Quantitative Information Analysis III

Collection Editor:
Jeffrey Stanton

Quantitative Information Analysis III

Collection Editor:
Jeffrey Stanton

Authors:
Susan Dean
Barbara Illowsky, Ph.D.

Online:
< http://cnx.org/content/col11155/1.1/ >

CONNEXIONS

Rice University, Houston, Texas

Collection structure revised: December 25, 2009
PDF generated: August 8, 2013

Table of Contents

12 The Chi-Square Distribution

13 Linear Regression and Correlation

Chapter 1

Sampling and Data

1.1 Sampling and Data[1]

1.1.1 Student Learning Outcomes

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

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- 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 Statistics[2]

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

[1] This content is available online at <http://cnx.org/content/m16008/1.9/>.
[2] This content is available online at <http://cnx.org/content/m16020/1.16/>.

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:

Frequency of Average Time (in Hours) Spent Sleeping per Night

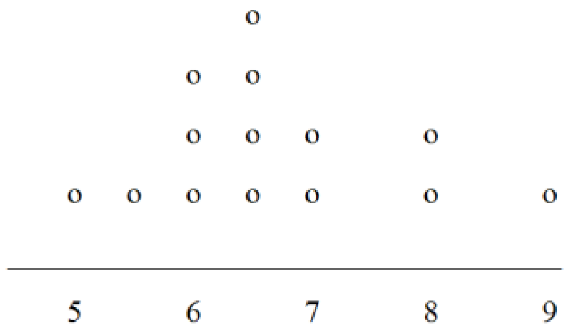


Figure 1.1

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, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that 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 difference. 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 differences 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 differences 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 definite ordering but there is a difference between data. The differences 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. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10^o F and -15^o 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^o C (nor is 80^o F 4 times as hot as 20^o 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 final 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 differences 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 Probability[3]

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

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

1.4 Key Terms[4]

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.

[3] This content is available online at <http://cnx.org/content/m16015/1.11/>.

[4] This content is available online at <http://cnx.org/content/m16007/1.17/>.

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

A **variable**, 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 affiliation, 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 affiliation makes no sense).

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

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and 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 $\frac{22}{40}$ 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 1.1

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.

1.4.1 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.

1.5 Data[5]

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.

[5] This content is available online at <http://cnx.org/content/m16005/1.18/>.

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 $\frac{\pi}{6}$, $\frac{\pi}{3}$,$\frac{\pi}{2}$, π , $\frac{3\pi}{4}$, 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.

De Anza College

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

Table 1.1

Foothill College

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

Table 1.2

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.

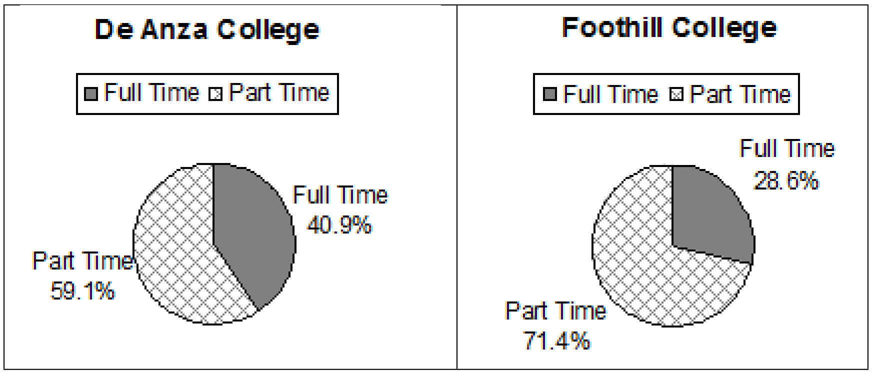


Table 1.3

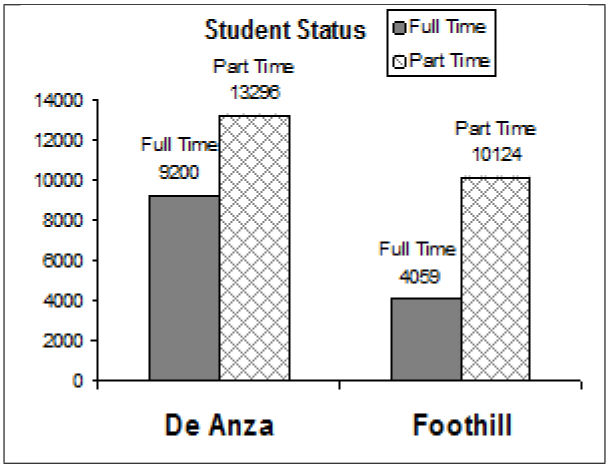


Table 1.4

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%.

De Anza College Spring 2010

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 1.5

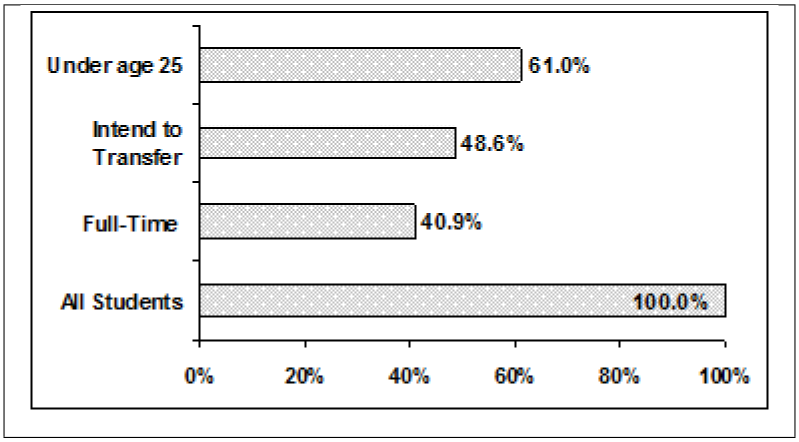


Table 1.6

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.

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

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.7

Bar graph Without Other/Unknown Category

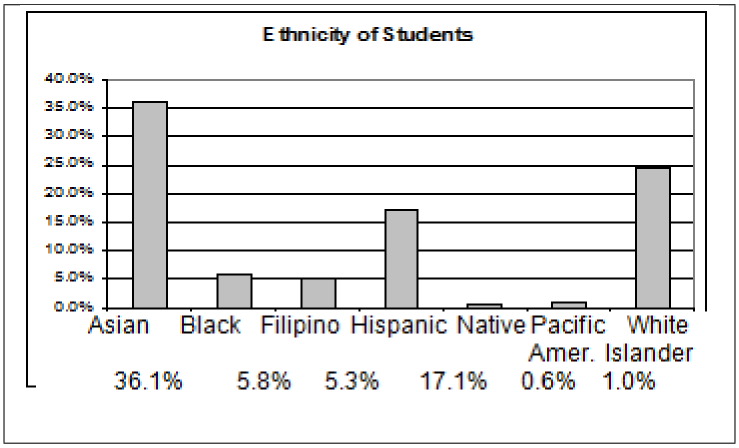


Table 1.8

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%, Pacific 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.

Bar Graph With Other/Unknown Category

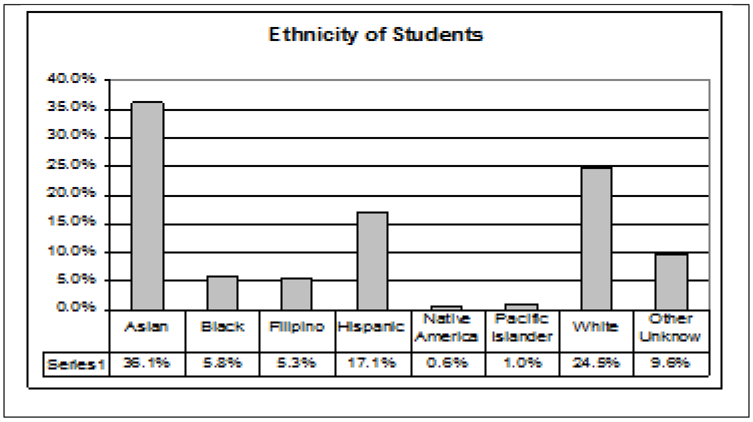


Table 1.9

Pareto Chart With Bars Sorted By Size

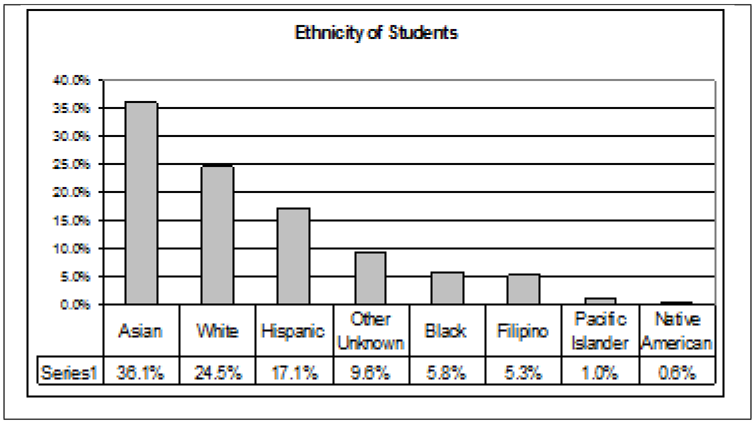


Table 1.10

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.

