Principles of Business Statistics
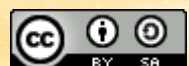
Contents

Chapter 1 Sampling and Data

1.1 Sampling and Data: Introduction

1.1.1 Student Learning Outcomes

Learning Objectives
• Recognize and differentiate between key terms. • Apply various types of sampling methods to data collection. • Create and interpret frequency tables.

1.1.2 Introduction

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

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

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

1.2 Sampling and Data: Statistics

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

1.2.1 Optional Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5; 5.5; 6; 6; 6; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:
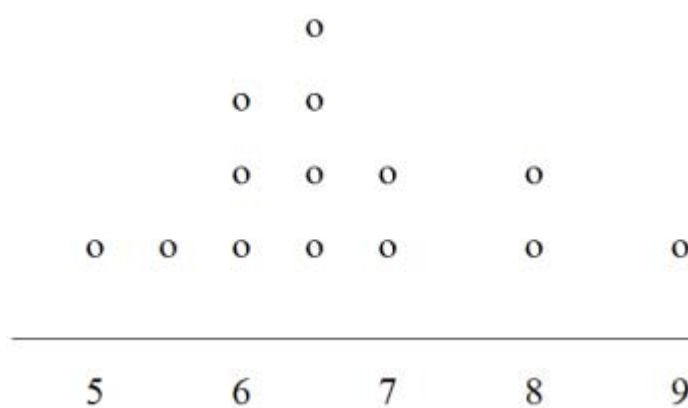
Figure 1.1 Frequency of Average Time (in Hours) Spent Sleeping per Night

Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

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

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

1.2.2 Levels of Measurement and Statistical Operations

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- Nominal scale level
- Ordinal scale level
- Interval scale level
- Ratio scale level

Data that is measured using a **nominal scale** is qualitative. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big diference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure diferences between the data.

Another example using the ordinal scale is a cruise survey where the responses to questions about the cruise are "excellent," "good," "satisfactory" and "unsatisfactory." These responses are ordered from the most desired response by the cruise lines to the least desired. But the diferences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a def inite ordering but there is a diference between data. The diferences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40 degrees is equal to 100 degrees minus 60 degrees. Diferences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations but one type of comparison cannot be done. Eighty degrees C is not 4 times as hot as 20° C (nor is 80° F 4 times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or 4 to 1).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data but, in addition, it has a 0 point and ratios can be calculated. For example, four multiple choice statistics fnal exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams were machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The diferences between the data have meaning. The score 92 is more than the score 68 by 24 points.

Ratios can be calculated. The smallest score for ratio data is 0. So 80 is 4 times 20. The score of 80 is 4 times better than the score of 20.

Exercises

What type of measure scale is being used? Nominal, Ordinal, Interval or Ratio.

1. High school men soccer players classified by their athletic ability: Superior, Average, Above average.
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box.
4. Social security numbers.
5. Incomes measured in dollars
6. A satisfaction survey of a social website by number: 1 very satisfied, 2 somewhat satisfied, 3 not satisfied.
7. Political outlook: extreme left, left-of-center, right-of-center, extreme right.
8. Time of day on an analog watch.
9. The distance in miles to the closest grocery store.
10. The dates 1066, 1492, 1644, 1947, 1944.
11. The heights of 21 65 year-old women.
12. Common letter grades A, B, C, D, F.

Answers 1. ordinal, 2. interval, 3. nominal, 4. nominal, 5. ratio, 6. ordinal, 7. nominal, 8. interval, 9. ratio, 10. interval, 11. ratio, 12. ordinal

---

1.3 Sampling and Data: Key Terms

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger

population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

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

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

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

A **variable**, notated by capital letters like **X** and **Y** , is a characteristic of interest for each person or thing in a population. Variables may be numerical or categorical. Numerical variables take on values with equal units such as weight in pounds and time in hours. Categorical variables place the person or thing into a category. If we let **X** equal the number of points earned by one math student at the end of a term, then **X** is a numerical variable. If we let **Y** be a person's party afliation, then examples of **Y** include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of **X** (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party afliation makes no sense). **Data** are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{44}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note: The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example

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

Solution

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

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

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

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

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

1.3.1 Optional Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

1.4 Sampling and Data: Data

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

- Qualitative
- Quantitative

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

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

All data that are the result of counting are called **quantitative discrete data.** These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in the numbers $\frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \pi, \frac{3\pi}{4}$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

> **Note:** In this course, the data used is mainly quantitative. It is easy to calculate statistics (like the mean or proportion) from numbers. In the chapter Descriptive Statistics, you will be introduced to stem plots, histograms and box plots all of which display quantitative data. Qualitative data is discussed at the end of this section through graphs.

Example 1.2: Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example 1.3: Data Sample of Quantitative Continuous Data

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

Example 1.4: Data Sample of Qualitative Data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

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

Example 1.5

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.

10. Weight of sumo wrestlers.
11. Amount of money won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Qualitative Data Discussion

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

	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

Table 1.1 De Anza College

	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%
Total	14,183	100%

Table 1.2 Foothill College

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

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

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

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

Look at the graphs and determine which graph (pie or bar) you think displays the comparisons better. This is a matter of preference.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.
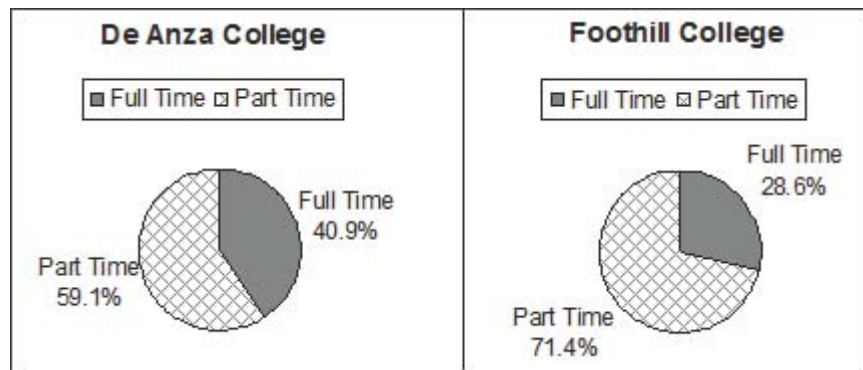
Figure 1.2 Pie Chart
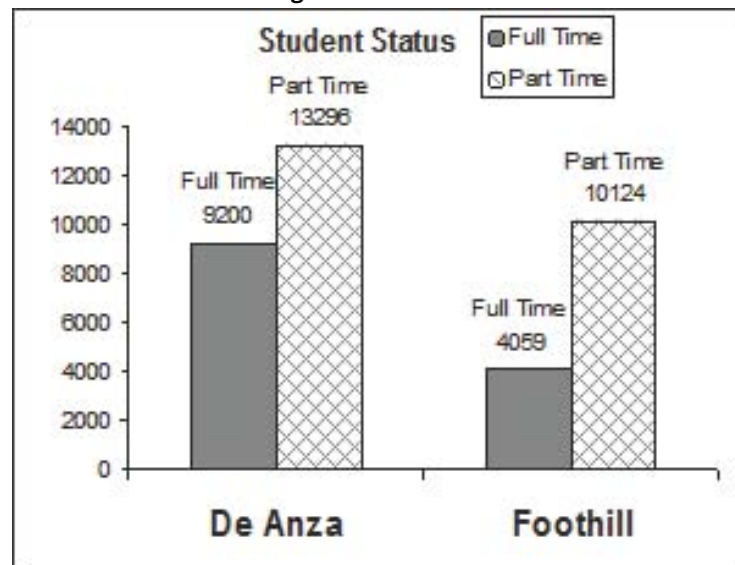
Figure 1.3 Bar Chart

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

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

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

Table 1.3 De Anza College Spring 2010
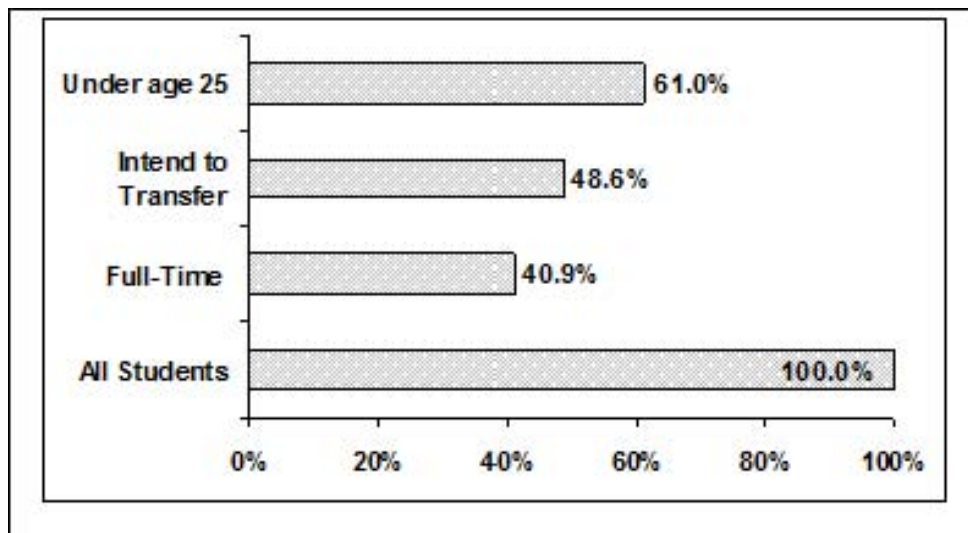
Figure 1.4 De Anza College Spring 2010

Omitting Categories/Missing Data

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

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%

Table 1.4 Missing Data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)

	Frequency	Percent
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacifc Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.4 Missing Data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)
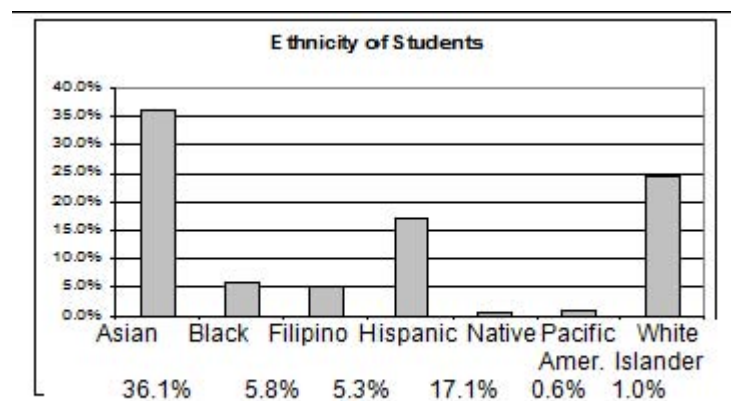
Figure 1.5 Bar graph Without Other/Unknown Category

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been added back in. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacifc Islander 1.0% particularly). This is important to know when we think about what the data are telling us.

This particular bar graph can be hard to understand visually. The graph below it is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.
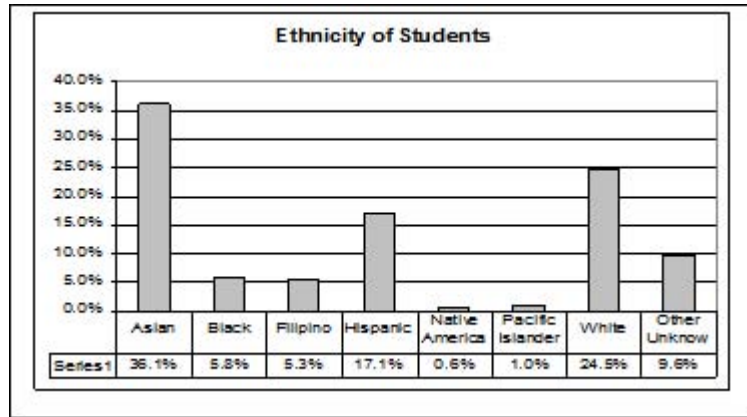
Figure 1.6 Bar Graph With Other/Unknown Category
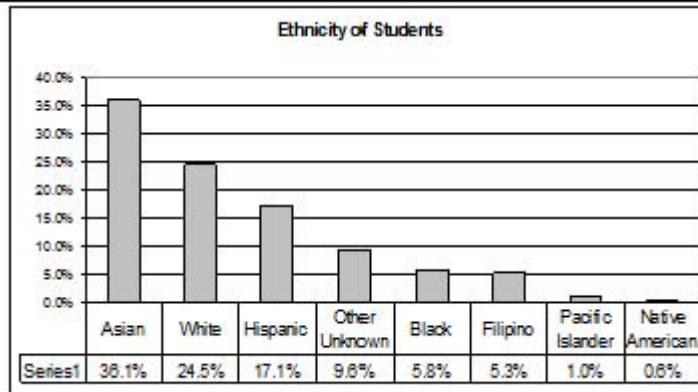
Figure 1.7 Pareto Chart With Bars Sorted By Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category added back in (since the percentages must add to 100%). The chart on the right is organized having the wedges by size and makes for a more visually informative graph than the unsorted, alphabetical graph on the left.
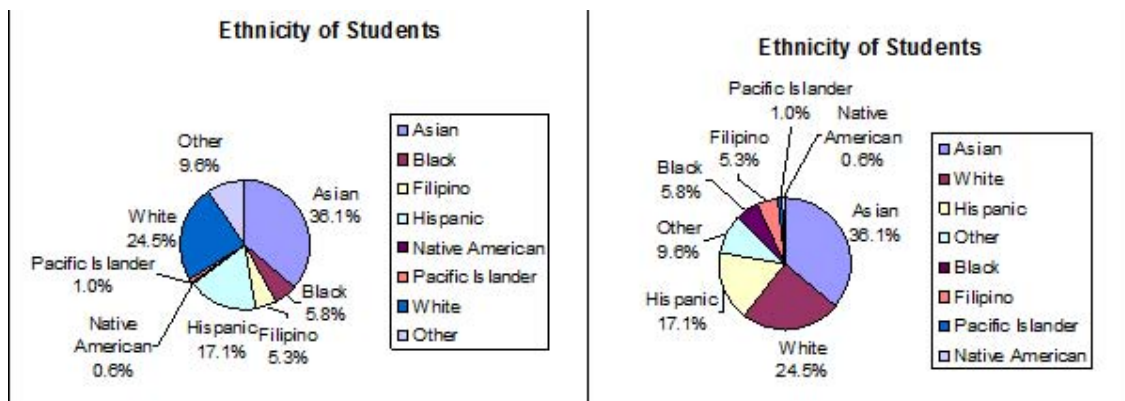
Figure 1.8 Pie Charts with "Other/Unknown" category

1.5 Sampling and Data: Variation and Critical Evaluation

1.5.1 Variation in Data

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

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

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

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

1.5.2 Variation in Samples

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

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

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

1.5.2.1 Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few

hundred observations, or even smaller, are sufcient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

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

1.5.2.2 Optional Collaborative Classroom Exercise

Exercise 1.5.1

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice**. Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fves, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.5 First Experiment (20 rolls)

Face on Die	Frequency
1	
2	

Table 1.6 Second Experiment (20 rolls)

Face on Die	Frequency
3	
4	
5	
6	

Table 1.6 Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

1.5.3 Critical Evaluation

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

- Problems with Samples: A sample should be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-Selected Samples: Responses only by people who choose to respond, such as call-in surveys are often unreliable.
- Sample Size Issues: Samples that are too small may be unreliable. Larger samples are better if possible. In some situations, small samples are unavoidable and can still be used to draw conclusions, even though larger samples are better. Examples: Crash testing cars, medical testing for rare conditions.
- Undue infuence: Collecting data or asking questions in a way that infuences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can afect the results.

