scores. Since SAT scores are designed to be normally distributed, the Spearman rank correlation 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
405	42
680	64
490	45
565	56
580	59
615	56
435	42
440	38
515	50
380	37
510	42
565	53

Solution:

To calculate the Spearman rank correlation coefficient, we determine the ranks of each of the variables in the data set (above), calculate the difference and then calculate the squared difference for each of these ranks.

Table 12.8:

Math SAT Score (X)	Final Exam Score (Y)	X Rank	Y Rank	d	d^2
595	68	4	1	3	9

Table 12.8: (continued)

Math SAT Score (X)	Final Exam Score (Y)	X Rank	Y Rank	d	d^2
520	55	8	7	1	1
715	65	1	2	$-$	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 that:

$$\rho = 1 - 6\sum \frac{d^2}{n(n^2-1)} = 1 - \frac{6(36.50)}{15(225-1)} = 1 - 0.07 = 0.93$$

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

1. We use the rank sum test (also known as the Mann-Whitney υ 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.

2. When performing the rank sum test there are several different conditions that need to be met including that the population not be normally distributed, we have continuously distributed observations, there be an independence of samples, the samples are greater than 5 observations, and that the observations be on a numeric or ordinal scale.

3. 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.

4. To calculate the test statistic for the rank sum test, we first must calculate something known as the U statistic which 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

$$z = \frac{U - \mu_U}{\sigma_U}$$

5. We use the Spearman rank correlation coefficient (also known as simply the 'rank correlation' coefficient) to measure the strength, magnitude and direction of the relationship between two variables from non-normal distributions.

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

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 using 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 homogeneous 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 does not meet the assumptions of normality or homogeneity of variance. However, this test has its own requirements: it is essential that the data 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 distribution 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 (H) is the non-parametric alternative to the F-statistic. This test statistic is defined by the formula:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{k} \frac{R_k^2}{n_k} - 3(N+1)$$

where

$$N = \sum n_k$$

n_k = number of observations in the k^{th} sample

R_k = sum of the ranks in the kth sample

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, etc. Similar to this procedure in the other 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 $K - 1$ Degrees of Freedom where K 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 in the example below.

Example:

Suppose that the 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:

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

Please determine if there is a difference between the final exam scores of the three teachers.

Solution:

Our hypothesis for the Kruskal-Wallis test is that there is no difference in the distribution 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 $\alpha = .05$.

To test this hypothesis, we need to calculate our test statistic (H). To calculate this statistic, it is necessary to assign and sum the ranks for each of the scores in the table above:

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:

$$H = \frac{12}{N(N+1)} \sum_{k=1} \frac{R_k^2}{n_k} - 3(N+1) = \frac{12}{17 \times 18} \left(\frac{29^2}{6} + \frac{46^2}{5} + \frac{78^2}{6} \right) - 3(17+1) = 7.86$$

Using the Chi-Square distribution, we determined that with 2 Degrees of Freedom (3 samples –1), our critical value at $\alpha = .05$ is 5.991. Since our test statistic ($H = 7.86$) exceeds the critical value, we can reject the null hypothesis that stated there is no difference in the final exam scores between students from three different classrooms.

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 the grouping and the pattern of observations. For example, the sequence "$++++---+++--+++++++---$" has six 'runs.' Three of these runs are designated by the positive sign and three of the runs are designated by the negative sign.

We often use the run 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 appear 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 use the 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 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 a runs test is to compute the mean of the data and then designate each observations 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 ($R = 5$). 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 and 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:

$$z = \# \text{ of observed runs} - \mu/\sigma$$

$$\mu = \text{expected number of runs} = 1 + \frac{2n_1n_2}{n_1 + n_2}$$
$$\sigma^2 = \text{variance number of runs} = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

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 are random. He records the seating pattern of his students and records the following sequence:

$$\text{MFMMFFFFMMMFMFMMMMFFMFFMFFFF}$$

Is the seating arrangement random? Use a $\alpha = .05$.

Solution:

To answer this question, we first generate the null hypothesis that the seating arrangement is random and independent. Our alternate hypothesis states that the seating arrangement is *not* random or independent. With a $\alpha = .05$, 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:

$$R = 14$$

1. $M\Box(n\Box_{\downarrow}1) = 13$
2. $F\Box(n\Box_{\downarrow}2) = 15$

With these data, we can easily compute the test statistic:

$$\begin{aligned}
\mu &= \text{expected number of runs} = 1 + \frac{2(13)(15)}{13+15} = 1 + \frac{390}{28} = 14.9 \\
\sigma^2 &= \text{variance number of runs} = \frac{2(13)(15)(2*13*15-13-15)}{(13*15)^2(13+15-1)} = \frac{390(362)}{(152100)(27)} = .0034 \\
\sigma &= 0.05 \\
z &= \#\text{ of observed runs} - \mu/\sigma = \frac{14-14.9}{.05} = -18.0
\end{aligned}$$

Since the calculated test statistic is extremely high ($z = 18.0$) and exceeds our critical value we can reject the null hypothesis and conclude that the seating arrangement of males and females is not random.

Lesson Summary

1. The Kruskal-Wallis test is used when we are assessing the one way variance of a specific variable in non-normal distributions.

2. The test statistic for the Kruskal-Wallis test (H) is the non-parametric alternative to the F-statistic. This test statistic is defined by the formula

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{k} \frac{R_k^2}{n_k} - 3(N+1)$$

3. 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.

Image Sources