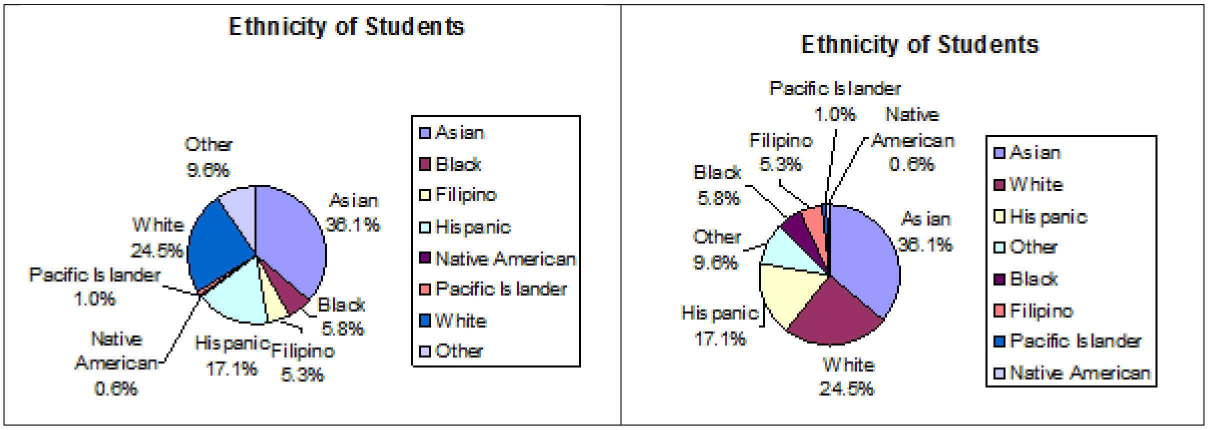


Table 1.11

1.6 Sampling[6]

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is

[6] This content is available online at <http://cnx.org/content/m16014/1.17/>.

equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

Class Roster

ID	Name
00	Anselmo
01	Bautista
02	Bayani
03	Cheng
04	Cuarismo
05	Cuningham
06	Fontecha
07	Hong
08	Hoobler
09	Jiao
10	Khan
11	King
12	Legeny
13	Lundquist
14	Macierz
15	Motogawa
16	Okimoto
17	Patel
18	Price
19	Quizon
20	Reyes
21	Roquero
22	Roth
23	Rowell
24	Salangsang
25	Slade
26	Stracher
27	Tallai
28	Tran
29	Wai
30	Wood

Table 1.12

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360; .99832; .14669; .51470; .40581; .73381; .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

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

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. For example, divide your college faculty by department. The departments are the clusters. Number each department and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

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

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

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

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

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

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To 4 decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To 4 decimal places, these numbers are not equivalent.

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

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

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

Example 1.6

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

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.

3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution

1. stratified
2. cluster
3. stratified
4. systematic
5. simple random
6. convenience

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

Example 1.7

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

Suppose we take two different samples.

First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:

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

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:

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

Problem 1

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Problem 2
Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

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

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

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

Problem 3
Is the sample biased?

Solution
The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

1.6.1 Optional Collaborative Classroom Exercise

Exercise 1.6.1
As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

1.7 Variation[7]

1.7.1 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.

1.7.2 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 different. Neither would be wrong, however.

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

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

1.7.2.1 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 sufficient 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.

[7] This content is available online at <http://cnx.org/content/m16021/1.15/>.

1.7.2.2 Optional Collaborative Classroom Exercise

Exercise 1.7.1

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, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.13

Second Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.14

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.

1.7.3 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 influence: Collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may 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.

1.8 Answers and Rounding Off[8]

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round only the final answer. Do not round any intermediate results, if possible. If it becomes necessary to round intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores 4, 6, 9 is 6.3, rounded to the nearest tenth, because the data are whole numbers. Most answers will be rounded in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics (Section 3.1), the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

1.9 Frequency[9]

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.

[8] This content is available online at <http://cnx.org/content/m16006/1.8/>.

[9] This content is available online at <http://cnx.org/content/m16012/1.20/>.

Frequency Table of Student Work Hours

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

Table 1.15

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.

Frequency Table of Student Work Hours w/ Relative Frequency

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

Table 1.16

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.

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

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

continued on next page

2	3	$\frac{3}{20}$ 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
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 1.17

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.

Frequency Table of Soccer Player Height

HEIGHTS (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 1.18

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 1.8

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 1.9

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 1.10

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.

1.9.1 Optional Collaborative Classroom Exercise

Exercise 1.9.1

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 1.11
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:

Frequency of Commuting Distances

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

Table 1.19

Problem ***(Solution on p. 27.)***

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)?

1.10 Summary[10]

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
    - Discrete (You count it.)
    - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement**: A member of the population may be chosen more than once
- **Without Replacement**: A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
    - Simple random sample
    - Stratified sample
    - Cluster sample
    - Systematic sample
- Not Random
    - Convenience sample

Frequency (freq. or f)

- The number of times an answer occurs

Relative Frequency (rel. freq. or RF)

- The proportion of times an answer occurs
- Can be interpreted as a fraction, decimal, or percent

Cumulative Relative Frequencies (cum. rel. freq. or cum RF)

- An accumulation of the previous relative frequencies

[10] This content is available online at <http://cnx.org/content/m16023/1.10/>.

Solutions to Exercises in Chapter 1

Solution to Example 1.5, Problem (p. 7)
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 1.10, Problem (p. 24)

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 1.11, Problem (p. 25)

1. No. Frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. 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.
3. $\frac{5}{19}$
4. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Chapter 2

Descriptive Statistics

2.1 Descriptive Statistics[1]

2.1.1 Student Learning Outcomes

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 Displaying Data[2]

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.

[1] This content is available online at <http://cnx.org/content/m16300/1.9/>.

[2] This content is available online at <http://cnx.org/content/m16297/1.9/>.

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 briefly look at stem-and-leaf plots, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

2.3 Stem and Leaf Graphs (Stemplots)[3]

One simple graph, the **stem-and-leaf graph** or **stem plot**, comes from the field of exploratory data analysis.It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.1
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

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

Stem-and-Leaf Diagram

Stem	Leaf
3	3
4	299
5	355
6	1378899
7	2348
8	03888
9	0244446
10	0

Table 2.1

The stem plot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stem plot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example 2.2
Create a stem plot using the data:

[3] This content is available online at <http://cnx.org/content/m16849/1.17/>.

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

The data are the distance (in kilometers) from a home to the nearest supermarket.

Problem ***(Solution on p. 60.)***

1. Are there any values that might possibly be outliers?
2. Do the data seem to have any concentration of values?

HINT: The leaves are to the right of the decimal.

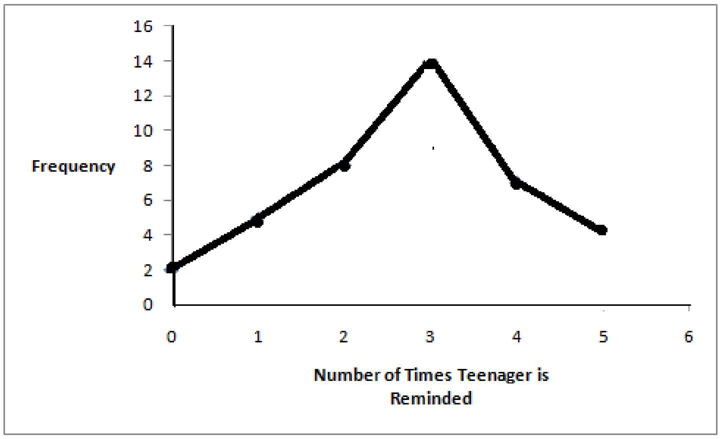
Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequency points**. The frequency points are connected.

Example 2.3

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

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

Table 2.2



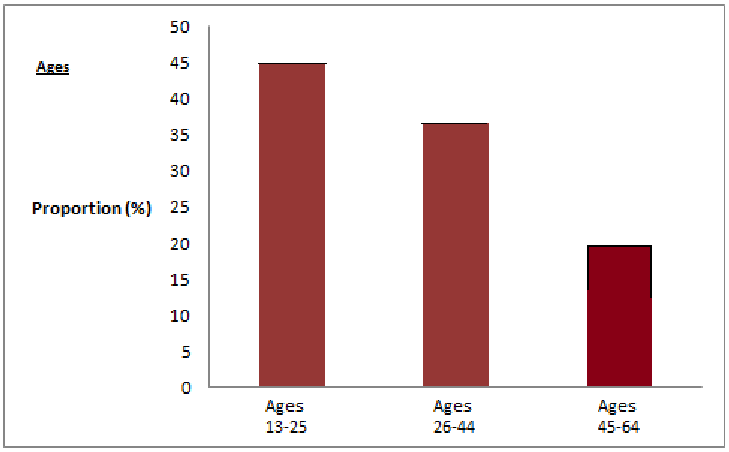
Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal.

The **bar graph** shown in **Example 4** has age groups represented on the **x-axis** and proportions on the **y-axis**.

Example 2.4
By the end of 2011, in the United States, Facebook had over 146 million users. The table shows three age groups, the number of users in each age group and the proportion (%) of users in each age group. **Source: *http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/***

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

Table 2.3



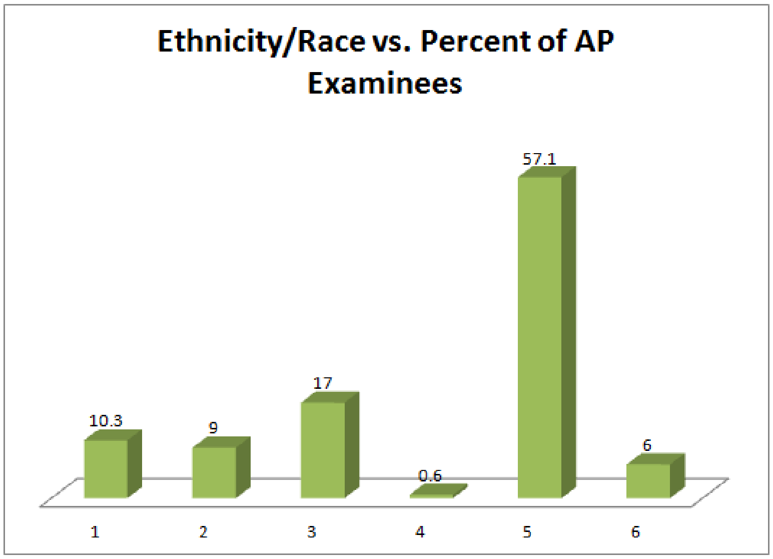
Example 2.5
The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2011, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools (qualitative data) on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**. (**Source: http://www.collegeboard.com** and **Source: http://apreport.collegeboard.org/goals-and-findings/promoting-equity**)

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%

continued on next page

2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Table 2.4



Go to Outcomes of Education Figure 22[4] for an example of a bar graph that shows unemployment rates of persons 25 years and older for 2009.