- Causality: A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship through a different variable.
- Self-Funded or Self-Interest Studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading Use of Data: Improperly displayed graphs, incomplete data, lack of context.
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

1.6 Sampling and Data: Frequency Relative Frequency and Cumulative Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3

Below is a frequency table listing the different data values in ascending order and their frequencies.

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

Table 1.7 Frequency Table of Student Work Hours

A **frequency** is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3

hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the fraction or proportion of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

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

Table 1.8 Frequency Table of Student Work Hours w/ Relative Frequency

The sum of the relative frequency column is $\frac{20}{20}$ or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

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

Table 1.9 Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	2	$\frac{2}{20}$ or 0.10	0.85+0.10=0.95
7	1	$\frac{1}{20}$ or 0.05	0.95+0.05=1.00

Table 1.9 Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

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

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

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHT (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95 – 61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95 – 63.95	3	$\frac{3}{100} = 0.03$	0.05+0.03=0.08
63.95 – 65.95	15	$\frac{15}{100} = 0.15$	0.08+0.15=0.23
65.95 – 67.95	40	$\frac{40}{100} = 0.40$	0.23+0.40=0.63
67.95 – 69.95	17	$\frac{17}{100} = 0.17$	0.63+0.17=0.80
69.95 – 71.95	12	$\frac{12}{100} = 0.12$	0.80+0.12=0.92
71.95 – 73.95	7	$\frac{7}{100} = 0.07$	0.92+0.07=0.99
73.95 – 75.95	1	$\frac{1}{100} = 0.01$	0.99+0.01=1.00
	Total = 100	Total = 1.00	

Table 1.10 Frequency Table of Soccer Player Height

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

- 59.95 - 61.95 inches
- 61.95 - 63.95 inches
- 63.95 - 65.95 inches
- 65.95 - 67.95 inches
- 67.95 - 69.95 inches
- 69.95 - 71.95 inches
- 71.95 - 73.95 inches
- 73.95 - 75.95 inches

Note: This example is used again in the Descriptive Statistics (Section 2.1) chapter, where the method used to compute the intervals will be explained.

In this sample, there are **5** players whose heights are between 59.95 - 61.95 inches, **3** players whose heights fall within the interval 61.95 - 63.95 inches, **15** players whose heights fall within the interval 63.95 - 65.95 inches, 40 players whose heights fall within the interval 65.95 - 67.95 inches, **17** players whose heights fall within the interval 67.95 - 69.95 inches, **12** players whose heights fall within the interval 69.95 - 71.95, 7 players whose height falls within the interval 71.95 - 73.95, and 1 player whose height falls within the interval 73.95 - 75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example 1.6

From the table, find the percentage of heights that are less than 65.95 inches.

Solution

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

Example 1.7

From the table, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution

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

Example 1.8

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

1. The percentage of heights that are from 67.95 to 71.95 inches is:
2. The percentage of heights that are from 67.95 to 73.95 inches is:
3. The percentage of heights that are more than 65.95 inches is:
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

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

1.6.1 Optional Collaborative Classroom Exercise

Exercise 1.6.1

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class has 0 siblings?
2. What percentage of the students has from 1 to 3 siblings?
3. What percentage of the students has fewer than 3 siblings?

Example 1.9

Nineteen people were asked how many miles, to the nearest mile they commute to work each day.

The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10

The following table was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

Table 1.11 Frequency of Commuting Distances

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

Table 1.11 Frequency of Commuting Distances

Problem
1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute 3 miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute 5 or 7 miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (does not include 5 and 13 miles)?

1.6.2 Solutions to Exercises in Chapter 1

Solution to Example 1.5, Problem

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

Solution to Example 1.8, Problem

1. 29%
2. 36%

3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Solution to Example 1.9, Problem

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

1. False. Frequency for 3 miles should be 1; for 2 miles (left out), 2. Cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.
2. $\dfrac{5}{19}$
3. $\dfrac{7}{19}, \dfrac{12}{19}, \dfrac{7}{19}$

Chapter 2 Descriptive Statistics

2.1 Descriptive Statistics: Introduction

2.1.1 Student Learning Outcomes

> ## Learning Objectives
>
> By the end of this chapter, the student should be able to:
>
> - Display data graphically and interpret graphs: stemplots, histograms and boxplots.
> - Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
> - Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
> - Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

2.1.2 Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "**Descriptive Statistics**". You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

2.2 Descriptive Statistics: Displaying Data

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of

numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

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

2.3 Descriptive Statistics: Histogram

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

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

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on Sampling and Data (Section 1.1), we defined frequency as the number of times an answer occurs.) If:

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

then:

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

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$$f = 3, n = 40, and RF = \frac{f}{n} = \frac{3}{40} = 0.075$$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 - 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 - 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 - .0005 = 0.9995). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 (2 - 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example 2.1

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5;

67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

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

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

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

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.

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

Note: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

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

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.
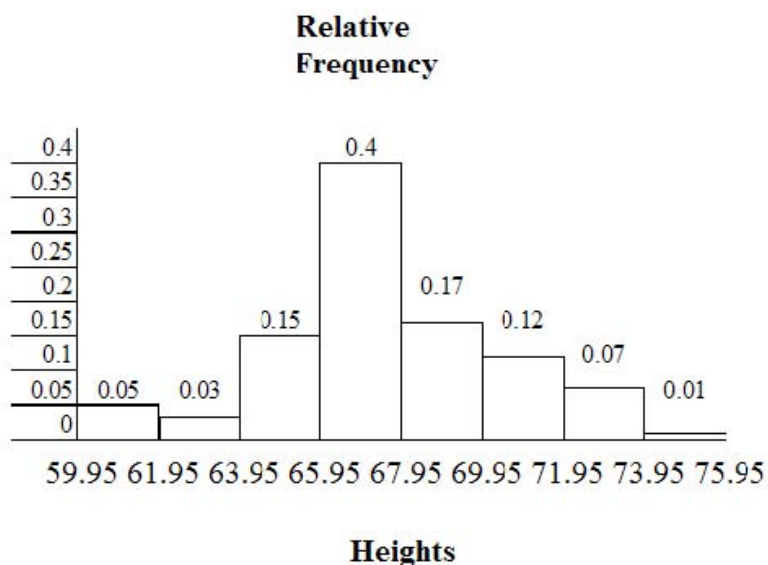
Example 2.2

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.

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

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

Calculate the number of bars as follows:
$$\frac{6.5 - 0.5}{bars} = 1$$

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

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.
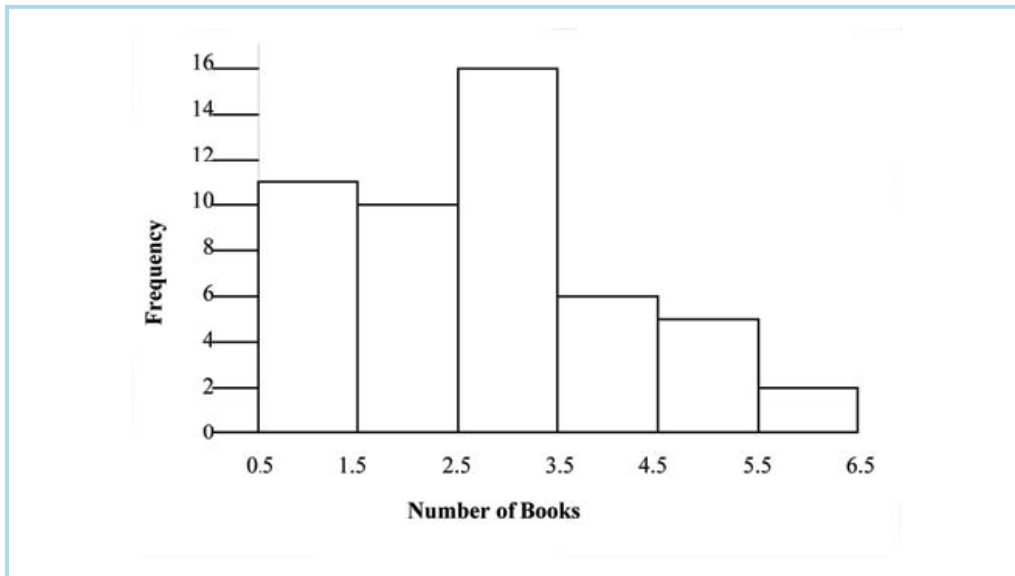
Using the TI-83, 83+, 84, 84+ Calculator Instructions

Go to the Appendix (14:Appendix) in the menu on the left. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.

- Press Y=. Press CLEAR to clear out any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6
- Into L2, enter 11, 10, 16, 6, 5, 2
- Press WINDOW. Make Xmin = .5, Xmax = 6.5, Xscl = (6.5 -.5)/6, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1
- Press 2nd Y =. Start by pressing 4:Plotsof ENTER.
- Press 2nd Y =. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH
- Use the TRACE key and the arrow keys to examine the histogram.

2.3.1 Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

2.4 Descriptive Statistics: Measuring the Center of the Data

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

> **Note:** The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced "**x** bar"): $\bar{x}$.

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

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

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

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

In the second calculation for the sample mean, the frequencies are 3, 2, 1, and 5. You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st

values. The location of the median and the value of the median are **not** the same. The upper case letter **M** is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.3

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

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

Calculate the mean and the median.

Solution

The calculation for the mean is:

$$\bar{x} = \frac{[3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + ... + 35 + 37 + 40 + (44)(2) + 47]}{40} = 23.6$$

To find the median, **M**, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

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

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

The median is 24.

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

Calculator Instructions are located in the menu item 14:Appendix (Notes for the TI-83, 83+, 84, 84+ Calculators).

- Enter data into the list editor. Press STAT 1:EDIT
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and ENTER.
- Press the down and up arrow keys to scroll.

$$\bar{x} = 23.6, M = 24$$

Example 2.4

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

Solution

$$\bar{x} = \frac{5000000 + 49 \times 30000}{50} = 129400$$

$$M = 30000$$

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

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

Another measure of the center is the mode. The mode is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example 2.5

Statistics exam scores for 20 students are as follows:

50 ; 53 ; 59 ; 59 ; 63 ; 63 ; 72 ; 72 ; 72 ; 72 ; 72 ; 76 ; 78 ; 81 ; 83 ; 84 ; 84 ; 84 ; 90 ; 93

Problem
Find the mode.

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

Example 2.6

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

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

> **Note:** The mode can be calculated for qualitative data as well as for quantitative data.
>
> Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

2.4.1 The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample is very likely to get closer and closer to **μ**. This is discussed in more detail in **The Central Limit Theorem**.

> **Note:** The formula for the mean is located in the Summary of Formulas section course.

2.4.2 Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\bar{x}$ is an example of a statistic which estimates the population mean μ.

2.5 Descriptive Statistics: Skewness and the Mean, Median, and Mode

Consider the following data set:

4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.
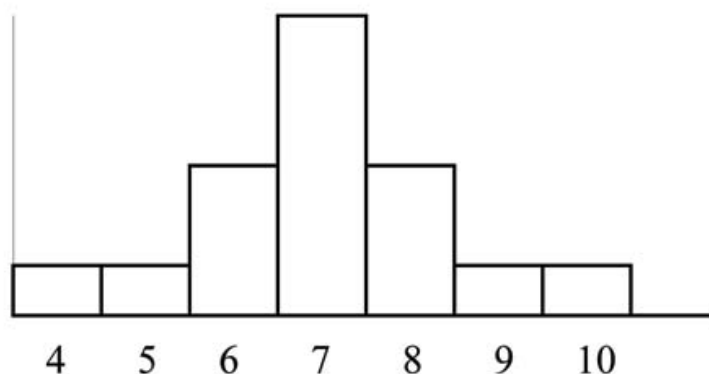
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same**. This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.
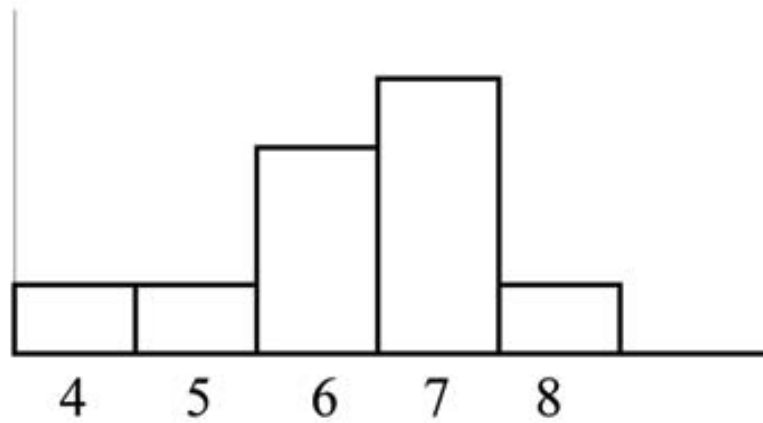
The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode**. The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10

is also not symmetrical. It is **skewed to the right.**
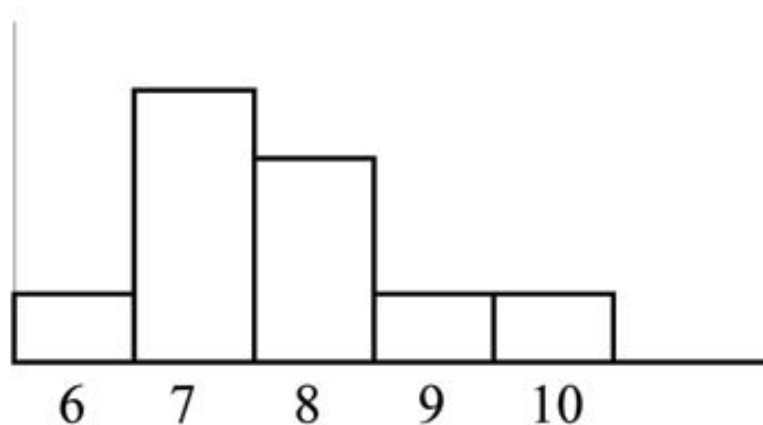
The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest.** Again, the mean refects the skewing the most.