> NOTE: This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the Texas Instruments (TI) website[5] .

2.4 Histograms[6]

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.)

[4] http://nces.ed.gov/pubs2011/2011015_5.pdf
[5] http://education.ti.com/educationportal/sites/US/sectionHome/support.html
[6] This content is available online at <http://cnx.org/content/m16298/1.14/>.

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:

$$\text{RF} = \frac{f}{n} \tag{2.1}$$

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

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 2.6

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; 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.

$$\frac{74.05 - 59.95}{8} = 1.76 \tag{2.2}$$

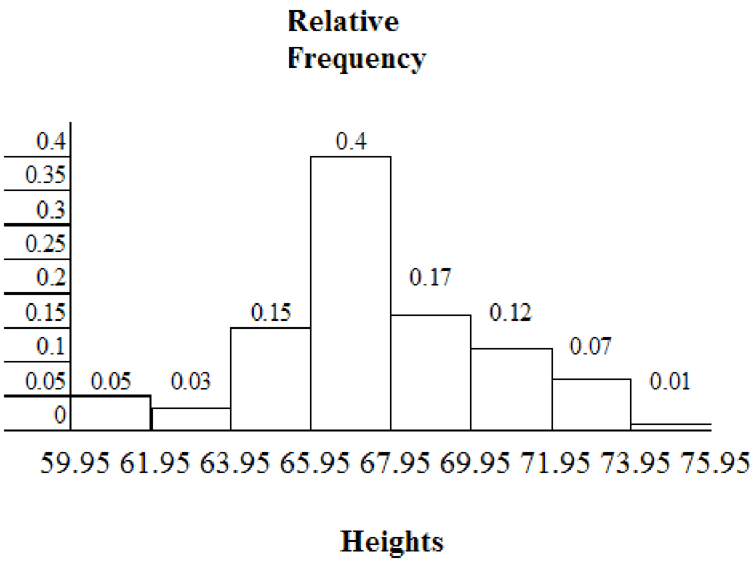
NOTE: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also 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 2.7

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 ***(Solution on p. 60.)***

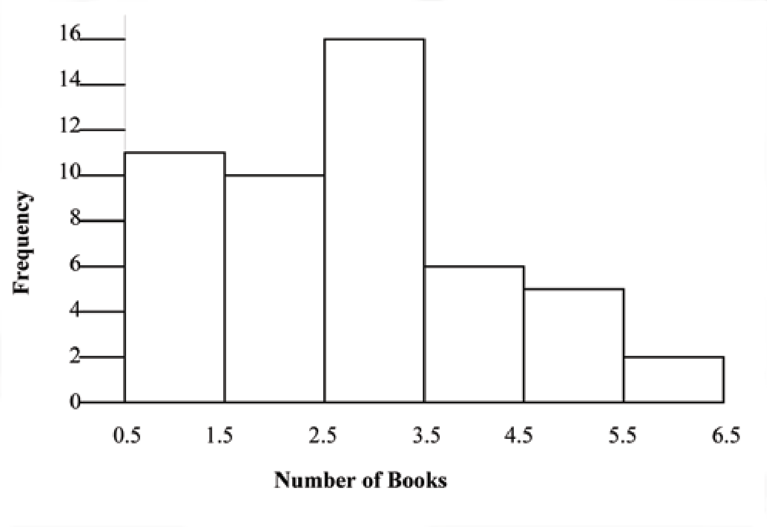
Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting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

$$\frac{6.5 - 0.5}{bars} = 1 \tag{2.3}$$

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:Plotsoff 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.

2.4.1 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.

2.5 Box Plots[7]

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The **median**, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger. For example, consider the following data:

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

Ordered from smallest to largest:

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

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.

$$\frac{6.8+7.2}{2}=7 \tag{2.4}$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of the data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

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

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

1; 1; 2; **2**; 4; 6; 6.8

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

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

7.2; 8; 8.3; **9**; 10; 10; 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

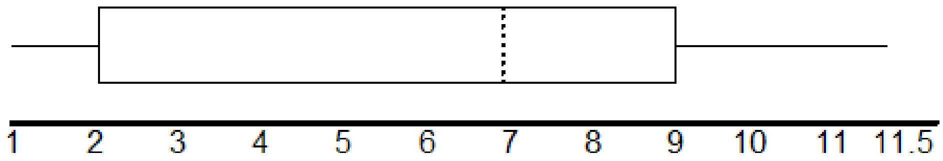
[7] This content is available online at <http://cnx.org/content/m16296/1.13/>.

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

Consider the following data:

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

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the TI web site[8]):



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

Example 2.8

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

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

Construct a box plot:

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

- Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, arrow down.
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
- Press ENTER
- Use the down and up arrow keys to scroll.

- Smallest value = 59
- Largest value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

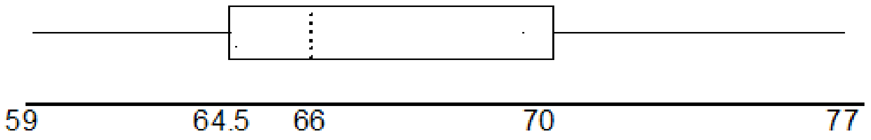
Using the TI-83, 83+, 84, 84+ to Construct the Box Plot

Go to 14:Appendix for Notes for the TI-83, 83+, 84, 84+ Calculator. To create the box plot:

- Press Y=. If there are any equations, press CLEAR to clear them.
- Press 2nd Y=.
- Press 4:Plotsoff. Press ENTER

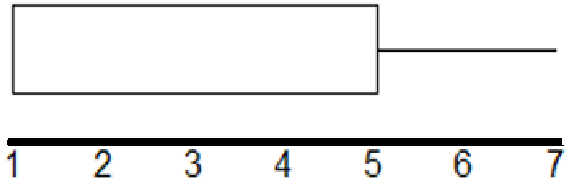
[8] http://education.ti.com/educationportal/sites/US/sectionHome/support.html

- Press 2nd Y=
- Press 1:Plot1. Press ENTER.
- Arrow down and then use the right arrow key to go to the 5th picture which is the box plot. Press ENTER.
- Arrow down to Xlist: Press 2nd 1 for L1
- Arrow down to Freq: Press ALPHA. Press 1.
- Press ZOOM. Press 9:ZoomStat.
- Press TRACE and use the arrow keys to examine the box plot.



a. Each quarter has 25% of the data.
b. The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
d. The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
e. The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example 2.9
Test scores for a college statistics class held during the day are:

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

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

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

Problem ***(Solution on p. 60.)***

- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

The first data set (the top box plot) has the widest spread for the middle 50% of the data. $IQR = Q3 - Q1$ is $82.5 - 56 = 26.5$ for the first data set and $89 - 78 = 11$ for the second data set. So, the first set of data has its middle 50% of scores more spread out.

25% of the data is between M and $Q3$ and 25% is between $Q3$ and $Xmax$.

2.6 Measures of the Location of the Data[9]

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3, is the same as the 75th percentile (75th %ile). The median, M, is called both the second quartile and the 50th percentile (50th %ile).

NOTE: Quartiles are given special attention in the Box Plots module in this chapter.

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

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

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

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

$$IQR = Q_3 - Q_1 \tag{2.5}$$

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

[9] This content is available online at <http://cnx.org/content/m16314/1.18/>.

Example 2.10
For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

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

Solution
Order the data from smallest to largest.

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

$M = 488,800$

$Q_1 = \frac{230500+387000}{2} = 308750$

$Q_3 = \frac{639000+659000}{2} = 649000$

$IQR = 649000 - 308750 = 340250$

$(1.5)\,(IQR) = (1.5)\,(340250) = 510375$

$Q_1 - (1.5)\,(IQR) = 308750 - 510375 = -201625$

$Q_3 + (1.5)\,(IQR) = 649000 + 510375 = 1159375$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2.11
For the two data sets in the test scores example (p. 40), find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.
c. The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Example 2.12: Finding Quartiles and Percentiles Using a Table
Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

continued on next page

4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Table 2.5

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

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

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example 2.13
Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile.
4. What is another name for the first quartile?

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct a table of the data.
4. Construct 2 different histograms. For each, starting value = _____ ending value = ____.
5. Use the table to find the median, first quartile, and third quartile.
6. Construct a box plot.
7. Use the table to find the following:
    - The 10th percentile
    - The 70th percentile
    - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median
A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. p% of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

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

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

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

Guideline:

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

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example 2.14
On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 2.15
On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example 2.16
At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles

Exercise 2.6.1 ***(Solution on p. 61.)***

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?

b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

c. A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 2.6.2 ***(Solution on p. 62.)***

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?

b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Exercise 2.6.3 ***(Solution on p. 62.)***

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

Exercise 2.6.4 ***(Solution on p. 62.)***

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

Exercise 2.6.5 ***(Solution on p. 62.)***

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

Exercise 2.6.6 ***(Solution on p. 62.)***

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Exercise 2.6.7 ***(Solution on p. 62.)***

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

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

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

Exercise 2.6.8 ***(Solution on p. 62.)***

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

**With contributions from Roberta Bloom

2.7 Measures of the Center of the Data[10]

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"): $\overline{x}$.

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

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{2.6}$$

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

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 2.17
AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

[10] This content is available online at <http://cnx.org/content/m17102/1.13/>.

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:

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

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.

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

Example 2.18
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
$\overline{x} = \frac{5000000+49\times30000}{50} = 129400$

$M = 30000$

(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 2.19: Statistics exam scores for 20 students are as follows
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 2.20

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.

2.7.1 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 $\overline{x}$ 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 2.10) section course.

2.7.2 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

Table 2.6

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

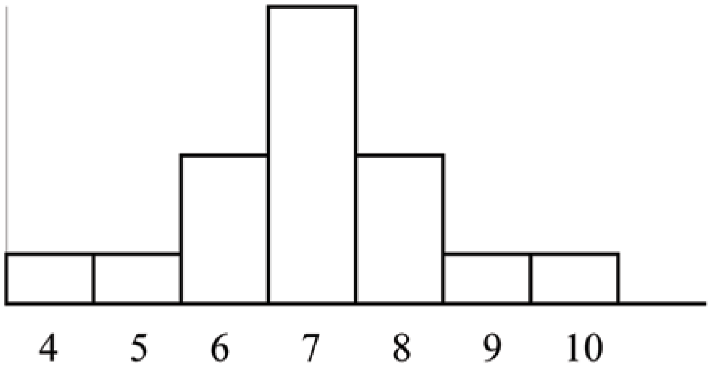
A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\overline{x}$ is an example of a statistic which estimates the population mean μ.

2.8 Skewness and the Mean, Median, and Mode[11]

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.



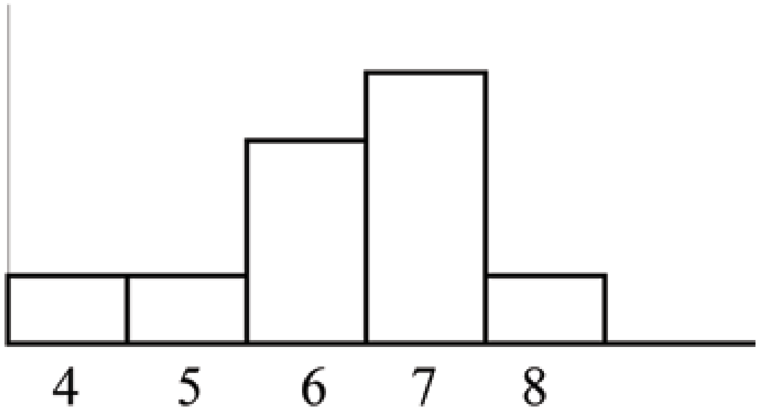
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 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.

[11] This content is available online at <http://cnx.org/content/m17104/1.9/>.

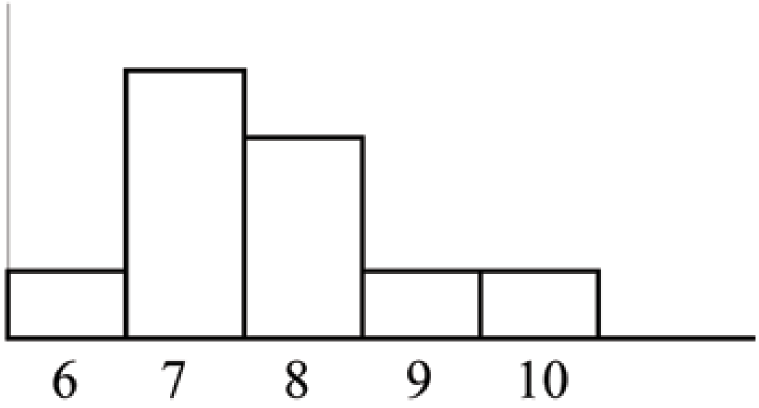


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

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 reflects the skewing the most.

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

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

2.9 Measures of the Spread of the Data[12]

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** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- 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.

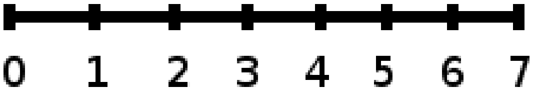
Binh waits for 1 minute.

- 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 affects 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
$5 + (1)(2) = 7$.

[12] This content is available online at <http://cnx.org/content/m17103/1.15/>.

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:** $x = \overline{x} + (\#ofSTDEV)(s)$
- **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 $\overline{x}$ is the sample mean and the Greek symbol μ is the population mean.

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

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 $x- \overline{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **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

- $s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f\cdot(x-\overline{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\overline{\mu})^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f\cdot(x-\overline{\mu})^2}{N}}$
- 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** is 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 $\frac{\sigma}{\sqrt{n}}$ 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 2.21
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

$$\overline{x} = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525 \tag{2.8}$$

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 calculating s.

Data	**Freq.**	**Deviations**	*Deviations*^{2}	**(Freq.)(***Deviations*^{2}**)**
x	f	$(x-\overline{x})$	$(x-\overline{x})^2$	$(f)\,(x-\overline{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

Table 2.7

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.
- $\overline{x}$=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 = $\overline{x}$ + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = $\overline{x}$ + (#ofSTDEVs)(s) because the data is from a sample

Problem 2
Find the value that is 1 standard deviation above the mean. Find $(\overline{x} + 1s)$.

Solution
$(\overline{x} + 1s) = 10.53 + (1)\,(0.72) = 11.25$

Problem 3
Find the value that is two standard deviations below the mean. Find $(\overline{x} - 2s)$.

Solution
$(\overline{x} - 2s) = 10.53 - (2)\,(0.72) = 9.09$

Problem 4
Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

- $(\overline{x} - 1.5s) = 10.53 - (1.5)\,(0.72) = 9.45$

- $\left(\overline{x}+1.5s\right)=10.53+\left(1.5\right)\left(0.72\right)=11.61$

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 2.22
Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

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

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

i. The sample mean
ii. The sample standard deviation
iii. The median

   - **iv.** The first quartile
   - **v.** The third quartile
   - **vi.** IQR

c. 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

a.

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
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?)

Table 2.8

b. **i.** The sample mean = 73.5
   - **ii.** The sample standard deviation = 17.9
   - **iii.** The median = 73
   - **iv.** The first quartile = 61
   - **v.** The third quartile = 90
   - **vi.** IQR = 90 - 61 = 29

c. 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.

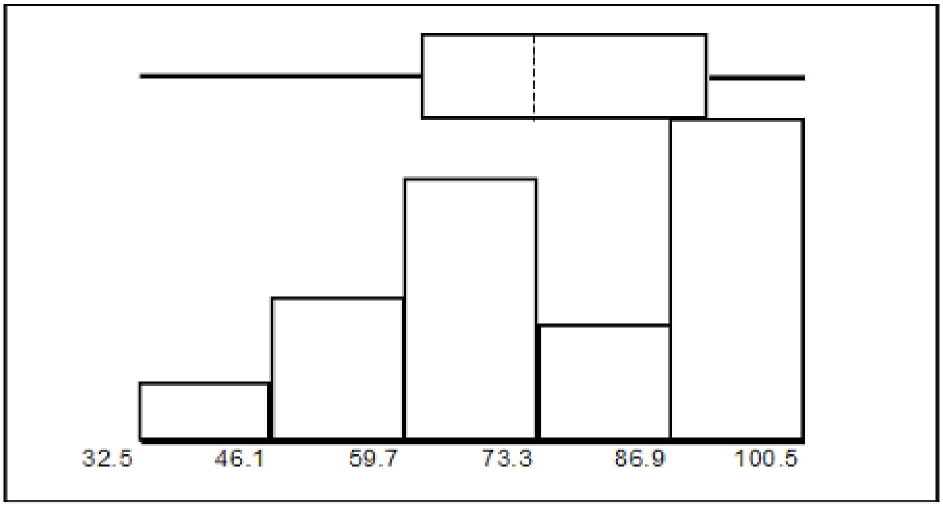


Figure 2.1

The long left whisker in the box plot is reflected 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 reflect 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	$x = \overline{x} + z\,s$	$z = \frac{x-\overline{x}}{s}$
Population	$x = \mu + z\,\sigma$	$z = \frac{x-\mu}{\sigma}$

Table 2.9

Example 2.23
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

Table 2.10

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

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

2.10 Summary of Formulas[13]

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- $\overline{x}$ = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x * f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x * f$ = The sum of values multiplied by their respective frequencies
- $(x - \overline{x})$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean)
- $(x - \overline{x})^2$ or $(x - \mu)^2$ = Deviations squared
- $f\ (x - \overline{x})^2$ or $f\ (x - \mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $\overline{x} = \frac{\sum x}{n}$ or $\overline{x} = \frac{\sum f \cdot x}{n}$
- $\mu = \frac{\sum x}{N}$ or $\mu = \frac{\sum f \cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x-\overline{x})^2}{n-1}}$
- $\sigma = \sqrt{\frac{\Sigma(x-\overline{\mu})^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot (x-\overline{\mu})^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- x = $\overline{x}$+ (#ofSTDEVs)(s)
- x = μ + (#ofSTDEVs)(σ)

[13] This content is available online at <http://cnx.org/content/m16310/1.9/>.

Solutions to Exercises in Chapter 2

Solution to Example 2.2, Problem (p. 31)
The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Table 2.11

Solution to Example 2.7, Problem (p. 36)

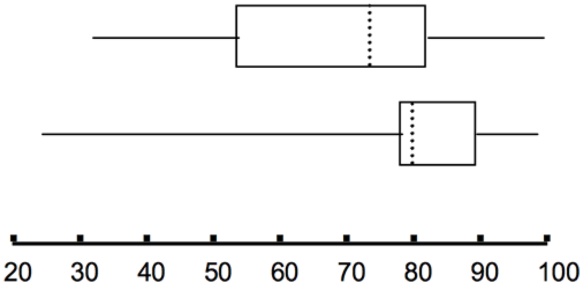
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Solution to Example 2.9, Problem (p. 40)
First Data Set

- $Xmin = 32$
- $Q1 = 56$
- $M = 74.5$
- $Q3 = 82.5$
- $Xmax = 99$

Second Data Set

- $Xmin = 25.5$
- $Q1 = 78$
- $M = 81$
- $Q3 = 89$
- $Xmax = 98$



Solution to Example 2.11, Problem (p. 42)
For the IQRs, see the answer to the test scores example (Solution to Example 2.9: p. 60). The first data set has the larger IQR, so the scores between $Q3$ and $Q1$ (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot (IQR) = \left(\frac{3}{2}\right) \cdot (26.5) = 39.75$
- $Xmax - Q3 = 99 - 82.5 = 16.5$
- $Q1 - Xmin = 56 - 32 = 24$

$\left(\frac{3}{2}\right) \cdot (IQR) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (IQR) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $Xmax - Q3 = 98 - 89 = 9$
- $Q1 - Xmin = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (IQR) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and Data Chapter (Section 1.9)). Get the percentiles from that chart.