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

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

2.6 Descriptive Statistics: Measuring the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The **standard deviation i**s a number that measures how far data values are from their mean. **The standard deviation**

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

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

The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

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

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data afects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is one standard deviation to the **right** of 5 because 5 + (1)(2) = 7.

If 1 were also part of the data set, then 1 is two standard deviations to the **left** of 5 because 5 + (−2)(2) = 1.
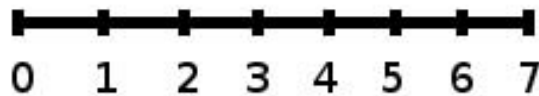
- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7 = 5 + **(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1 = 5 + **(−2)**(2)

The equation **value** = **mean** + (**#ofSTDEVs**)(**standard deviation**) can be expressed for a sample and for a population:

- **sample**: $x = \bar{x} + (\#ofSTDEV)(s)$
- **Population**: $x = \mu + (\#ofSTDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\bar{x}$ is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

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

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

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

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n-1**, one less than the number of items in the sample. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

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

  - For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

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

  - For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is 1. If a value appears three times in the data set or population, *f* is 3.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean i**s an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The

notation for the standard error of the mean is $\dfrac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note: In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example 2.7

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

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

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

The average age is 10.53 years, rounded to 2 places.

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

Data	Freq.	Deviations	Deviations²	(Freq.)(Deviations²)
X	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	9 − 10.525 = −1.525	$(-1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 − 10.525 = −1.025	$(-1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 − 10.525 = −0.525	$(-0.525)^2$ = 0.275625	4 × .275625 = 1.1025

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
10.5	4	10.5 – 10.525 = –0.025	$(-0.025)^2$ = 0.000625	4 × .000625 = .0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × .225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × .950625 = 2.851875

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

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

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

$$s = \sqrt{0.5125} = .0715891$$

Rounded to two decimal places, s =0.72

Typically, **you do the calculation for the standard deviation on your calculator or computer.** The intermediate results are not rounded. This is done for accuracy.

Problem 1

Verify the mean and standard deviation calculated above on your calculator or computer.

Solution

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

- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x} = 10.525$
- Use Sx because this is sample data (not a population): Sx = 0.715891
- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: $x = \bar{x} + (\#ofSTDEVs)(s)$
- For a population: $x = \mu + (\#ofSTDEVs)(\sigma)$
- For this example, use $x = \bar{x} + (\#ofSTDEVs)(s)$ because the data is from a sample

Problem 2

Find the value that is 1 standard deviation above the mean. Find $(\bar{x} + 1s)$.

Solution

$$(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$$

Problem 3

Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.

Solution

$$(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$$

Problem 4

Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

$$(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$$
$$(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero.** (For this example, there are n=20 deviations.) So you cannot simply add the

deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

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

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

> **Note:** Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**.

> **Note:** The formula for the standard deviation is at the end of the chapter.

Example 2.8

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

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

1. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
2. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:

1. The sample mean
2. The sample standard deviation
3. The median
4. The first quartile
5. The third quartile
6. IQR

3. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution

1.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548

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

2.
   1. The sample mean = 73.5
   2. The sample standard deviation = 17.9
   3. The median = 73
   4. The first quartile = 61
   5. The third quartile = 90
   6. IQR = 90 – 61 = 29

3. The x-axis goes from 32.5 to 100.5; y-axis goes from –2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 –32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 the ending value; No data values fall on an interval boundary.
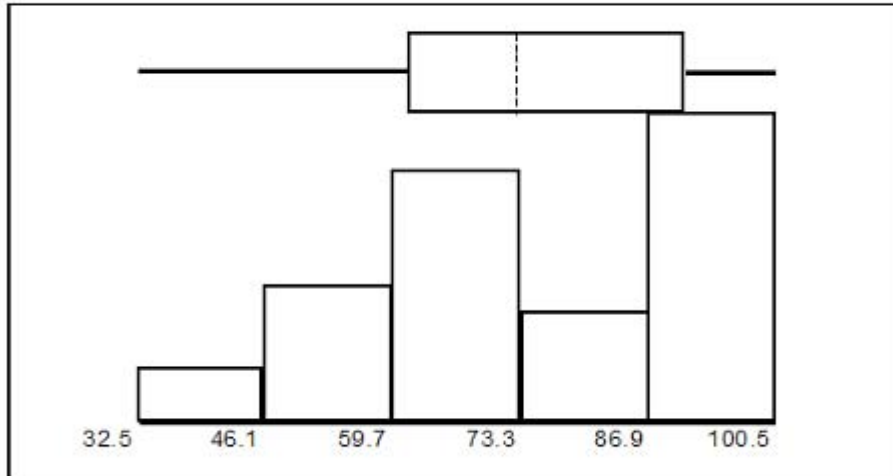
The long left whisker in the box plot is refected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 − 33 = 40) than the spread in the upper 50% (100 − 73 = 27). The histogram, box plot, and chart all refect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

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

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.

$$\#ofSTDEVs = \frac{value - mean}{standard\ deviation}$$

- Compare the results of this calculation.

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

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

Example 2.9

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

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

Solution

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

$$\#ofSTDEVs = \frac{value - mean}{standard\ deviation}; z = \frac{x - \mu}{\sigma}$$

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

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

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

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

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.

- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

2.6.1 Solutions to Exercises in Chapter 2

Solution to Example 2.2, Problem

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

Chapter 3 The Normal Distribution

3.1 Normal Distribution: Introduction

3.1.1 Student Learning Outcomes

> ## Learning Objectives
>
> By the end of this chapter, the student should be able to:
>
> - Recognize the normal probability distribution and apply it appropriately.
> - Recognize the standard normal probability distribution and apply it appropriately.
> - Compare normal probabilities by converting to the standard normal distribution.

3.1.2 Introduction

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real estate prices ft a normal distribution. The normal distribution is extremely important but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal, and applications associated with them.

3.1.3 Optional Collaborative Classroom Activity

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the x-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the

approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If **X** is a quantity to be measured that has a normal distribution with mean (μ) and the standard deviation (σ), we designate this by writing

$$NORMAL: X \sim N(\mu, \sigma)$$
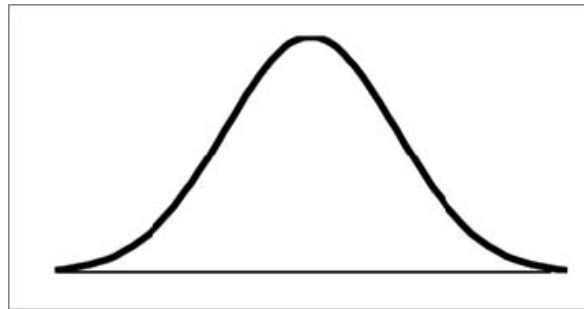
The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

$$f(x) = \frac{1}{\alpha \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot (\frac{x-\mu}{\sigma})^2}$$

The cumulative distribution function is **P (X<x)** . It is calculated either by a calculator or a computer or it is looked up in a table. Technology has made the tables basically obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions in this chapter. See the NOTE in this chapter in **Calculation of Probabilities**.

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median since the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infnite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

3.2 Normal Distribution: Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z-scores**. **A z-score is measured in units of the standard deviation**. For example, if the mean of a normal distribution is 5 and the standard deviation is 2, the value 11 is 3 standard deviations above (or to the right of) the mean. The calculation is:

$$x = \mu + (z)\sigma = 5 + (3)(2) = 11$$

The z-score is 3.

The mean for the standard normal distribution is 0 and the standard deviation is 1.

The transformation $z = \dfrac{x - \mu}{\sigma}$ produces the distribution Z~ N (0, 1) . The value *x* comes from a normal distribution with mean μ and standard deviation σ.

3.3 Normal Distribution: Z-scores

If **X** is a normally distributed random variable and X~N(μ, σ), then the z-score is:

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

The z-score tells you how many standard deviations that the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of 0.

> # Example 3.1
>
> Suppose **X ~ N (5, 6)**. This says that **X** is a normally distributed random variable with mean μ = 5 and standard deviation σ = 6. Suppose x = 17. Then:
>
> $$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$
>
> This means that x = 17 is **2 standard deviations** (2σ) above or to the right of the mean μ = 5. The standard deviation is σ = 6.
> Notice that:
>
> 5+2 · 6 = 17 (The pattern is μ + zσ = x.)
>
> Now suppose x = 1. Then:
>
> $$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$
>
> (rounded to two decimal places)
>
> **This means that x = 1 is 0.67 standard deviations (– 0.67σ) below or to the left of the mean μ = 5. Notice that:**
>
> 5+(−0.67) (6) is approximately equal to 1 (This has the pattern μ +(−0.67) σ = 1) Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ.

Example 3.2

Some doctors believe that a person can lose 5 pounds, on the average, in a month by reducing his/her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let **X** = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of 2 pounds. **X~N (5, 2)**. Fill in the blanks.

Problem 1

Suppose a person lost 10 pounds in a month. The z-score when x = 10 pounds is z = 2.5 (verify). This z-score tells you that x = 10 is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Problem 2

Suppose a person **gained** 3 pounds (a negative weight loss). Then z _____. This z-score tells you that x = -3 is _____ standard deviations to the _____ (right or left) of the mean.

Suppose the random variables X and Y have the following normal distributions: X ~ N (5, 6) and Y ~ N (2, 1). If x = 17, then z = 2. (This was previously shown.) If y = 4, what is z?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$$

where μ = 2 and σ = 1.

The z-score for y = 4 is z = 2. This means that 4 is z = 2 standard deviations to the right of the mean. Therefore, x = 17 and y = 4 are both 2 (of **their**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose **X ~N (5, 6)** represents weight gains for one group of people who are trying to gain weight in a 6 week period and **Y ~N (2, 1)** measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since **x = 17** and **y = 4** are each **2** standard deviations to the right of their means, they represent the same weight gain **relative to their means**.

The Empirical Rule

If **X** is a random variable and has a normal distribution with mean μ and standard deviation σ then the **Empirical Rule** says (See the Figure 3.1 below)

- About 68.27% of the x values lie between -1σ and +1σ of the mean μ (within 1 standard deviation of the mean).

- About 95.45% of the x values lie between -2σ and +2σ of the mean μ (within 2 standard deviations of the mean).
- About 99.73% of the x values lie between -3σ and +3σ of the mean μ (within 3 standard deviations of the mean). Notice that almost all the x values lie within 3 standard deviations of the mean.
- The z-scores for +1σ and 1σ are +1 and -1, respectively.
- The z-scores for +2σ and 2σ are +2 and -2, respectively.
- The z-scores for +3σ and 3σ are +3 and -3 respectively.
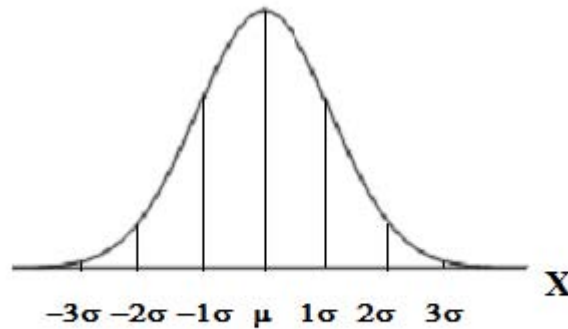
Figure 3.1 The Empirical Rule

The Empirical Rule is also known as the 68-95-99.7 Rule.

Example 3.3

Suppose X has a normal distribution with mean 50 and standard deviation 6.

- About 68.27% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within 1 standard deviation of the mean 50. The z-scores are -1 and $+1$ for 44 and 56, respectively.
- About 95.45% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$ of the mean 50. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within 2 standard deviations of the mean 50. The z-scores are -2 and 2 for 38 and 62, respectively.
- About 99.73% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within 3 standard deviations of the mean 50. The z-scores are -3 and $+3$ for 32 and 68, respectively.

3.4 Normal Distribution: Areas to the Left and Right of x

The arrow in the graph below points to the area to the left of x. This area is represented by the probability P (X<x). Normal tables, computers, and calculators provide or calculate the probability P (X<x).
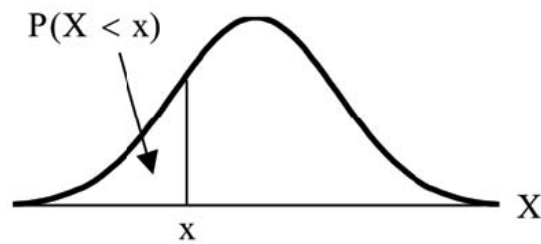
The area to the right is then P (X>x) = 1 − P (X<x).

Remember, P (X<x)= **Area to the left** of the vertical line through x.

P (X>x) = 1 − P (X<x) = **Area to the right** of the vertical line through x.

P (X<x) is the same as P (X ≤ x) and P (X>x) is the same as P (X ≥ x) for continuous distributions.

3.5 Normal Distribution: Calculations of Probabilities

Probabilities are calculated by using technology. There are instructions in the chapter for the TI-83+ and TI-84 calculators.

Note: In the Table of Contents for Collaborative Statistics, entry 15. Tables has a link to a table of normal probabilities. Use the probability tables if so desired, instead of a calculator. The tables include instructions for how to use then.

Example 3.4

If the area to the left is 0.0228, then the area to the right is 1 - 0.0228 = 0.9772.

Example 3.5

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

Please see Problem 1-4.

Problem 1

Find the probability that a randomly selected student scored more than 65 on the exam.

Solution
Let X = a score on the final exam. $X \sim N$ (63, 5), where μ = 63 and σ = 5
Draw a graph.
Then, find P (x> 65).
P (x> 65) = 0.3446 (calculator or computer)
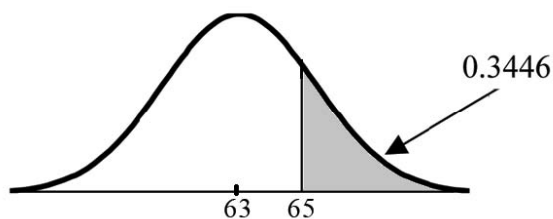
The probability that one student scores more than 65 is 0.3446. Using the TI-83+ or the TI-84 calculators, the calculation is as follows. Go into 2nd DISTR. After pressing 2nd DISTR, press 2:normalcdf. The syntax for the instructions are shown below. normalcdf(lower value, upper value, mean, standard deviation) For this problem: normal cdf(65,1E99,63,5) = 0.3446. You get 1E99 (10^{99}) by pressing 1, the **EE** key (a 2nd key) and then 99. Or, you can enter 10-99 instead. The number 1099 is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99} . In some instances, the lower number of the area might be –1E99 (–10^{99}). The number –10^{99} is way out in the left tail of the normal curve.