First Data Set

- 30th %ile (between the 6th and 7th values) $= \frac{(56 + 59)}{2} = 57.5$
- 80th %ile (between the 16th and 17th values) $= \frac{(84 + 84.5)}{2} = 84.25$

Second Data Set

- 30th %ile (7th value) $= 78$
- 80th %ile (18th value) $= 90$

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Solution to Example 2.13, Problem (p. 43)

1. $\frac{(8 + 9)}{2} = 8.5$
   Look where cum. rel. freq. = 0.80. 80% of the data is 8 or less. 80th %ile is between the last 8 and first 9.
2. 9
3. 6
4. First Quartile = 25th %ile

Solution to Exercise 2.6.1 (p. 45)

a. For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.

b. INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.

c. He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Solution to Exercise 2.6.2 (p. 45)

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

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

Solution to Exercise 2.6.3 (p. 45)

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Solution to Exercise 2.6.4 (p. 45)

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

Solution to Exercise 2.6.5 (p. 45)

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Solution to Exercise 2.6.6 (p. 45)

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

Solution to Exercise 2.6.7 (p. 45)

a. The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.

b. The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Solution to Exercise 2.6.8 (p. 45)

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

Chapter 3

Probability Topics

3.1 Probability Topics[1]

3.1.1 Student Learning Outcomes

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

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams (optional).
- Construct and interpret Tree Diagrams (optional).

3.1.2 Introduction

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

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

3.1.3 Optional Collaborative Classroom Exercise

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

[1] This content is available online at <http://cnx.org/content/m16838/1.11/>.

Use the class data as estimates of the following probabilities. P(change) means the probability that a randomly chosen person in your class has change in his/her pocket or purse. P(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find P(change).
- Find P(bus).
- Find P(change and bus) Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find P(change | bus) Find the probability that a randomly chosen student has change given that he/she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

3.2 Terminology[2]

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

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

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between 0 and 1, inclusive** (includes 0 and 1 and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative fequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head(H) and a Tail(T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes $\{HT, TH\}$. $P(A) = \frac{2}{4}$.

Suppose you roll one fair six-sided die, with the numbers {1,2,3,4,5,6} on its faces. Let event E = rolling a number that is at least 5. There are two outcomes $\{5, 6\}$. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, 2/6 of the rolls would result in an

[2] This content is available online at <http://cnx.org/content/m16845/1.13/>.

outcome of "at least 5". You would not expect exactly 2/6. The long-term relative frequency of obtaining this result would approach the theoretical probability of 2/6 as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is the known as the **Law of Large Numbers**: as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes don't happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.) The Law of Large Numbers will be discussed again in Chapter 7.

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased** . Two math professors in Europe had their statistics students test the Belgian 1 Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos have a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later in this chapter we will learn techniques to use to work with probabilities for events that are not equally likely.

"OR" Event:
An outcome is in the event *A OR B* if the outcome is in *A* or is in *B* or is in both *A* and *B*. For example, let $A = \{1,\ 2,\ 3,\ 4,\ 5\}$ and $B = \{4,\ 5,\ 6,\ 7,\ 8\}$. $A\ OR\ B = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8\}$. Notice that 4 and 5 are NOT listed twice.

"AND" Event:
An outcome is in the event A AND B if the outcome is in both *A* and *B* at the same time. For example, let *A* and *B* be $\{1,\ 2,\ 3,\ 4,\ 5\}$ and $\{4,\ 5,\ 6,\ 7,\ 8\}$, respectively. Then A AND B = $\{4, 5\}$.

The **complement** of event *A* is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in *A*. Notice that $P(A) + P(A') = 1$. For example, let $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ and let $A = \{1,\ 2,\ 3,\ 4\}$. Then, $A' = \{5,\ 6\}$. $P(A) = \frac{4}{6},\ P(A') = \frac{2}{6},\ and\ P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

The **conditional probability** of *A* given *B* is written $P(A|B)$. $P(A|B)$ is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate $P(A|B)$ is

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

where $P(B)$ is greater than 0.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$. Let *A* = face is 2 or 3 and *B* = face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2,\ 4,\ 6\}$. Then we divide that by the number of outcomes in *B* (and not *S*).

We get the same result by using the formula. Remember that *S* has 6 outcomes.

$$P(A|B) = \frac{\text{P(A and B)}}{\text{P(B)}} = \frac{\text{(the number of outcomes that are 2 or 3 and even in S) / 6}}{\text{(the number of outcomes that are even in S) / 6}} = \frac{1/6}{3/6} = \frac{1}{3}$$

Understanding Terminology and Symbols
It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Exercise 3.2.1 ***(Solution on p. 78.)***
In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts (a) through (j) below. (Note that you can't find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

**With contributions from Roberta Bloom

3.3 Independent and Mutually Exclusive Events[3]

Independent and mutually exclusive do **not** mean the same thing.

3.3.1 Independent Events

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A\ AND\ B) = P(A) \cdot P(B)$

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

[3] This content is available online at <http://cnx.org/content/m16837/1.14/>.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:**: When sampling is done without replacement, then each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether A and B are independent or dependent, **assume they are dependent until you can show otherwise**.

3.3.2 Mutually Exclusive Events

A and B are **mutually exclusive** events if they cannot occur at the same time. This means that A and B do not share any outcomes and P(A AND B) $= 0$.

For example, suppose the sample space $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\}$. Let $A = \{1,\ 2,\ 3,\ 4,\ 5\}$, $B = \{4,\ 5,\ 6,\ 7,\ 8\}$, and $C = \{7,\ 9\}$. A AND B $= \{4,5\}$. P(A AND B) = $\frac{2}{10}$and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so P(A AND C) = 0. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**.

The following examples illustrate these definitions and terms.

Example 3.1
Flip two fair coins. (This is an experiment.)

The sample space is $\{HH,\ HT,\ TH,\ TT\}$ where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means 0 or 1 tail.) Then A can be written as $\{HH,\ HT,\ TH\}$. The outcome HH shows 0 tails. HT and TH each show 1 tail.
- Let B = the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A. So, $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for A and for B are $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{4}$.
- Let C = the event of getting all heads. $C = \{HH\}$. Since $B = \{TT\}$, $P(B\ AND\ C) = 0$. B and C are mutually exclusive. (B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. $D = \{TT\}$. $P(D) = \frac{1}{4}$.
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$.
- Find the probability of getting **at least one** (1 or 2) tail in two flips. Let F = event of getting at least one tail in two flips. $F = \{HT, TH, TT\}$. P(F) $= \frac{3}{4}$

Example 3.2
Roll one fair 6-sided die. The sample space is $\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$. Let event A = a face is odd. Then $A = \{1,\ 3,\ 5\}$. Let event B = a face is even. Then $B = \{2,\ 4,\ 6\}$.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. P(A) + P(B) = P(A) + P(A') = 1. Also, P(A) $= \frac{3}{6}$ and P(B) $= \frac{3}{6}$

- Let event C = odd faces larger than 2. Then $C = \{3,5\}$. Let event D = all even faces smaller than 5. Then $D = \{2,4\}$. P(C and D) $= 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than 5. $E = \{1,2,3,4\}$.

  Problem ***(Solution on p. 78.)***
  Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?
- Find P(C | A). This is a conditional. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find P(C | A), find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, P(C | A) $= \frac{2}{3}$

Example 3.3
Let event G = taking a math class. Let event H = taking a science class. Then, $G\ AND\ H$ = taking a math class and a science class. Suppose P(G) $= 0.6$, P(H) $= 0.5$, and P(G AND H) $= 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- P(G | H) = P(G)
- P(H | G) = P(H)
- P(G AND H) = P(G) · P(H)

NOTE: **The choice you make depends on the information you have.** You could choose any of the methods here because you have the necessary information.

Problem 1
Show that P(G | H) = P(G).

Solution
$$\text{P(G | H)} = \frac{\text{P(G AND H)}}{\text{P(H)}} = \frac{0.3}{0.5} = 0.6 = \text{P(G)}$$

Problem 2
Show P(G AND H) = P(G) · P(H).

Solution
$$P(G) \cdot P(H) = 0.6 \cdot 0.5 = 0.3 = \text{P(G AND H)}$$

Since G and H are independent, then, knowing that a person is taking a science class does not change the chance that he/she is taking math. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he/she is taking math. For practice, show that P(H | G) = P(H) to show that G and H are independent events.

Example 3.4
In a box there are 3 red cards and 5 blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.

The sample space S = R1, R2, R3, B1, B2, B3, B4, B5. S has 8 outcomes.

- P(R) $= \frac{3}{8}$. P(B) $= \frac{5}{8}$. P(R AND B) $= 0$. (You cannot draw one card that is both red and blue.)

- P(E) $= \frac{3}{8}$. (There are 3 even-numbered cards, *R*2, *B*2, and *B*4.)
- P(E | B) $= \frac{2}{5}$. (There are 5 blue cards: *B*1, *B*2, *B*3, *B*4, and *B*5. Out of the blue cards, there are 2 even cards: *B*2 and *B*4.)
- P(B | E) $= \frac{2}{3}$. (There are 3 even-numbered cards: *R*2, *B*2, and *B*4. Out of the even-numbered cards, 2 are blue: *B*2 and *B*4.)
- The events *R* and *B* are mutually exclusive because P(R AND B) $= 0$.
- Let *G* = card with a number greater than 3. $G \ = \{B4, B5\}$. P(G) $= \frac{2}{8}$. Let *H* = blue card numbered between 1 and 4, inclusive. $H \ = \{B1, B2, B3, B4\}$. P(G | H) $= \frac{1}{4}$. (The only card in H that has a number greater than 3 is *B*4.) Since $\frac{2}{8} \ = \frac{1}{4}$, P(G) $=$ P(G | H) which means that *G* and *H* are independent.

Example 3.5

In a particular college class, 60% of the students are female. 50 % of all students in the class have long hair. 45% of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that the student is female. Let L be the event that the student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- P(F) $= 0.60$; P(L) $= 0.50$
- P(F AND L) $= 0.45$
- P(L | F) $= 0.75$

NOTE: **The choice you make depends on the information you have.** You could use the first or last condition on the list for this example. You do not know P(F | L) yet, so you can not use the second condition.

Solution 1

Check whether P(F and L) = P(F)P(L): We are given that P(F and L) = 0.45 ; but P(F)P(L) = (0.60)(0.50)= 0.30 The events of being female and having long hair are not independent because P(F and L) does not equal P(F)P(L).

Solution 2

check whether P(L | F) equals P(L): We are given that P(L | F) = 0.75 but P(L) = 0.50; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

**Example 5 contributed by Roberta Bloom

3.4 Two Basic Rules of Probability[4]

3.4.1 The Multiplication Rule

If *A* and *B* are two events defined on a **sample space**, then: P(A AND B) $=$ P(B) $\cdot$ P(A | B).

This rule may also be written as : $P\left(A|B\right) = \frac{P(A\ AND\ B)}{P(B)}$

[4] This content is available online at <http://cnx.org/content/m16847/1.11/>.

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are **independent**, then P(A | B) $=$ P(A). Then P(A AND B) $=$ P(A | B) P(B) becomes P(A AND B) $=$ P(A) P(B).

3.4.2 The Addition Rule

If A and B are defined on a sample space, then: P(A OR B) $=$ P(A) $+$ P(B) $-$ P(A AND B).

If A and B are **mutually exclusive**, then P(A AND B) $=$ 0. Then P(A OR B) $=$ P(A) $+$ P(B) $-$ P(A AND B) becomes P(A OR B) $=$ P(A) $+$ P(B).

Example 3.6

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is P(A) $= 0.6$ and the probability that he chooses B is P(B) $= 0.35$.
- P(A and B) $= 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is P(A OR B) $=$ P(A) $+$ P(B) $= 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example 3.7

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game.

A = the event Carlos is successful on his first attempt. P(A) $= 0.65$. B = the event Carlos is successful on his second attempt. P(B) $= 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

Problem 1

What is the probability that he makes both goals?

Solution

The problem is asking you to find P(A AND B) $=$ P(B AND A). Since P(B | A) $= 0.90$:

$$\text{P(B AND A)} = \text{P(B | A) P(A)} = 0.90 * 0.65 = 0.585 \tag{3.1}$$

Carlos makes the first and second goals with probability 0.585.

Problem 2

What is the probability that Carlos makes either the first goal or the second goal?

Solution

The problem is asking you to find P(A OR B).

$$\text{P(A OR B)} = \text{P(A)} + \text{P(B)} - \text{P(A AND B)} = 0.65 + 0.65 - 0.585 = 0.715 \tag{3.2}$$

Carlos makes either the first goal or the second goal with probability 0.715.

Problem 3

Are A and B independent?

Solution
No, they are not, because P(B AND A) $=$ 0.585.

$$\text{P(B)} \cdot \text{P(A)} = (0.65) \cdot (0.65) = 0.423 \tag{3.3}$$

$$0.423 \neq 0.585 = \text{P(B AND A)} \tag{3.4}$$

So, P(B AND A) is **not** equal to P(B) $\cdot$ P(A).

Problem 4
Are *A* and *B* mutually exclusive?

Solution
No, they are not because P(A and B) = 0.585.

To be mutually exclusive, P(A AND B) must equal 0.

Example 3.8
A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice 4 times a week. **Thirty** of the intermediate swimmers practice 4 times a week. **Ten** of the novice swimmers practice 4 times a week. Suppose one member of the swim team is randomly chosen. Answer the questions (Verify the answers):

Problem 1
What is the probability that the member is a novice swimmer?

Solution
$\frac{28}{150}$

Problem 2
What is the probability that the member practices 4 times a week?

Solution
$\frac{80}{150}$

Problem 3
What is the probability that the member is an advanced swimmer and practices 4 times a week?

Solution
$\frac{40}{150}$

Problem 4
What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

Solution
P(advanced AND intermediate) $=$ 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Problem 5
Are being a novice swimmer and practicing 4 times a week independent events? Why or why not?

Solution
No, these are not independent events.

$$\text{P(novice AND practices 4 times per week)} = 0.0667 \tag{3.5}$$

$$\text{P(novice)} \cdot \text{P(practices 4 times per week)} = 0.0996 \tag{3.6}$$

$$0.0667 \neq 0.0996 \tag{3.7}$$

Example 3.9
Studies show that, if she lives to be 90, about 1 woman in 7 (approximately 14.3%) will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

Problem 1
What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

Solution
$\text{P(B)} = 0.143$; $\text{P(N)} = 0.85$

Problem 2
Given that the woman has breast cancer, what is the probability that she tests negative?

Solution
$\text{P(N} \mid \text{B)} = 0.02$

Problem 3
What is the probability that the woman has breast cancer AND tests negative?

Solution
$\text{P(B AND N)} = \text{P(B)} \cdot \text{P(N} \mid \text{B)} = (0.143) \cdot (0.02) = 0.0029$

Problem 4
What is the probability that the woman has breast cancer or tests negative?

Solution
$\text{P(B OR N)} = \text{P(B)} + \text{P(N)} - \text{P(B AND N)} = 0.143 + 0.85 - 0.0029 = 0.9901$

Problem 5
Are having breast cancer and testing negative independent events?

Solution
No. P(N) = 0.85; P(N | B) = 0.02. So, P(N | B) does not equal P(N)

Problem 6
Are having breast cancer and testing negative mutually exclusive?

Solution
No. P(B AND N) = 0.0029. For *B* and *N* to be mutually exclusive, P(B AND N) must be 0.

3.5 Contingency Tables[5]

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner. Contingincy tables provide a way of portraying data that can facilitate calculating probabilities.

Example 3.10
Suppose a study of speeding violations and drivers who use car phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Car phone user	25	280	305
Not a car phone user	45	405	450
Total	70	685	755

Table 3.1

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that $305 + 450 = 755$ and $70 + 685 = 755$.

Calculate the following probabilities using the table

Problem 1
P(person is a car phone user) =

Solution
$\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$

Problem 2
P(person had no violation in the last year) =

[5] This content is available online at <http://cnx.org/content/m16835/1.12/>.

Solution
$\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

Problem 3
P(person had no violation in the last year AND was a car phone user) =

Solution
$\frac{280}{755}$

Problem 4
P(person is a car phone user OR person had no violation in the last year) =

Solution
$\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

Problem 5
P(person is a car phone user GIVEN person had a violation in the last year) =

Solution
$\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

Problem 6
P(person had no violation last year GIVEN person was not a car phone user) =

Solution
$\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Example 3.11
The following table shows a random sample of 100 hikers and the areas of hiking preferred:

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Table 3.2

Problem 1 ***(Solution on p. 78.)***
Complete the table.

Problem 2 ***(Solution on p. 78.)***
Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

a. P(F AND C) =

b. $P(F) \cdot P(C) =$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Problem 3 ***(Solution on p. 78.)***

Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male and let L = prefers hiking near lakes and streams.

a. What word tells you this is a conditional?
b. Fill in the blanks and calculate the probability: P(___ | ___) = ___.
c. Is the sample space for this problem all 100 hikers? If not, what is it?

Problem 4 ***(Solution on p. 78.)***

Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female and let P = prefers mountain peaks.

a. P(F) =
b. P(P) =
c. P(F AND P) =
d. Therefore, P(F OR P) =

Example 3.12

Muddy Mouse lives in a cage with 3 doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

Table 3.3

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Caught).
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Not Caught).

Verify the remaining entries.

Problem 1 ***(Solution on p. 78.)***

Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Problem 2

What is the probability that Alissa does not catch Muddy?

Solution

$\frac{41}{60}$

Problem 3

What is the probability that Muddy chooses Door One **OR** Door Two given that Muddy is caught by Alissa?

Solution

$\frac{9}{19}$

NOTE: You could also do this problem by using a probability tree. See the Tree Diagrams (Optional)[6] section of this chapter for examples.

[6] "Probability Topics: Tree Diagrams (optional)" <http://cnx.org/content/m16846/latest/>

3.6 Summary of Formulas[7]

Formula 3.1: Complement
If A and A' are complements then $P(A) + \mathrm{P(A')} = 1$

Formula 3.2: Addition Rule
$\mathrm{P(A\ OR\ B)} = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A\ AND\ B)}$

Formula 3.3: Mutually Exclusive
If A and B are mutually exclusive then $\mathrm{P(A\ AND\ B)} = 0$; so $\mathrm{P(A\ OR\ B)} = \mathrm{P(A)} + \mathrm{P(B)}$.

Formula 3.4: Multiplication Rule

- $\mathrm{P(A\ AND\ B)} = \mathrm{P(B)P(A \mid B)}$
- $\mathrm{P(A\ AND\ B)} = \mathrm{P(A)P(B \mid A)}$

Formula 3.5: Independence
If A and B are independent then:

- $\mathrm{P(A \mid B)} = \mathrm{P(A)}$
- $\mathrm{P(B \mid A)} = \mathrm{P(B)}$
- $\mathrm{P(A\ AND\ B)} = \mathrm{P(A)P(B)}$

[7] This content is available online at <http://cnx.org/content/m16843/1.5/>.