Historical Note: The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65 - 63}{5} = 0.4$$

Area to the left is 0.6554. P (x> 65) = P (z> 0.4) = 1 – 0.6554 = 0.3446

Problem 2

Find the probability that a randomly selected student scored less than 85.

Solution
Draw a graph. Then find P (x< 85). Shade the graph. P (x< 85) = 1 (calculator or computer) The probability that one student scores less than 85 is approximately 1 (or 100%). The TI-instructions and answer are as follows: normalcdf(0,85,63,5) = 1 (rounds to 1)

Problem 3

Find the 90th percentile (that is, find the score k that has 90 % of the scores below k and 10% of the scores above k).

Solution

Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. k is located on the x-axis. P (x<k) is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k and 10% are the same or higher. k is often called a **critical value.**

k = 69.4 (calculator or computer)
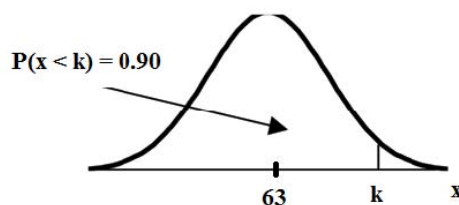
The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. For the TI-83+ or TI-84 calculators, use invNorm in 2nd DISTR. invNorm(area to the left, mean, standard deviation) For this problem, invNorm(0.90,63,5) = 69.4

Problem 4

Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution

Find the 70th percentile.
Draw a new graph and label it appropriately. k = 65.6
The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.
invNorm(0.70,63,5) = 65.6

Example 3.6

A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Please see problem 1-2.

Problem 1

Find the probability that a household personal computer is used between 1.8 and 2.75 hours per day.

Solution
Let X = the amount of time (in hours) a household personal computer is used for entertainment. $x \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.
Find $P(1.8 < x < 2.75)$.
The probability for which you are looking is the area **between** x = 1.8 and x = 2.75. $P(1.8 < x < 2.75) = 0.5886$
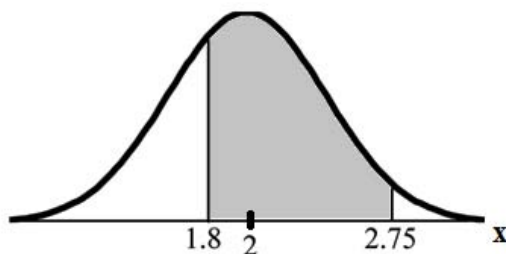
normalcdf(1.8,2.75,2,0.5) = 0.5886
The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Problem 2

Find the maximum number of hours per day that the bottom quartile of households use a personal computer for entertainment.

Solution
To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile**, k, where $P(x < k) = 0.25$.
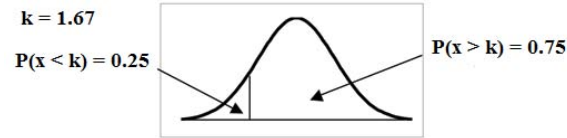
invNorm(0.25,2,.5) = 1.66
The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

3.6 Central Limit Theorem: Central Limit Theorem for Sample Means

Suppose **X** is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

1. μ_X = the mean of X
2. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\bar{X}$ which consists of sample means, tends to be **normally distributed** and

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

The Central Limit Theorem for Sample Means says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by **n**, the sample size. n is the number of values that are averaged together not the number of times the experiment is done.

To put it more formally, if you draw random samples of size **n**, the distribution of the random variable $\bar{X}$, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as **n**, the sample size, increases.

The random variable $\bar{X}$ has a different z-score associated with it than the random variable **X**. $\bar{x}$ is the value of $\bar{X}$ in one sample.

$$z = \frac{\bar{x} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

μ_X is both the average of X and of $\bar{X}$.

$\sigma_{\bar{X}} = \dfrac{\sigma_X}{\sqrt{n}}$ = standard deviation of $\bar{X}$ and is called the **standard error of the mean.**

Example 3.7

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population.

See Problem 1-2.

Problem 1

Find the probability that the **sample mean** is between 85 and 92.

Solution
Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean.**

Let $\bar{X}$ = the mean of a sample of size 25. Since μX = 90, σX = 15, and n = 25;
then

$$\bar{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$$

Find P (85 <$\bar{x}$< 92) Draw a graph.
$$P(85 < \bar{x} < 92) = 0.6997$$

The probability that the sample mean is between 85 and 92 is 0.6997.
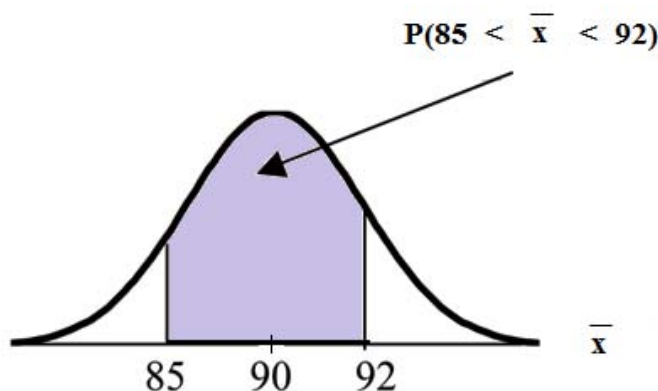
TI-83 or 84: `normalcdf` (lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\dfrac{a}{\sqrt{n}}$)

$$normalcdf\left(85, 92, 90, \frac{15}{\sqrt{25}}\right) = 0.6997$$

Problem 2

Find the value that is 2 standard deviations above the expected value (it is 90) of the sample mean.

Solution

To find the value that is 2 standard deviations above the expected value 90, use the formula

$$value = \mu_X + (\#ofSTDEVs)\left(\frac{\sigma_X}{\sqrt{n}}\right)$$

$$value = 90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$$

So, the value that is 2 standard deviations above the expected value is 96.

Example 3.8

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2** hours and a **standard deviation of 0.5 hours**. A **sample of size n = 50** is drawn randomly from the population.

See Problem below.

Problem

Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution

Let X = the time, in hours, it takes to play one soccer match.
The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.
Let $\bar{X}$ the **mean** time , in hours, it takes to play one soccer match.
If μ_X = _____ , σ_X = _____ , and n = _____ , then
$\bar{X} \sim N$ (_____ , _____) by the Central Limit Theorem for Means.

μ_X = 2, σ_X = 0.5, n = 50, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \bar{x} < 2.3)$. Draw a graph.
$$P(1.8 < \bar{x} < 2.3) = 0.9977$$

$$normalcdf\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$$

> The probability that the mean time is between 1.8 hours and 2.3 hours is _____.

3.7 Central Limit Theorem: Using the Central Limit Theorem

It is important for you to understand when to use the **CLT**. If you are being asked to find the probability of the mean, use the CLT for the mean. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for means and sums.

> **Note:** If you are being asked to find the probability of an individual value, do not use the CLT. **Use the distribution of its random variable.**

3.7.1 Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean $\bar{x}$ of the sample tends to get closer and closer to μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean and Sum Examples

> # Example 3.9
>
> A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:
>
> 1. The probability that the **mean stress score** for the 75 students is less than 2.
> 2. The 90th percentile for the **mean stress score** for the 75 students.
> 3. The probability that the **total of the 75 stress scores** is less than 200.
> 4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.
Problems 1 and 2 ask you to find a probability or a percentile for a **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.
Since the individual stress scores follow a uniform distribution, X ~ U (1, 5) where a = 1 and b = 5 (See Continuous Random Variables for the uniform).

$$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$$

For problems 1. and 2., let $\bar{X}$ = the mean stress score for the 75 students. Then,

$$\bar{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$$

where n = 75.

See Problem 1-4.

Problem 1

Find $P\left(\bar{x} < 2\right)$. Draw the graph.
Solution

$$P\left(\bar{X} < 2\right) = 0$$

The probability that the mean stress score is less than 2 is about 0.

$$P\left(\bar{x} < 2\right)$$
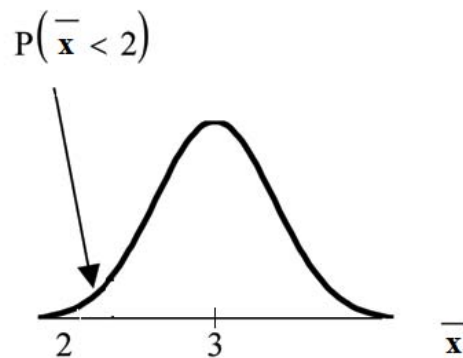
$$normalcdf\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$$

REMINDER: The smallest stress score is 1. Therefore, the smallest mean for 75 stress scores is 1.

Problem 2

Find the 90th percentile for the mean of 75 stress scores. Draw a graph.
Solution

Let k = the 90th percentile.
Find k where

$$P\left(\bar{x} < k\right) = 0.90.$$
$$k = 3.2$$

$$P\left(\overline{x} < k\right) = 0.90$$
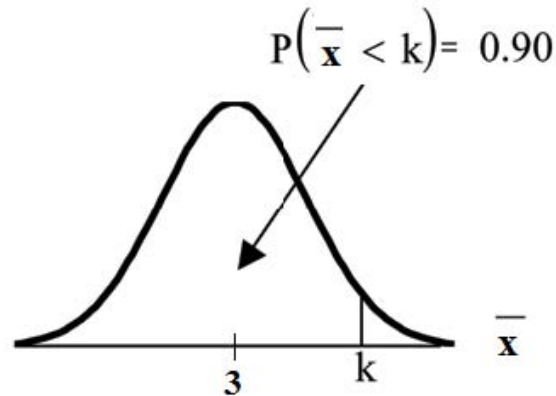
The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2 and 10% are at least 3.2.

$$invNorm\left(.90, 3, \sqrt{75}\right) = 3.2$$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then,

$$\sum X \sim N[(75) \cdot (3), \sqrt{75} \cdot 1.15]$$

Problem 3

Find P (Σx < 200). Draw the graph.

Solution

The mean of the sum of 75 stress scores is 75 · 3 = 225
The standard deviation of the sum of 75 stress scores is

$$\sqrt{75} \cdot 1.15 = 9.96$$

P (Σx < 200) = 0

$$P\left(\sum x < 200\right)$$
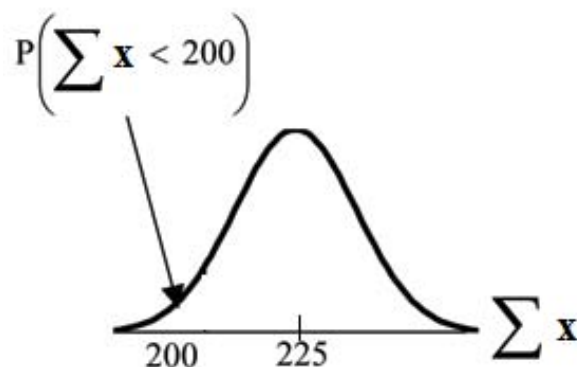
The probability that the total of 75 scores is less than 200 is about 0.

$$normalcdf\left(75, 200, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 0$$

REMINDER: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Problem 4

Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution

Let k = the 90th percentile.

Find k where P (Σx <k) = 0.90.

k = 237.8

$$P\left(\sum x < k\right) = 0.90.$$
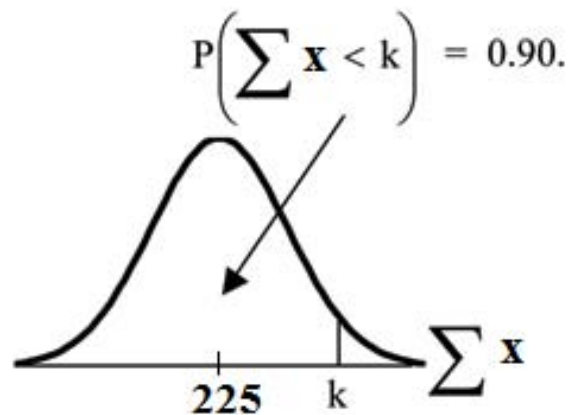
The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

$$invNorm(.90, 75 \cdot 3, \sqrt{75} \cdot 1.15) = 237.8$$

Example 3.10

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the excess time used follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one **INDIVIDUAL** cell phone customer who exceeds his contracted time allowance.

$$X \sim Exp\left(\frac{1}{22}\right)$$

From Chapter 5, we know that μ = 22 and σ = 22.

Let $\bar{X}$ = the mean excess time used by a sample of n = 80 customers who exceed their contracted time allowance.

$$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$$

by the CLT for Sample Means.

See Problems 5–6.

Problem 5

Using the CLT to find Probability:

1. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$ Draw the graph.
2. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find P (x > 20)
3. Explain why the probabilities in (a) and (b) are different.

Solution

Part a.

Find: $P(\bar{x} > 20)$

$$P(\bar{x} > 20) = 0.7919$$

using

$$normalcdf\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.

$P\left(\overline{X} > 20\right)$

20 22
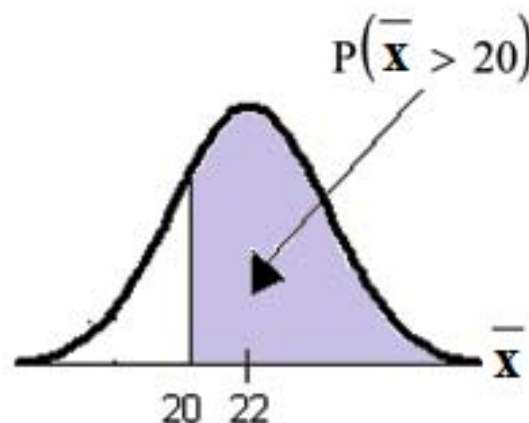
REMINDER: 1E99 = 10^{99} and−1E99 = -10^{99} . Press the EE key for E. Or just use 10^99 instead of 1E99.

Part b.