Solutions to Exercises in Chapter 3

Solution to Exercise 3.2.1 (p. 66)

a. P(L′)=P(S)
b. P(M or S)
c. P(F and L)
d. P(M | L)
e. P(L | M)
f. P(S | F)
g. P(F | L)
h. P(F or L)
i. P(M and S)
j. P(F)

Solution to Example 3.2, Problem (p. 68)
No. $C = \{3,\ 5\}$ and $E = \{1,\ 2,\ 3,\ 4\}$. $P\left(C\ AND\ E\right) = \frac{1}{6}$. To be mutually exclusive, $P\left(C\ AND\ E\right)$ must be 0.
Solution to Example 3.11, Problem 1 (p. 74)

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Table 3.4

Solution to Example 3.11, Problem 2 (p. 74)

a. P(F AND C) $= \frac{18}{100} = 0.18$
b. $P\left(F\right)\ \cdot\ P\left(C\right) = \frac{45}{100} \cdot \frac{34}{100} = 0.45\ \cdot\ 0.34 = 0.153$

P(F AND C) $\neq P\left(F\right)\ \cdot\ P\left(C\right)$, so the events F and C are not independent.
Solution to Example 3.11, Problem 3 (p. 75)

a. The word 'given' tells you that this is a conditional.
b. P(M | L) $= \frac{25}{41}$
c. No, the sample space for this problem is 41.

Solution to Example 3.11, Problem 4 (p. 75)

a. P(F) $= \frac{45}{100}$
b. P(P) $= \frac{25}{100}$
c. P(F AND P) $= \frac{11}{100}$
d. P(F OR P) $= \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Solution to Example 3.12, Problem 1 (p. 75)

Door Choice

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Table 3.5

Chapter 4

Discrete Random Variables

4.1 Discrete Random Variables[1]

4.1.1 Student Learning Outcomes

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

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately (optional).
- Recognize the geometric probability distribution and apply it appropriately (optional).
- Recognize the hypergeometric probability distribution and apply it appropriately (optional).
- Classify discrete word problems by their distributions.

4.1.2 Introduction

A student takes a 10 question true-false quiz. Because the student had such a busy schedule, he or she could not study and randomly guesses at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

In this chapter, you will study probability problems involving discrete random distributions. You will also study long-term averages associated with them.

4.1.3 Random Variable Notation

Upper case letters like X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If X **is a random variable, then** X **is written in words.** and x **is given as a number.**

[1] This content is available online at <http://cnx.org/content/m16825/1.14/>.

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is TTT; THH; HTH; HHT; HTT; THT; TTH; HHH. Then, x = 0, 1, 2, 3. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

4.1.4 Optional Collaborative Classroom Activity

Toss a coin 10 times and record the number of heads. After all members of the class have completed the experiment (tossed a coin 10 times and counted the number of heads), fill in the chart using a heading like the one below. Let X = the number of heads in 10 tosses of the coin.

x	**Frequency of x**	**Relative Frequency of x**

Table 4.1

- Which value(s) of x occurred most frequently?
- If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
- What does the relative frequency column sum to?

4.2 Probability Distribution Function (PDF) for a Discrete Random Variable[2]

A discrete **probability distribution function** has two characteristics:

- Each probability is between 0 and 1, inclusive.
- The sum of the probabilities is 1.

Example 4.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times a newborn wakes its mother after midnight. For this example, x = 0, 1, 2, 3, 4, 5.

P(x) = probability that X takes on a value x.

[2] This content is available online at <http://cnx.org/content/m16831/1.14/>.

x	P(x)
0	P(x=0) $= \frac{2}{50}$
1	P(x=1) $= \frac{11}{50}$
2	P(x=2) $= \frac{23}{50}$
3	P(x=3) $= \frac{9}{50}$
4	P(x=4) $= \frac{4}{50}$
5	P(x=5) $= \frac{1}{50}$

Table 4.2

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because

1. Each P(x) is between 0 and 1, inclusive.
2. The sum of the probabilities is 1, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1 \tag{4.1}$$

Example 4.2
Suppose Nancy has classes **3 days** a week. She attends classes 3 days a week **80%** of the time, **2 days 15%** of the time, **1 day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Problem 1 ***(Solution on p. 94.)***
Let X = the number of days Nancy ____________________ .

Problem 2 ***(Solution on p. 94.)***
X takes on what values?

Problem 3 ***(Solution on p. 94.)***
Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in the previous example. The table should have two columns labeled x and P(x). What does the P(x) column sum to?

4.3 Mean or Expected Value and Standard Deviation[3]

The **expected value** is often referred to as the **"long-term"average or mean** . This means that over the long term of doing an experiment over and over, you would **expect** this average.

The **mean** of a random variable X is μ. If we do an experiment many times (for instance, flip a fair coin, as Karl Pearson did, 24,000 times and let X = the number of heads) and record the value of X each time, the average is likely to get closer and closer to μ as we keep repeating the experiment. This is known as the **Law of Large Numbers**.

> NOTE: To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

[3] This content is available online at <http://cnx.org/content/m16828/1.16/>.

A Step-by-Step Example
A men's soccer team plays soccer 0, 1, or 2 days a week. The probability that they play 0 days is 0.2, the probability that they play 1 day is 0.5, and the probability that they play 2 days is 0.3. Find the long-term average, μ, or expected value of the days per week the men's soccer team plays soccer.

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table, adding a column $xP(x)$. In this column, you will multiply each x value by its probability.

Expected Value Table

x	**P(x)**	x**P(x)**
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Table 4.4: This table is called an expected value table. The table helps you calculate the expected value or long-term average.

Add the last column to find the long term average or expected value: $(0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1$.

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$

Example 4.3
Find the expected value for the example about the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times a newborn wakes its mother after midnight.

x	**P(X)**	x**P(X)**
0	P(x=0) $= \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$
1	P(x=1) $= \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$
2	P(x=2) $= \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$
3	P(x=3) $= \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$
4	P(x=4) $= \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$
5	P(x=5) $= \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$

Table 4.5: You expect a newborn to wake its mother after midnight 2.1 times, on the average.

Add the last column to find the expected value. μ = Expected Value = $\frac{105}{50} = 2.1$

Problem
Go back and calculate the expected value for the number of days Nancy attends classes a week. Construct the third column to do so.

Solution
2.74 days a week.

Example 4.4
Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from 0 to 9 with replacement. You pay $2 to play and could profit $100,000 if you match all 5 numbers in order (you get your $2 back plus $100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and -2 dollars.

To win, you must get all 5 numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are 10 numbers. You may choose a number more than once. The probability of choosing all 5 numbers correctly and in order is:

$$\frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \frac{1}{10} * \; = \; 1 * 10^{-5} = 0.00001 \tag{4.2}$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

$$1 - 0.00001 = 0.99999 \tag{4.3}$$

The expected value table is as follows.

	x	P(x)	xP(x)
Loss	-2	0.99999	(-2)(0.99999)=-1.99998
Profit	100,000	0.00001	(100000)(0.00001)=1

Table 4.6: Add the last column. -1.99998 + 1 = -0.99998

Since -0.99998 is about -1, you would, on the average, expect to lose approximately one dollar for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Example 4.5
Suppose you play a game with a biased coin. You play each game by tossing the coin once. P(heads) $= \frac{2}{3}$ and P(tails) $= \frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

Problem 1 ***(Solution on p. 94.)***
Define a random variable X.

Problem 2 ***(Solution on p. 94.)***
Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

Table 4.7

Problem 3 ***(Solution on p. 94.)***

What is the expected value, μ? Do you come out ahead?

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root . To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x-\mu)^2 \cdot P(x)$ and take the square root.

x	P(x)	xP(x)	(x -μ)^{2}P(x)
0	0.2	(0)(0.2) = 0	$(0-1.1)^2(.2) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1-1.1)^2(.5) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2-1.1)^2(.3) = 0.243$

Table 4.8

Add the last column in the table. $0.242 + 0.005 + 0.243 = 0.490$. The standard deviation is the square root of 0.49. $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas that calculate μ and σ.

4.4 Common Discrete Probability Distribution Functions[4]

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

4.5 Binomial[5]

The characteristics of a binomial experiment are:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only 2 possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial. $p+q=1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true - false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose

[4] This content is available online at <http://cnx.org/content/m16821/1.6/>.

[5] This content is available online at <http://cnx.org/content/m16820/1.17/>.

Joe always guesses correctly on any statistics true - false question with probability $p = 0.6$. Then, $q = 0.4$ .This means that for every true - false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable $X =$ the number of successes obtained in the n independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution is $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then $\sigma = \sqrt{npq}$.

Any experiment that has characteristics 2 and 3 and where n = 1 is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example 4.6
At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable is X = the number of students who withdraw from the randomly selected elementary physics class.

Example 4.7
Suppose you play a game that you can only either win or lose. The probability that you win any game is 55% and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, what is the probability that you win 15 of the 20 games? Here, if you define X = the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Example 4.8
A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than 10 heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, p = 0.5 and q = 0.5. The number of trials is n = 15. The probability question can be stated mathematically as $P(x > 10)$.

Example 4.9
Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Problem 1 ***(Solution on p. 94.)***
This is a binomial problem because there is only a success or a __________, there are a definite number of trials, and the probability of a success is 0.70 for each trial.

Problem 2 ***(Solution on p. 94.)***
If we are interested in the number of students who do their homework, then how do we define X?

Problem 3 ***(Solution on p. 94.)***
What values does x take on?

Problem 4 ***(Solution on p. 94.)***
What is a "failure", in words?

The probability of a success is p = 0.70. The number of trial is n = 50.

Problem 5 ***(Solution on p. 94.)***
If $p + q = 1$, then what is q?

Problem 6 ***(Solution on p. 94.)***

The words "at least" translate as what kind of inequality for the probability question $P\left(x\_\_\_\_40\right)$.

The probability question is $P\left(x\geq 40\right)$.

4.5.1 Notation for the Binomial: B = Binomial Probability Distribution Function

$X\sim B\left(n,p\right)$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p. n = number of trials p = probability of a success on each trial

Example 4.10

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where n = 20 and p = 0.41. q = 1 - 0.41 = 0.59. $X\sim B\left(20,0.41\right)$

Find $P\left(x\leq 12\right)$. $P\left(x\leq 12\right)=0.9738$. (calculator or computer)

Using the TI-83+ or the TI-84 calculators, the calculations are as follows. Go into 2nd DISTR. The syntax for the instructions are

To calculate (x = value): binompdf(n, p, number) If "number" is left out, the result is the binomial probability table.