Find P(x>20) . Remember to use the exponential distribution for an individual: **X~Exp(1/22)**.

$$P(x > 20) = e^{\wedge}\left((1/22)^{*} 20\right) \quad or \quad e^{\wedge}(.04545^{*}20) = 0.4029$$

Part c. Explain why the probabilities in (a) and (b) are different.

$$P(x > 20) = 0.4029 \quad but \quad P(\bar{x} > 20) = 0.7919$$

The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.

When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the CLT. Use the CLT with the normal distribution when you are being asked to find the probability for an mean.

Problem 6

Using the CLT to find Percentiles:

Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution

Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$

k = 26.0 using $invNorm\left(.95, 22, \dfrac{22}{\sqrt{80}}\right) = 26.0$

$$P(\overline{x} < k) = 0.95$$
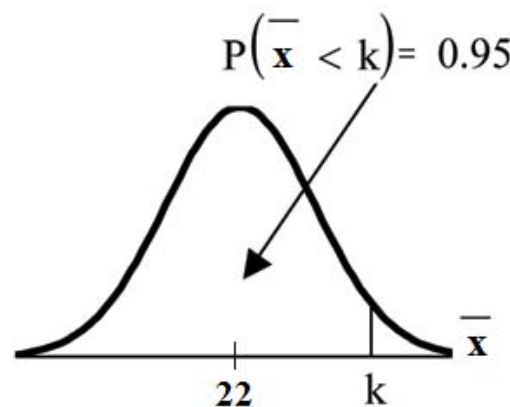
The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time. 95% of such samples would have means under 26 minutes; only 5% of such samples would have means above 26 minutes.

Note: (HISTORICAL): Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for n (say, 20). To calculate the probabilities with large values of n, you had to use the binomial formula which could be very complicated. Using the **Normal Approximation to the Binomial** simplifed the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials

- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then X~B (n, p). The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five (np > 5 and nq > 5; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean μ = np and standard deviation $\sigma = \sqrt{npq}$. Remember that q = 1 – p. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use x + 0.5 or x – 0.5). The number 0.5 is called the **continuity correction factor**.

Example 3.11

Suppose in a local Kindergarten through 12th grade (K – 12) school district, 53 percent of the population favor a charter school for grades K – 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let X = the number that favor a charter school for grades K – 5. X~B (n, p) where n = 300 and p = 0.53. Since np > 5 and nq > 5, use the normal approximation to the binomial. The formulas for the mean and standard deviation are μ = np and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is **Y .Y ~ N (159, 8.6447)**. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so P (x ≥ 150) has normal approximation P (Y ≥ 149.5) = 0.8641.
$$normalcdf\,(149.5, 1099, 159, 8.6447) = 0.8641$$
For Problem 2., you **include 160** so P (x ≤ 160) has normal approximation P (Y ≤ 160.5) = 0.5689.
$$normalcdf\,(0, 160.5, 159, 8.6447) = 0.5689$$
For Problem 3., you **exclude 155** so P (x > 155) has normal approximation P (y > 155.5) = 0.6572.
$$normalcdf\,(155.5, 1099, 159, 8.6447) = 0.6572$$
For Problem 4., you **exclude 147** so P (x < 147) has normal approximation P (Y < 146.5) = 0.0741.
$$normalcdf\,(0, 146.5, 159, 8.6447) = 0.0741$$
For Problem 5., P (x = 175) has normal approximation P (174.5 < y < 175.5) = 0.0083.
$$normalcdf\,(174.5, 175.5, 159, 8.6447) = 0.0083$$

Because of calculators and computer software that easily let you calculate binomial probabilities for large values of n, it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial. In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.

For **Example 3**, the probabilities are calculated using the binomial (n = 300 and p = 0.53) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.

$P (x \geq 150)$: 1 − binomialcdf (300, 0.53, 149) = 0.8641
$P (x \leq 160)$: binomialcdf (300, 0.53, 160) = 0.5684
$P (x > 155)$: 1 − binomialcdf (300, 0.53, 155) = 0.6576
$P (x < 147)$: binomialcdf (300, 0.53, 146) = 0.0742
$P (x = 175)$: (You use the binomial pdf.) binomialpdf (175, 0.53, 146) = 0.0083

**Contributions made to Example 2 by Roberta Bloom

3.7.2 Solutions to Exercises in Chapter 3

Solution to Example 3.2, Problem 1

This z-score tells you that x = 10 is **2.5** standard deviations to the **right** of the mean 5.

Solution to Example 3.2, Problem 2

z = **-4**. This z-score tells you that x= −3 is 4 standard deviations to the **left** of the mean.

Chapter 4 Confidence Interval

4.1 Confidence Intervals: Introduction

4.1.1 Student Learning Outcomes

> ## Learning Objectives
>
> By the end of this chapter, the student should be able to:
>
> - Calculate and interpret confidence intervals for one population mean and one population proportion.
> - Interpret the student-t probability distribution as the sample size changes.
> - Discriminate between problems applying the normal and the student-t distributions.

4.1.2 Introduction

Suppose you are trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percent of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics. The sample data help us to make an estimate of a population parameter.** We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct confidence intervals in which we believe the parameter lies.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of compact discs (CD's) a consumer buys per month. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the

sample standard deviation, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, *s*, is the point estimate for the population standard deviation, *σ*.

Each of $\bar{x}$ and s is also called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the CD example we do not know the population mean μ but we do know that the population standard deviation is σ =1 and our sample size is 100. Then by the Central Limit Theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. For our CD example, two standard deviations is (2) (0.1) = 0.2. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations ((2) (0.1)) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the CD example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \ and \ \bar{x} + 0.2 = 2 + 0.2 = 2.2$$

We say that we are **95% confident** that the unknown population mean number of CDs is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (100% - 95%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form

(point estimate - margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate + or - the margin of error. These are two ways of expressing the same concept.

Note: Although the text only covers symmetric confidence intervals, there are non-symmetric confidence intervals (for example, a confidence interval for the standard deviation).

4.1.3 Optional Collaborative Classroom Activity

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be 3 meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval

$$\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}} \right)$$

We say we are approximately 95% confident that the true average number of meals that students eat out in a week is between _____ and_____.

4.2 Confidence Intervals: Confidence Interval, Single Population Mean, Population Standard Deviation Known , Normal

4.2.1 Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**, we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error is called the **error bound for a population mean** (abbreviated **EBM**). The sample mean $\bar{x}$ is the **point estimate** of the unknown population mean μ

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or, in symbols,

$$(\bar{x} - EBM, \ \bar{x} + EBM)$$

The margin of error depends on the **confidence level** (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a

confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level CL. α is the probability that the interval does not contain the unknown population parameter. Mathematically, α + CL = 1.

Example 4.1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population. The sample mean is 7 and the error bound for the mean is 2.5.

$$\bar{x} = 7 \quad and \quad EBM = 2.5$$

The confidence interval is (7 − 2.5, 7 + 2.5); calculating the values gives (4.5, 9.5).

If the confidence level (CL) is 95%, then we say that "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5." A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of x = 10 and we have constructed the 90% confidence interval (5, 15) where EBM = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the normal distribution.

Confidence Level (CL) = 0.90

$\bar{x} = 10$

$EBM = 5$

$\bar{x} - EBM = 5$

$\bar{x} + EBM = 15$
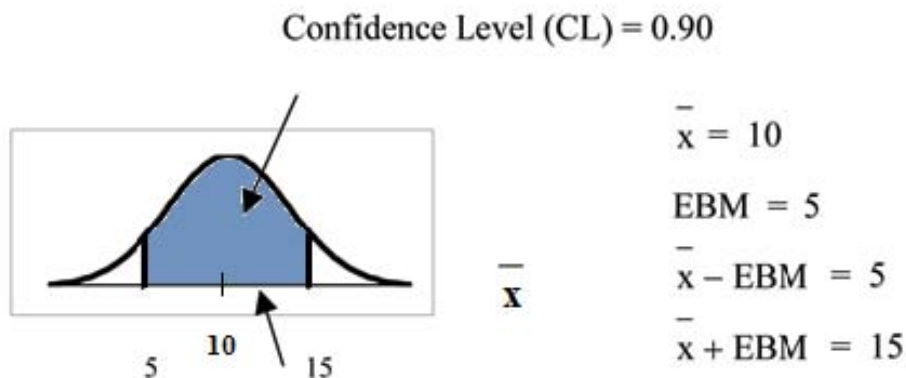
μ is believed to be in the interval (5, 15) with 90% confidence.

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. 1.645 is the z-score from a Standard Normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating. So in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. $\frac{\sigma}{\sqrt{n}}$ is commonly called the "standard error of the mean" in order to clearly distinguish the standard deviation for a mean from the population standard deviation σ.

In summary, as a result of the Central Limit Theorem:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N\left(\mu_X \cdot \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a Normal distribution to calculate the error bound.**

Calculating the Confidence Interval:

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section, we already know the population standard deviation σ.
- Find the Z-score that corresponds to the confidence level.
- Calculate the error bound EBM
- Construct the confidence interval
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding z for the stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution Z ~ N (0,1).

The confidence level, CL, is the area in the middle of the standard normal distribution. CL = 1 − α. So α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$

For example, when CL = 0.95 then α = 0.05 and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{.025}$

The area to the right of z.$_{025}$ is 0.025 and the area to the left of z.$_{025}$ is 1-0.025 = 0.975

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a Standard Normal probability table.

Using the TI83, TI83+ or TI84+ calculator: invNorm(0.975, 0, 1) = 1.96

CALCULATOR NOTE: Remember to use area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs 2 in the invNorm command are 0,1 because you are using a Standard Normal Distribution Z~N(0,1)

EBM: Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$EMB = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(\bar{x} - EBM, \bar{x} + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$$
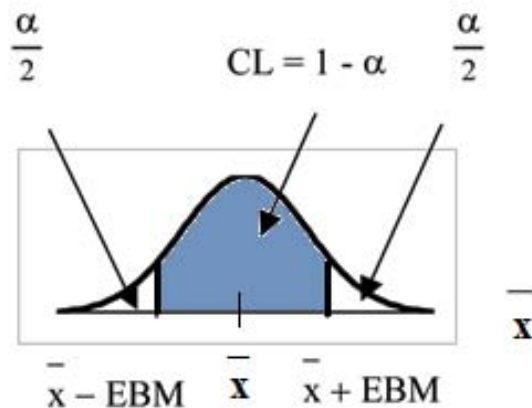
Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and should state the confidence interval (both endpoints). "We estimate with _____ % confidence that the true population mean (include context of the problem) is between _____ and (include appropriate units)."

Example 4.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Problem

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution

- You can use technology to directly calculate the confidence interval
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+ and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$$

σ =3 ; n = 36 ; The confidence level is 90% (CL = 0.90)
CL = 0.90 so α = 1 – CL = 1 – 0.90 = 0.10

$$\frac{\alpha}{2} = 0.05$$

$$z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 1–0.05 =0.95

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using invNorm(0.95,0,1) on the TI-83,83+,84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.

$$EBM = 1.645 \cdot \left(\frac{3}{\sqrt{36}} \right) = 0.8225$$
$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is (67.1775, 68.8225).

Solution B
Using a function of the TI-83, TI-83+ or TI-84 calculators:
Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.

Arrow to Stats and press ENTER.
Arrow down and enter 3 for σ, 68 for $\bar{x}$, 36 for n, and .90 for C-level.
Arrow down to Calculate and press ENTER.
The confidence interval is (to 3 decimal places) (67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

4.2.2 Changing the Confidence Level or Sample Size

Example 4.3: Changing the Confidence Level

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution
To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$$\bar{x} = 68$$

$$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{a}{\sqrt{n}} \right)$$

σ =3 ; n = 36 ; The confidence level is 95% (CL = 0.95)
CL = 0.95 so α = 1 − CL = 1 − 0.95 = 0.05

$$\frac{\alpha}{2} = 0.025$$

$$z_{\frac{\alpha}{2}} = z_{.025}$$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 1−0.025 = 0.975

$$z_{\frac{\alpha}{2}} = z_{.025} = 1.96$$

using invnorm(.975,0,1) on the TI−83,83+,84+ calculators. (This can

also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.)

$$EMB = 1.96 \cdot \left(\frac{3}{\sqrt{36}} \right) = 0.98$$
$$\bar{x} - EBM = 68 - 0.98 = 67.02$$
$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

95% of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider.
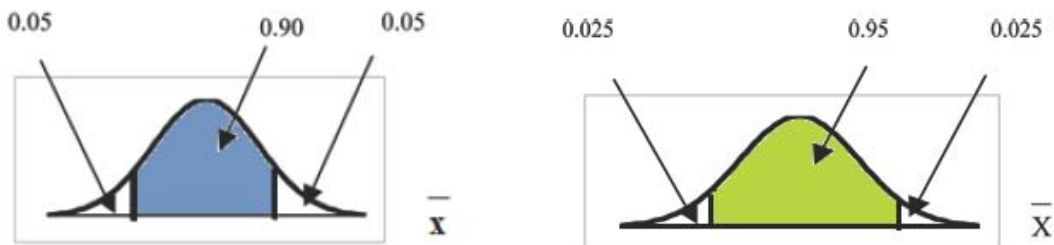
Figure 4.1 Comparing the results

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Example 4.4: Changing the Sample Size:

Suppose we change the original problem to see what happens to the error bound if the sample size is changed.

See the following Problem.

Problem

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use n=100 instead of n=36? What happens if we decrease the sample size to n=25 instead of n=36?

- $\bar{x} = 68$
- $EBM = z_{\frac{\alpha}{2}} \cdot \left(\dfrac{\sigma}{\sqrt{n}} \right)$
- σ = 3 ; The confidence level is 90% (CL = 0.90) ; $z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.
When n = 100 :

$$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = 1.645 \cdot \left(\frac{3}{\sqrt{100}} \right) = 0.4935$$

Solution B

If we **decrease** the sample size n to 25, we **increase** the error bound.
When n = 25 :

$$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = 1.645 \cdot \left(\frac{3}{\sqrt{25}} \right) = 0.987$$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

4.2.3 Working Backwards to Find the Error Bound or Sample Mean

Working Backwards to find the Error Bound or the Sample Mean

When we calculate a confidence interval, we find the sample mean and calculate the error bound and use them to calculate the confidence interval. But sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean
- OR, From the upper value for the interval, subtract the lower value. Then divide the diference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval
- OR, Average the upper and lower endpoints of the confidence interval

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 4.5

Suppose we know that a confidence interval is (67.18, 68.82) and we want to find the error bound.

We may know that the sample mean is 68. Or perhaps our source only gave the confidence interval and did not tell us the value of the the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: EBM = 68.82 – 68 = 0.82
- If we don't know the sample mean:

$$EBM = \frac{(68.82 - 67.18)}{2} = 0.82$$

Calculate the Sample Mean:

- If we know the error bound: $\bar{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\bar{x} = \frac{67.18 + 68.82}{2} = 68$

4.2.4 Calculating the Sample Size n

If researchers desire a specifc margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is

$$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$$

The formula for sample size is $n = \frac{z^2 \sigma^2}{EBM^2}$, found by solving the error bound formula for n

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who 2 wants a specifed confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 4.6

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within 2 years of the true population mean age of Foothill College students , how many randomly selected Foothill College students must be surveyed?

From the problem, we know that σ = 15 and EBM = 2
z = z$_{.025}$ = 1.96, becuase the confidence level is 95%.

$$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.96^2 15^2}{2^2} = 216.09$$

using the sample size equation.

Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within 2 years of the true population mean age of Foothill College students.

**With contributions from Roberta Bloom

4.3 Confidence Intervals: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's-t

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gossett (1876-1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This

problem led him to "discover" what is called the **Student's-t distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid 1970s, some statisticians used the **normal distribution** approximation for large sample sizes and only used the Student's-t distribution for sample sizes of at most 30. With the common use of graphing calculators and computers, the practice is to use the Student's-t distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

, then the t-scores follow a **Student's-t distribution with n − 1 degrees of freedom**. The t-score has the same interpretation as the z-score. It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's-t distribution.

The **degrees of freedom**, n − 1, come from the calculation of the sample standard deviation s. In Chapter 2, we used n deviations ($x - \bar{x}$ **values**) to calculate s. Because the sum of the deviations is 0, we can find the last deviation once we know the other n − 1 deviations. The other **n − 1** deviations can change or vary freely. **We call the number n − 1 the degrees of freedom (df).**

Properties of the Student's-t Distribution

- The graph for the Student's-t distribution is similar to the Standard Normal curve.
- The mean for the Student's-t distribution is 0 and the distribution is symmetric about 0.
- The Student's-t distribution has more probability in its tails than the Standard Normal distribution because the spread of the t distribution is greater than the spread of the Standard Normal. So the graph of the Student's-t distribution will be thicker in the tails and shorter in the center than the graph of the Standard Normal distribution.
- The exact shape of the Student's-t distribution depends on the "degrees of freedom". As the degrees of freedom increases, the graph Student's-t distribution becomes more like the graph of the Standard Normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed but it is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's-t probabilities. The TI-83,83+,84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specifed.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's-t distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's-t distribution.) When using t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's-t table gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's-t probabilities.**

The notation for the Student's-t distribution is (using T as the random variable) is

- $T \sim t_{df}$ where df = n − 1.
- For example, if we have a sample of size n = 20 items, then we calculate the degrees of freedom as df = n−1 = 20−1 = 19 and we write the distribution as $T \sim t_{19}$

If the population standard deviation is not known, the error bound for a population mean is:

$$EBM = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right)$$

- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$
- use df = n − 1 degrees of freedom
- s = sample standard deviation

The format for the confidence interval is:

$$(\bar{x} - EBM, \bar{x} + EBM)$$

The TI-83, 83+ and 84 calculators have a function that calculates the confidence interval directly. To get to it,

Press STAT
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 4.7

Suppose you do a study of acupuncture to determine how Effective it is in relieving pain. You measure sensory rates for 15 subjects with

the results given below. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+ and 84+ calculators. 8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses the Ti-83+ and Ti-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the EBM.

$\bar{x}$ = 8.2267   s = 1.6722   n = 15

df = 15 – 1 = 14

CL = 0.95 so α = 1 – CL = 1 – 0.95 = 0.05

$\dfrac{\alpha}{2} = 0.025$ $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of t$_{.025}$ is 0.025 and the area to the left of t$_{.025}$ is 1–0.025 = 0.975

$$t_{\frac{\alpha}{2}} = t_{.025} = 2.14$$

using invT(.975,14) on the TI-84+ calculator.

$$EBM = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right)$$

$$EBM = 2.14 \cdot \left(\frac{1.6722}{\sqrt{15}} \right) = 0.924$$

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is **(7.30, 9.15)**.

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press STAT and arrow over to TESTS.

Arrow down to 8:TInterval and press ENTER (or you can just press 8).

Arrow to Data and press ENTER.

Arrow down to List and enter the list name where you put the data.

Arrow down to Freq and enter 1.
Arrow down to C-level and enter .95
Arrow down to Calculate and press ENTER.
The 95% confidence interval is (7.3006, 9.1527)

Note: When calculating the error bound, a probability table for the Student's-t distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

**With contributions from Roberta Bloom

4.4 Confidence Intervals: Confidence Interval for a Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within 3 percentage points. Often, election polls are calculated with 95% confidence. So, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 : (0.40 − 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean. The formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying **distribution is binomial**. (There is no mention of a mean or average.) If **X** is a binomial random variable, then **X** ~ **B (n, p)** where n the number of trials and **p** the probability of a success. To form a proportion, take **X**, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable **P'** (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\hat{P}$, read "P hat".)

When n is large and p is not close to 0 or 1, we can use the **normal distribution** to approximate the binomial.

$$X \sim N\left(n \cdot p, \sqrt{n \cdot p \cdot q}\right)$$

If we divide the random variable by n, the mean by n, and the standard deviation by n, we get a normal distribution of proportions with P', called the estimated proportion, as the random variable. (Recall that a proportion = the number of successes divided by n.)

$$\frac{X}{n} = P' \sim N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right)$$

Using algebra to simplify :

$$\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$$

P' follows a normal distribution for proportions:

$$P' \sim N\left(P, \sqrt{\frac{p \cdot q}{n}}\right)$$

The confidence interval has the form (p' − **EBP**, p' + **EBP**).

$$p' = \frac{x}{n}$$

p' = the **estimated proportion** of successes (p' is a **point estimate** for p, the true proportion)

x = the **number** of successes.

n = the size of the sample

The error bound for a proportion is

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}}$$

where q' = 1 - p'

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$.

For a proportion, the appropriate standard deviation is $\sqrt{\frac{p \cdot q}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p' \cdot q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{p \cdot q}{n}}$.

However, in the error bound formula, the standard deviation is $\sqrt{\frac{p' \cdot q'}{n}}$.

In the error bound formula, the **sample proportions p' and q' are estimates of the unknown population proportions p and q**. The estimated proportions p' and q' are used because p and q are not known. p' and q' are calculated from the data. p' is the estimated proportion of successes. q' is the estimated proportion of failures.

The confidence interval can only be used if the number of successes np' and the number of failures nq' are both larger than 5.

Note: For the normal distribution of proportions, the z-score formula is as follows.

If

$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

then the z-score formula is

$$z = \frac{p' - p}{\sqrt{\frac{p \cdot q}{n}}}$$

Example 4.8

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adults residents of this city who have cell phones.

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution A
Let X = the number of people in the sample who have cell phones. X is binomial. $X \sim B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find p', q', and EBP.
n = 500   x = the number of successes = 421

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

p' =0.842 is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Since CL = 0.95, then α = 1 – CL = 1 – 0.95 = 0.05

$$\frac{a}{2} = 0.025.$$

Then

$$z_{\frac{a}{2}} = z_{.025} = 1.96$$

Use the TI-83, 83+ or 84+ calculator command invNorm(0.975,0,1) to find $z_{.025}$. Remember that the area to the right of $z_{.025}$ is 0.025 and the area to the left of $x_{.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = z_{\frac{a}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.96 \cdot \sqrt{\frac{(0.842) \cdot (0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is

$$(p' - EBP, p' + EBP) = (0.810, 0.874)$$

Interpretation
We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level
95% of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution B
Using a function of the TI-83, 83+ or 84 calculators:
Press STAT and arrow over to TESTS .
Arrow down toA :1-PropZint . Press ENTER .
Arrow down to x and enter 421.
Arrow down to n and enter 500.
Arrow down to C-Level and enter .95.
Arrow down to Calculate and press ENTER .
The confidence interval is (0.81003, 0.87397).

Example 4.9

For a class project, a political science student at a large university wants to estimate the percent of students that are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students that are registered voters and interpret the confidence interval.

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution A
x = 300 and n = 500.

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

$$Since \ CL = 0.90, \ then \ \alpha = 1 - CL = 1 - 0.90 = 0.10 \qquad \frac{\alpha}{2} = 0.05.$$

$$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$$

Use the TI-83, 83+ or 84+ calculator command invNorm(0.95,0,1) to find $z_{.05}$. Remember that the area to the right of $z_{.05}$ is 0.05 and the area to the left of z.05 is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.645 \cdot \sqrt{\frac{(0.60) \cdot (0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is (p' – EBP, p' + EBP) = (0.564, 0.636).

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level
90% of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered

voters.

Solution B
Using a function of the TI-83, 83+ or 84 calculators:
Press STAT and arrow over to TESTS .
Arrow down to A:1-PropZint . Press ENTER .
Arrow down to x and enter 300.
Arrow down to n and enter 500.
Arrow down to C-Level and enter .90.
Arrow down to Calculate and press ENTER .
The confidence interval is (0.564, 0.636).

4.4.1 Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. The error bound formula for a population proportion is

- $$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p'q'}{n}}$$
- Solving for n gives you an equation for the sample size.
- $$n = \frac{z_{\frac{\alpha}{2}}^2 \cdot p'q'}{EBP^2}$$

Example 4.10

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ that use text messaging on their cell phone. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ that use text messaging on their cell phone.

Solution
From the problem, we know that **EBP=0.03** (3% 0.03) and
$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$ because the confidence level is 90%

However, in order to find n , we need to know the estimated (sample) proportion p'. Remember that q' = 1-p'. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because p'q' = (.5)(.5) = .25 results in the largest possible product. (Try other products: (.6)(.4) = .24; (.3)(.7) = .21; (.2)(.8) = .16 and so on). The largest possible product gives us the largest n. This gives us

a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2} \quad gives \quad n = \frac{1.645^2 \cdot (.5)(.5)}{.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50+ that use text messaging on their cell phone.

**With contributions from Roberta Bloom.

Chapter 5 Hypothesis Testing

5.1 Hypothesis Testing of Single Mean and Single Proportion: Introduction

5.1.1 Student Learning Outcomes

> ## Learning Objectives
>
> By the end of this chapter, the student should be able to:
>
> - Diferentiate between Type I and Type II Errors
> - Describe hypothesis testing in general and in practice
> - Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
> - Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
> - Conduct and interpret hypothesis tests for a single population proportion.

5.1.2 Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing**." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufcient evidence based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or fail to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note: To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See the Table of Contents topic "Solution Sheets".

5.2 Hypothesis Testing of Single Mean and Single Proportion: Null and Alternate Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the alternate hypothesis. These hypotheses contain opposing viewpoints.

H_0: **The null hypothesis**: It is a statement about the population that will be assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt.

H_a: **The alternate hypothesis**: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Example 5.1

H_0: No more than 30% of the registered voters in Santa Clara County voted in the primary election.

H_a: More than 30% of the registered voters in Santa Clara County voted in the primary election.

Example 5.2

We want to test whether the mean grade point average in American colleges is different from 2.0 (out of 4.0).

$$H_0 : \mu = 2.0 \qquad H_\alpha : \mu \neq 2.0$$

Example 5.3

We want to test if college students take less than five years to graduate from college, on the average.

$$Ho : \mu \geqslant 5 \qquad H_\alpha : \mu < 5$$

Example 5.4

In an issue of **U. S. News and World Report**, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U. S. students take advanced placement exams and 4.4 % pass. Test if the percentage of U. S. students who take advanced placement exams is more than 6.6%.

$$Ho : p = 0.066 \qquad H_\alpha : p > 0.066$$

Since the null and alternate hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a decision. There are two options for a decision. They are "reject **H_o**" if the sample information favors the alternate hypothesis or "do not reject Ho" or "fail to reject **H_o**" if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in **H_o** and **H_a**:

H_o	H_a
equal (=)	not equal ($\neq$) **or** greater than (>) **or** less than (<)
greater than or equal to ($\geq$)	less than (<)
less than or equal to ($\leq$)	more than (>)

Note: H_o always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use = in the Null Hypothesis, even with > or < as the

5.2.1 Optional Collaborative Classroom Activity

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write a null and alternate hypotheses. Discuss your hypotheses with the rest of the class.

5.3 Hypothesis Testing of Single Mean and Single Proportion: Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual **probability of getting the test result, called the p-value. The p-value is the probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large p-value calculated from the data indicates that we should fail to reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 5.5: (to illustrate the p-value)

Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. and the distribution of heights is normal.

The null hypothesis could be H_o: $\mu \leq 15$
The alternate hypothesis is H_a: $\mu > 15$

The words **"is more than"** translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ is **known** (σ =0.5 cm.), the distribution for the population is known to be normal with mean μ= 15 and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16.$$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means from Chapter 7.
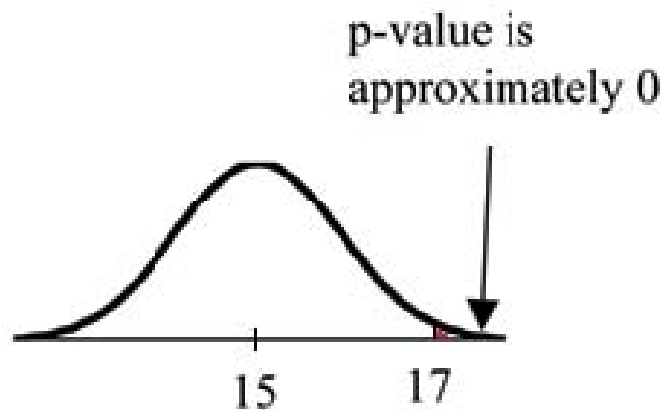
p-value = P ($\bar{x}$> 17) which is approximately 0.

A p-value of approximately 0 tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on the average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone),** we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufcient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

5.4 Hypothesis Testing of Single Mean and Single Proportion: Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the **p-value** and a **preset or preconceived α (also called a "significance level")**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_o, do as follows:

- If α > p-value, reject **H_o**. The results of the sample data are significant. There is sufficient evidence to conclude that **H_o** is an incorrect belief and that the **alternative hypothesis**, **H_a**, may be correct.
- If α ≤ p-value, do not reject **H_o**. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, **H_a**, may be correct.
- When you "do not reject **H_o**", it does not mean that you should believe that **H_o** is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of **H_o**.

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Chapter 6 Linear Regression and Correlation

6.1 Linear Regression and Correlation: Introduction

6.1.1 Student Learning Outcomes

> ## Learning Objectives
>
> By the end of this chapter, the student should be able to:
>
> - Discuss basic ideas of linear regression and correlation.
> - Create and interpret a line of best fit.
> - Calculate and interpret the correlation coefficient.
> - Calculate and interpret outliers.

6.1.2 Introduction

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is it and how strong is the relationship?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee. These are all examples in which regression can be used.

The type of data described in the examples is **bivariate** data -"bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

6.2 Linear Regression and Correlation: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:

$$y = a + bx$$

where a and b are constant numbers.

x **is the independent variable, and** *y* **is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 6.1

The following examples are linear equations.

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

The graph of a linear equation of the form y = a + bx is a **straight line**. Any line that is not vertical can be described by this equation.

Example 6.2

:
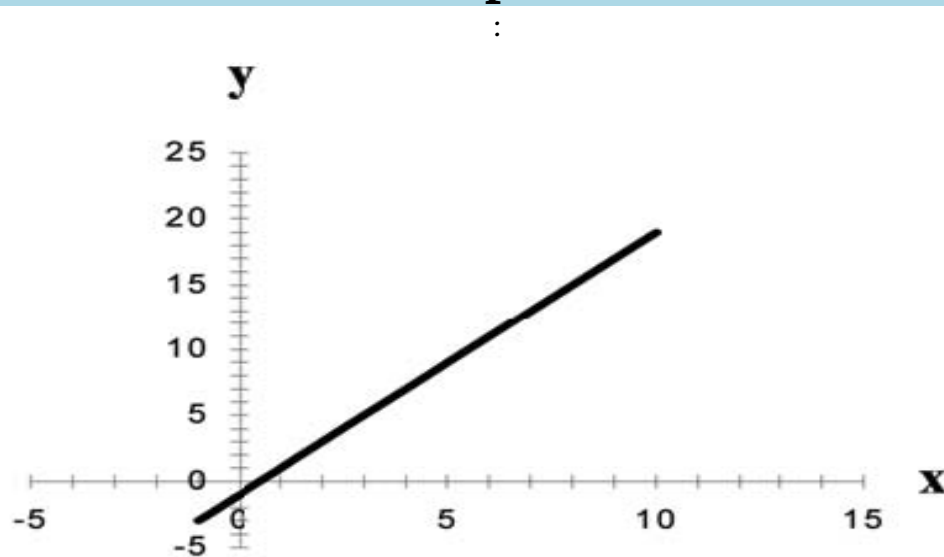
Figure 6.1 Graph of the equation y = −1+ 2x.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example 6.3

Aaron's Word Processing Service (AWPS) does word processing. Its rate is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.

See the following Problem.

Problem

Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

Solution

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is:

$$y = 31.50 + 32x$$

6.3 Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point (0,a) where the line crosses the y-axis.

(a)   (b)   (c)

Figure 6.2 Three possible graphs of y = a + bx. (a) If b> 0, the line slopes upward to the right. (b) If b =0, the line is horizontal. (c) If b< 0, the line slopes downward to the right.

Example 6.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is y = 25+ 15x.

See the following Problem.

Problem

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution
The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

6.4 Linear Regression and Correlation: Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables *x* and *y*. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example 6.5

From an article in the Wall Street Journal: In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

x (year)	y (# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Table 6.1 Table showing the number of m-commerce users (in millions) by year.
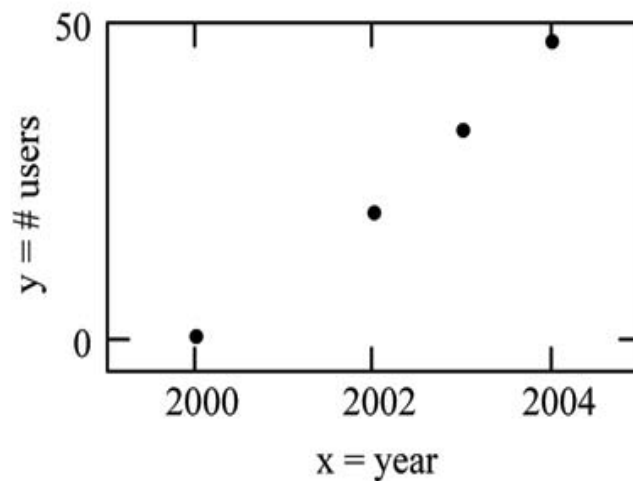
Figure 6.3 Scatter plot showing the number of m-commerce users (in millions) by year.

A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

Figure 6.4 (a) Positive Linear Pattern (Strong)   (b) Linear Pattern w/ One Deviation
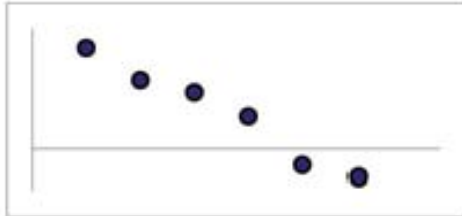
Figure 6.5 (a) Negative Linear Pattern (Strong)   (b) Negative Linear Pattern (Weak)
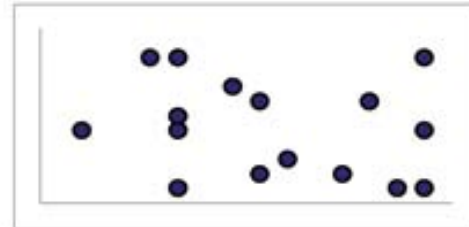
Figure 6.6 (a) Exponential Growth Pattern   (b) No Pattern

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If *x* is the independent variable and *y* the dependent variable, then we can use a regression line to predict *y* for a given value of *x*.

6.5 Linear Regression and Correlation: The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line.**

6.5.1 Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and use a ruler. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 6.6

A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

x (third exam score)	y (fnal exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163

Table 6.2 Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (fnal exam score)
71	159
69	151
69	159

Table 6.2 Table showing the scores on the final exam based on scores from the third exam.
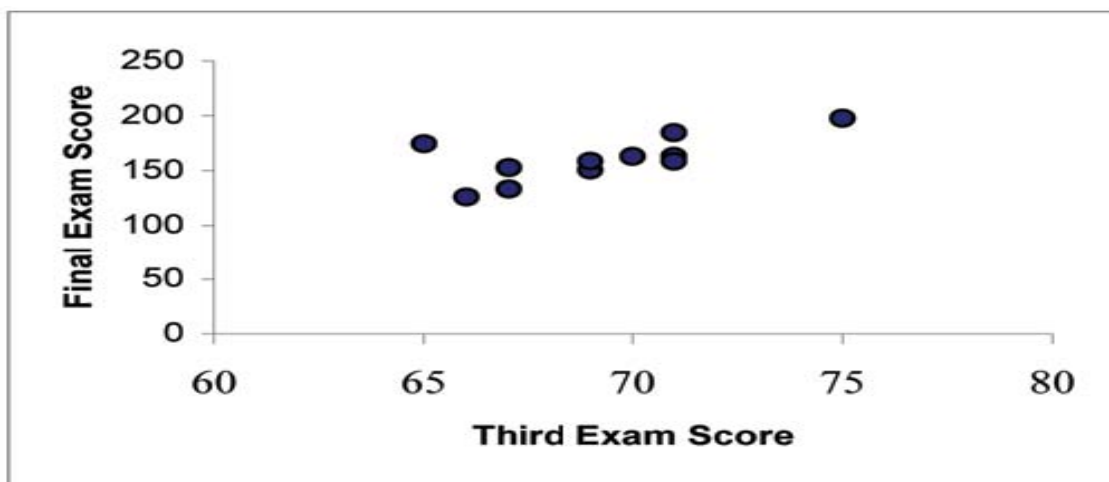
Figure 6.7 Scatter plot showing the scores on the fnal exam based on scores from the third exam.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line. Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least-squares linear regression has the form (x, $\hat{y}$).

The $\hat{y}$ is read **"y hat"** and is the **estimated value of y**. It is the value of y obtained using the regression line. It is not generally equal to *y* from data.
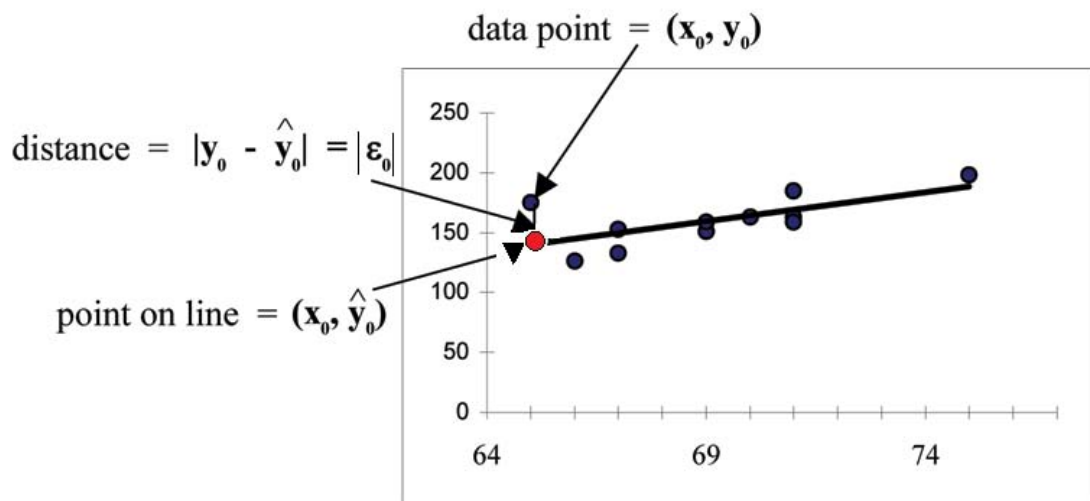
Figure 6.8

The term

$$y_0 - \hat{y}_0 = \epsilon_0$$

is called the **"error" or residual.** It is not an error in the sense of a mistake.
The **absolute value of a residual** measures the vertical distance between the actual value of *y* and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for *y*.

In the diagram above, $y_0 - \hat{y}_0 = \epsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ϵ = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors,

$$y_i - \hat{y}_i = \epsilon_i \ \ for \ \ i = 1, 2, 3, ..., 11.$$

Each $|\epsilon|$ is a vertical distance.

For the example about the third exam scores and the fnal exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ϵ values. If you square each ϵ and add, you get

$$(\epsilon_1)^2 + (\epsilon_2)^2 + ... + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$$

This is called the **Sum of Squared Errors (SSE).**

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the SSE a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx$$

where

$$a = \bar{y} - b \cdot \bar{x}$$

and

$$b = \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{\sum (x - \bar{x})^2}$$

$\bar{x}$ and $\bar{y}$ are the sample means of the x values and the y values, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope *b* can be written as

$$b = r \cdot \left(\frac{s_y}{s_x} \right)$$

where s_y = the standard deviation of the *y* values and s_x the standard deviation of the *x* values. *r* is the correlation coefficient which is discussed in the next section.

Least Squares Criteria for Best Fit

The process of fitting the best fit line is called linear regression. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line**.

> **Note:** Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

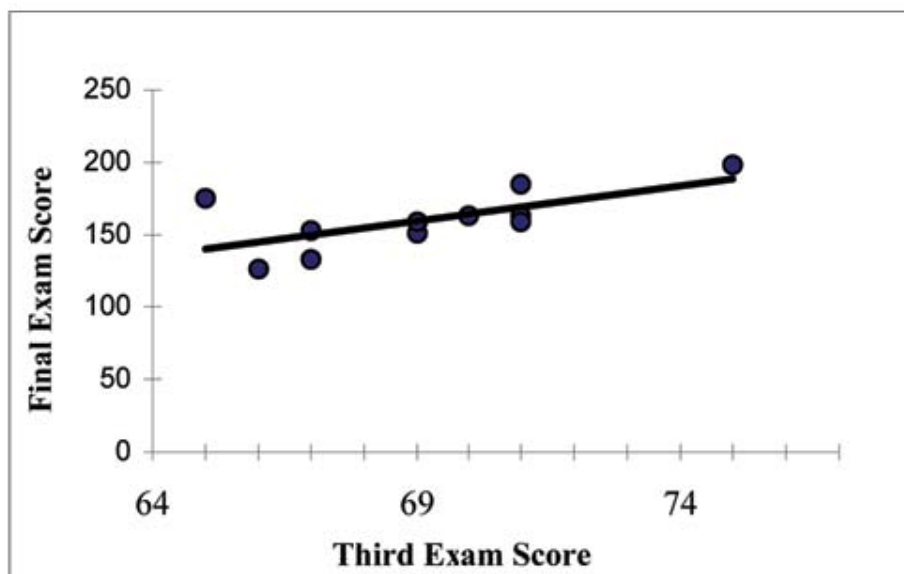The graph of the line of best fit for the third exam/final exam example is shown below:

Figure 6.9 Exam Score

The least squares regression line (best fit line) for the third exam/final exam example has the equation:

$$\hat{y} = -173.51 + 4.83x$$

Note: Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x–values in the sample data, but not necessarily for x–values outside that domain.

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, but not **necessarily for x-values outside that domain.**

You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.

You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

Slope: The slope of the line is b 4.83.
Interpretation: For a one point increase in the score on the third exam, the fnal exam score increases by 4.83 points, on average.

6.5.2 Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

Step 1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)

Step 2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)

Step 3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1

Step 4. On the next line, at the prompt β or ρ, highlight "$\neq 0$" and press ENTER

Step 5. Leave the line for "RegEq:" blank Step 6. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

LinRegTTest
Xlist: L1
Ylist: L2
Freq: 1
β or ρ : [≠0] <0 >0
RegEQ:
Calculate

TI-83+ and TI-84+
calculators

LinRegTTest
y = a + bx
β≠0 and ρ≠0
t = 2.657560155
p = .0261501512
df = 9
↓a = −173.513363
b = 4.827394209
s = 16.41237711
r^2 = .4396931104
r = .663093591
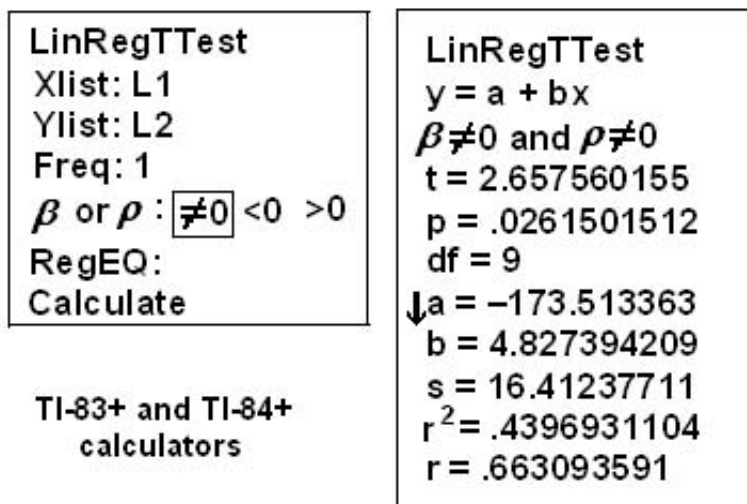
Figure 6.10 LinRegTTest Screen

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says y = a + bx. Scroll down to find the values a = -173.513, and b = 4.8273 ; the equation of

the best fit line is

$$\hat{y} = -173.51 + 4.83x$$

The two items at the bottom are r^2 = .43969 and *r* =.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

Step 1. We are assuming your X data is already entered in list L1 and your Y data is in list L2

Step 2. Press 2nd STATPLOT ENTER to use Plot 1

Step 3. On the input screen for PLOT 1, highlight **On** and press ENTER

Step 4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER

Step 5. Indicate Xlist: L1 and Ylist: L2

Step 6. For Mark: it does not matter which symbol you highlight.

Step 7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will ft the window to the data

Step 8. To graph the best ft line, press the "Y " key and type the equation -173.5+4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.

Step 9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

**With contributions from Roberta Bloom

6.6 Linear Regression and Correlation: Correlation Coefficient and Coefficient of Determination

6.6.1 The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient**, **r,** developed by Karl Pearson in the early 1900s, is a numerical measure of the strength of association between the independent variable x and the dependent variable y.
The correlation coefficient is calculated as

$$r = \frac{n \cdot \sum x \cdot y - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{\left[n \cdot \sum x^2 - \left(\sum x\right)^2\right] \cdot \left[n \cdot \sum y^2 - \left(\sum y\right)^2\right]}}$$

where *n* = the number of data points.

If you suspect a linear relationship between *x* and *y*, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of *r* is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between *x* and *y*. Values of r close to -1 or to +1 indicate a stronger linear relationship between *x* and *y*.
- If *r* = 0 there is absolutely no linear relationship between *x* and *y* (**no linear correlation**).
- If *r* = 1, there is perfect positive correlation. If *r* = −1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of *r* means that when x increases, *y* tends to increase and when *x* decreases, *y* tends to decrease **(positive correlation).**
- A negative value of *r* means that when x increases, *y* tends to decrease and when *x* decreases, *y* tends to increase **(negative correlation).**
- The sign of *r* is the same as the sign of the slope, *b*, of the best fit line.

Note: Strong correlation does not suggest that x causes y or y causes x. We say "correlation does not imply causation." For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!

Figure 6.11 (a) A scatter plot showing data with a positive correlation. 0 <r< 1 (b) A scatter plot showing data with a negative correlation. −1 <r< 0

Figure 6.12 (C) A scatter plot showing data with zero correlation. r=0

The formula for *r* looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate *r*. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

6.6.2 The Coefficient of Determination

r^2 is called the coefficient of determination. r^2 is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. r^2 has an interpretation in the context of the data:

- **r^2**, when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- **1-r^2**, when expressed as a percent, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

The line of best fit is: $\hat{y} = -173.51 + 4.83x$

The correlation coefficient is *r* = 0.6631

The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$

Interpretation of r^2 in the context of this example:

Approximately 44% of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best fit regression line.

Therefore approximately 56% of the variation (1 - 0.44 = 0.56) in the fnal exam grades can NOT be explained by the variation in the grades on the third exam, using the best fit regression line. (This is seen as the scattering of the points about the line.)

**With contributions from Roberta Bloom.

6.7 Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient

6.7.1 Testing the Significance of the Correlation Coefficient

The correlation coefficient, *r*, tells us about the strength of the linear relationship between *x* and *y*. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient *r* and the sample size *n*, together.

We perform a hypothesis test of the "**significance of the correlation coefficient**" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data is used to computer *r*, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we only have sample data, we can not calculate the population correlation coefficient. The sample correlation coefficient, *r*, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is *ρ*, the Greek letter "rho".

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient *ρ* is "close to 0" or "significantly different from 0". We decide this based on the sample correlation coefficient *r* and the sample size *n*.

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is "significant".

- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is significantly different from 0."

- What the conclusion means: There is a significant linear relationship between *x* and *y*. We can use the regression line to model the linear relationship between *x* and *y* in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is not significantly different from 0."
- What the conclusion means: There is not a significant linear relationship between *x* and *y*. Therefore we can NOT use the regression line to model a linear relationship between *x* and *y* in the population.

Note:
- If *r* is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If *r* is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If *r* is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST
SETTING UP THE HYPOTHESES:

- **Null Hypothesis:**

$$H_0 : \rho = 0$$

- **Alternate Hypothesis:**

$$H_a : \rho \neq 0$$

What the hypotheses mean in words:

- **Null Hypothesis H_o**: The population correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship(correlation) between *x* and *y* in the population.
- **Alternate Hypothesis Ha**: The population correlation coefficient IS significantly DIFFERENT FROM 0. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between *x* and *y* in the population.

DRAWING A CONCLUSION:

There are two methods to make the decision. Both methods are equivalent and give the same result.

Method 1: Using the p-value

Method 2: Using a table of critical values

In this chapter of this textbook, we will always use a significance level of 5%, α = 0.05

Note: Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using α = 0.05. But the table

of critical values provided in this textbook assumes that we are using a significance level of 5%, α = 0.05. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

The linear regression t-test LinRegTTEST on the TI-83+ or TI-84+ calculators calculates the p-value.

On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq 0$"
The output screen shows the p-value on the line that reads "p = ".
(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level (α = 0.05):

- Decision: REJECT the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."

If the p-value is NOT less than the significance level (α = 0.05):

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is NOT significantly different from 0."

Calculation Notes:

You will use technology to calculate the p-value. The following describe the calculations to compute the test statistics and the p-value:
The p-value is calculated using a t-distribution with n − 2 degrees of freedom.

The formula for the test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, *t*, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient *r*.

The p-value is the combined area in both tails.

An alternative way to calculate the p-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD EXAM vs FINAL EXAM EXAMPLE: p value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with r =0.6631 and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (*x* value), can we use the line to predict the final exam score (predicted *y* value)?**

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$
$$\alpha = 0.05$$

The p-value is 0.026 (from LinRegTTest on your calculator or from computer software)

The p-value, 0.026, is less than the signifcance level of $\alpha = 0.05$

Decision: Reject the Null Hypothesis H_0

Conclusion: There is sufcient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table at the end of this chapter (before the Summary) may be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 6.7

Suppose you computed $r = 0.801$ using $n = 10$ data points. df = $n - 2 = 10 - 2 = 8$. The critical values associated with df = 8 are -0.632 and $+0.632$. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

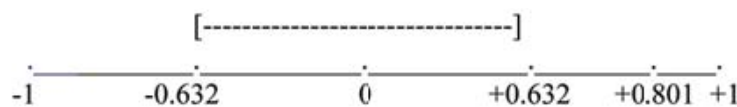**Figure 6.13 r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.**

Example 6.8

Suppose you computed $r = -0.624$ with 14 data points. df = $14 - 2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line may be used for prediction

Figure 6.14 $r = -0.624 < -0.532$. Therefore, r is significant.

Example 6.9

Suppose you computed $r = 0.776$ and $n = 6$. df $= 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant and the line should not be used for prediction.

Figure 6.15 $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

THIRD EXAM vs FINAL EXAM EXAMPLE: critical value method

- Consider the third exam final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$
$$\alpha = 0.05$$

Use the "95% Critical Value" table for r with df $= n - 2 = 11 - 2 = 9$

The critical values are -0.602 and $+0.602$

Since $0.6631 > 0.602$, r is significant.

Decision: Reject **H₀**:

Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example 6.10: Additional Practice Examples using Critical Values

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = -0.567$ and the sample size, n, is 19. The df $= n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
2. $r = 0.708$ and the sample size, n, is 9. The df $= n - 2 = 7$. The critical value is 0.666.
3. $0.708 > 0.666$ so r is significant.

4. r = 0.134 and the sample size, n, is 14. The df = 14 – 2 = 12. The critical value is 0.532.
5. 0.134 is between –0.532 and 0.532 so r is not significant.
6. r = 0 and the sample size, n, is 5. No matter what the dfs are, r = 0 is between the two critical values so r is not significant.

6.7.2 Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best fit line for our particular sample. We want to use this best fit line for the sample as an estimate of the best fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) above implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
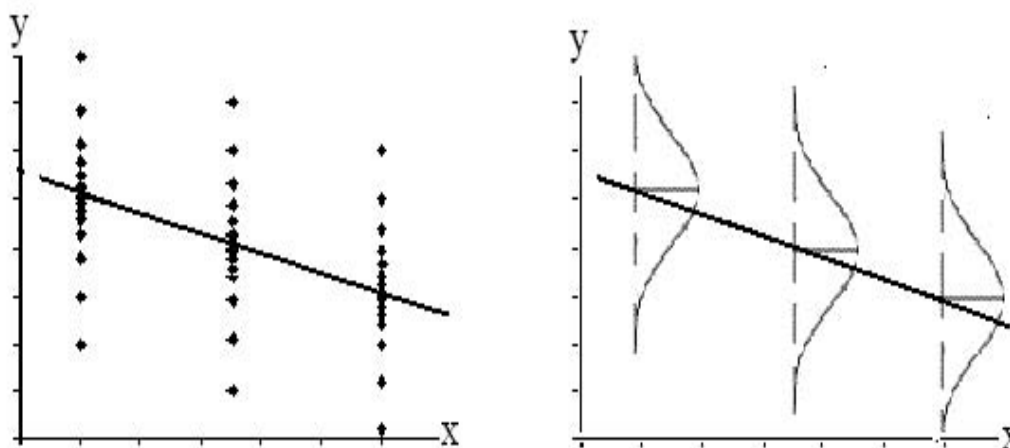- The residual errors are mutually independent (no pattern).

Figure 6.16 The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

**With contributions from Roberta Bloom

6.8 Linear Regression and Correlation: Prediction

Recall the third exam/final exam example.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best fit line for the final exam grade as a function of the grade on the third exam. We can now use the least squares regression line for prediction.

Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The exam scores (*x-values*) range from 65 to 75. **Since 73 is between the *x*-values 65 and 75**, substitute x = 73 into the equation. Then:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistic students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 6.11

Recall the third exam/final exam example.

See Problem 1 – 2.

Problem 1

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution

145.27

Problem 2

What would you predict the final exam score to be for a student who scored a 90 on the third exam?

**With contributions from Roberta Bloom

6.8.1 Solutions to Exercises in Chapter 6

Solution to Example 6.11, Problem 2

The x values in the data are between 65 and 75. 90 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter x into the equation and calculate a y value, you should not do so!)

To really understand how unreliable the prediction can be outside of the observed x values in the data, make the substitution x = 90 into the equation.

$$\hat{y} = -173.51 + 4.83(90) = 261.19$$

The final exam score is predicted to be 261.19. The largest the final exam score can be is 200.

Note: The process of predicting inside of the observed x values in the data is called interpolation. The process of predicting outside of the observed x values in the data is called extrapolation.

Chapter 7 Glossary

A

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

B

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, *n*, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV **X** is defined as the number of successes in n trials. The notation is: **X ~ B (n, p)**. The mean is **μ = np** and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly *x* successes in *n* trials is $P(X = x) = \binom{n}{x}p^x q^{n-x}$.

C

Central Limit Theorem

Given a random variable (RV) with known mean *μ* and known standard deviation **σ**. We are sampling with size n and we are interested in two new RVs - the sample mean, $\bar{X}$, and the sample sum, ΣX. If the size *n* of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$ and $\sum X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufciently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\dfrac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Coefficient of Correlation

A measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable. The formula is:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

where n is the number of data points. The coefficient cannot be more then 1 and less then -1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between *x* and *y*.

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Continuous Random Variable

A random variable (RV) whose outcomes are measured.

Example: The height of trees in the forest is a continuous RV.

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The Cumulative Relative Frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

D

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Degrees of Freedom (df)

The number of objects in a sample that are free to vary.

Discrete Random Variable

A random variable (RV) whose outcomes are counted.

E

Error Bound for a Population Mean (EBM)

The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion(EBP)

The margin of error. Depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

Exponential Distribution

A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: X~Exp (m). The mean is $\mu = \dfrac{1}{m}$ and the standard deviation is $\sigma = \dfrac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

F

Frequency

The number of times a value of the data occurs.

H

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

I

Inferential Statistics

Also called statistical inference or inductive statistics. This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if 4 out of the 100 calculators sampled are defective we might infer that 4 percent of the production is defective.

L

Level of Significance of the Test

Probability of a Type I error (reject the null hypothesis when it is true). Notation: α. In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

M

Mean

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

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

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

Median

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

Mode

The value that appears most frequently in a set of data.

N

Normal Distribution

A continuous random variable (RV) with pdf $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)2/2\sigma^2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N\ (\mu,\ \sigma)$. If μ =0 and σ =1, the RV is called the **standard normal distribution**.

P

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Parameter

A numerical characteristic of the population.

Point Estimate

A single number computed from a sample and used to estimate a population parameter.

Population

The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Proportion

- As a number: A proportion is the number of successes divided by the total number in the sample.
- As a probability distribution: Given a binomial random variable (RV), $X \sim B\ (n,\ p)$, consider the ratio of the number X of successes in n Bernouli trials to the number n of trials. $P' = \dfrac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, $P' \sim N\left(P, \dfrac{pq}{n}\right)$.

Q

Qualitative Data

See **Data.**

Quantitative Data

R

Relative Frequency

The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

S

Sample

A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Standard Deviation

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

Standard Error of the Mean

The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$.

Standard Normal Distribution

A continuous random variable (RV) *X-N* (0, 1) .. When X follows the standard normal distribution, it is often noted as *Z-N* (0, 1).

Statistic

A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

Student's-t Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and fatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

T

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

U

Uniform Distribution

A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: **X~U (a, b)**. The mean is $\mu = \dfrac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\dfrac{(b-a)^2}{12}}$ The probability density function is $f(x) = \dfrac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \dfrac{x-a}{b-a}.$

V

Variance

Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Z

z-score

The linear transformation of the form $z = \dfrac{x - \mu}{\sigma}$. If this transformation is applied to any normal distribution **X~N (μ, σ)** , the result is the standard normal distribution **Z~N (0, 1)**. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. Z-scores allow us to compare data that are normally distributed but scaled differently.