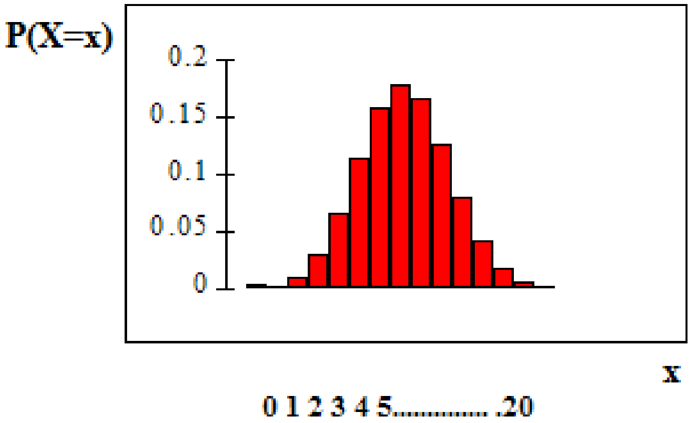
To calculate $P\left(x\leq \textbf{value}\right)$: binomcdf($n$, p, number) If "number" is left out, the result is the cumulative binomial probability table.

For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,.41,12). The result is $P\left(x\leq 12\right)=0.9738$.

NOTE: If you want to find $P\left(x=12\right)$, use the pdf (binompdf). If you want to find P(x>12), use 1 - binomcdf(20,.41,12).

The probability at most 12 workers have a high school diploma but do not pursue any further education is 0.9738

The graph of $x\sim B\left(20,0.41\right)$ is:



The y-axis contains the probability of x, where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)\,(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$. $\sigma = \sqrt{(20)\,(0.41)\,(0.59)} = 2.20$.

Example 4.11
The following example illustrates a problem that is **not** binomial. It violates the condition of independence. ABC College has a student advisory committee made up of 10 staff members and 6 students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? All names of the committee are put into a box and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw produces a student. The probability is $\frac{6}{15}$ when the first draw produces a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

4.6 Poisson[6]

Characteristics of a Poisson experiment:

1. The Poisson gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are 5 words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1000). You will verify the relationship in the homework exercises. n is the number of trials and p is the probability of a "success."

[6] This content is available online at <http://cnx.org/content/m16829/1.16/>.

Poisson probability distribution. The random variable X = the number of occurrences in the interval of interest. The mean and variance are given in the summary.

Example 4.12
The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in 5 minutes. The time interval of interest is 5 minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in 5 minutes is 3?

Let X = the number of loaves of bread put on the shelf in 5 minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, **then the average number of loaves put on the shelf in 5 minutes is**

$\left(\frac{5}{30}\right) \cdot 12 = 2$ loaves of bread

The probability question asks you to find $P(x = 3)$.

Example 4.13
A certain bank expects to receive 6 bad checks per day, on average. What is the probability of the bank getting fewer than 5 bad checks on any given day? Of interest is the number of checks the bank receives in 1 day, so the time interval of interest is 1 day. Let X = the number of bad checks the bank receives in one day. If the bank expects to receive 6 bad checks per day then the average is 6 checks per day. The probability question asks for $P(x < 5)$.

Example 4.14
You notice that a news reporter says "uh", on average, 2 times per broadcast. What is the probability that the news reporter says "uh" more than 2 times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

Problem 1 ***(Solution on p. 94.)***
What is the interval of interest?

Problem 2 ***(Solution on p. 94.)***
What is the average number of times the news reporter says "uh" during one broadcast?

Problem 3 ***(Solution on p. 94.)***
Let X = ____________. What values does X take on?

Problem 4 ***(Solution on p. 95.)***
The probability question is P(______).

4.6.1 Notation for the Poisson: P = Poisson Probability Distribution Function

$X \sim \mathrm{P}(\mu)$

Read this as "X is a random variable with a Poisson distribution." The parameter is μ (or λ). μ (or λ) = the mean for the interval of interest.

Example 4.15
Leah's answering machine receives about 6 telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than 1 call **in the next 15 minutes?**

Let X = the number of calls Leah receives in 15 minutes. (The **interval of interest** is 15 minutes or $\frac{1}{4}$ hour.)

x = 0, 1, 2, 3, ...

If Leah receives, on the average, 6 telephone calls in 2 hours, and there are eight 15 minutes intervals in 2 hours, then Leah receives

$\frac{1}{8} \cdot 6 = 0.75$

calls in 15 minutes, on the average. So, μ = 0.75 for this problem.

$X \sim$ P(0.75)

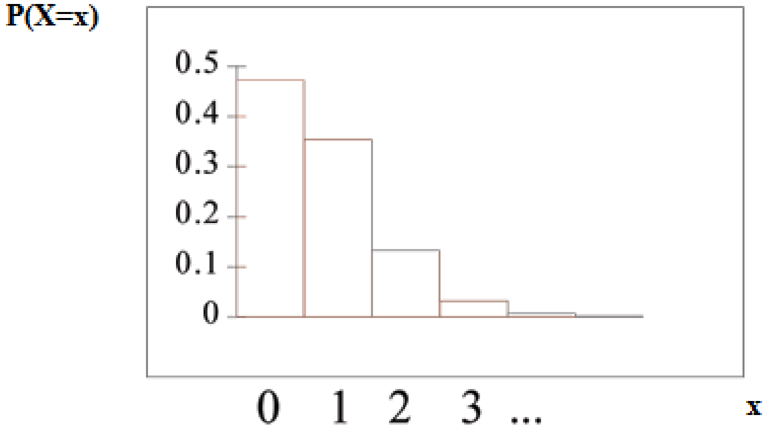
Find $P\,(x > 1)$. $P\,(x > 1) = 0.1734$ (calculator or computer)

TI-83+ and TI-84: For a general discussion, see **this example (Binomial)**. The syntax is similar. The Poisson parameter list is (μ for the interval of interest, number). **For this problem:**

Press 1- and then press 2nd DISTR. Arrow down to C:poissoncdf. Press ENTER. Enter .75,1). The result is $P\,(x > 1) = 0.1734$**. NOTE: The TI calculators use** λ **(lambda) for the mean.**

The probability that Leah receives more than 1 telephone call in the next fifteen minutes is about 0.1734.

The graph of $X \sim$ P(0.75) is:



The y-axis contains the probability of x where X = the number of calls in 15 minutes.

4.7 Summary of Functions[7]

Formula 4.1: Binomial
$X{\sim}B\left(n,p\right)$

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x=0{,}1,2,3,\ldots{,}n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p+q=1 \quad q=1-p$

The mean is $\mu=np$. The standard deviation is $\sigma=\sqrt{npq}$.

Formula 4.2: Geometric
$X{\sim}G\left(p\right)$

X = the number of independent trials until the first success (count the failures and the first success)

X takes on the values x= 1, 2, 3, ...

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p+q=1$

$q=1-p$

The mean is $\mu=\frac{1}{p}$

The standard deviation is $\sigma=\sqrt{\frac{1}{p}\left(\left(\frac{1}{p}\right)-1\right)}$

Formula 4.3: Hypergeometric
$X{\sim}H\left(r,b,n\right)$

X = the number of items from the group of interest that are in the chosen sample.

X may take on the values x= 0, 1, ..., up to the size of the group of interest. (The minimum value for X may be larger than 0 in some instances.)

r = the size of the group of interest (first group)

b= the size of the second group

n= the size of the chosen sample.

$n\leq r+b$

The mean is: $\mu=\frac{nr}{r+b}$

[7] This content is available online at <http://cnx.org/content/m16833/1.11/>.

The standard deviation is: $\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$

Formula 4.4: Poisson
$X \sim \mathrm{P}(\mu)$

X = the number of occurrences in the interval of interest

X takes on the values x = 0, 1, 2, 3, ...

The mean μ is typically given. (λ is often used as the mean instead of μ.) When the Poisson is used to approximate the binomial, we use the binomial mean $\mu = np$. n is the binomial number of trials. p = the probability of a success for each trial. This formula is valid when n is "large" and p "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to 0.05). If n is large enough and p is small enough then the Poisson approximates the binomial very well. The variance is $\sigma^2 = \mu$ and the standard deviation is $\sigma = \sqrt{\mu}$

Solutions to Exercises in Chapter 4

Solution to Example 4.2, Problem 1 (p. 83)
Let X = the number of days Nancy **attends class per week**.
Solution to Example 4.2, Problem 2 (p. 83)
0, 1, 2, and 3
Solution to Example 4.2, Problem 3 (p. 83)

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

Table 4.9

Solution to Example 4.5, Problem 1 (p. 85)
X = amount of profit
Solution to Example 4.5, Problem 2 (p. 85)

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	**-6**	$\frac{2}{3}$	$\frac{-12}{3}$

Table 4.10

Solution to Example 4.5, Problem 3 (p. 86)
Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.
Solution to Example 4.9, Problem 1 (p. 87)
failure
Solution to Example 4.9, Problem 2 (p. 87)
X = the number of statistics students who do their homework on time
Solution to Example 4.9, Problem 3 (p. 87)
0, 1, 2, . . ., 50
Solution to Example 4.9, Problem 4 (p. 87)
Failure is a student who does not do his or her homework on time.
Solution to Example 4.9, Problem 5 (p. 87)
$q = 0.30$
Solution to Example 4.9, Problem 6 (p. 88)
greater than or equal to ($\geq$)
Solution to Example 4.14, Problem 1 (p. 90)
One broadcast
Solution to Example 4.14, Problem 2 (p. 90)
2

Solution to Example 4.14, Problem 3 (p. 90)
Let X = **the number of times the news reporter says "uh" during one broadcast**.
$x = 0, 1, 2, 3, \ldots$
Solution to Example 4.14, Problem 4 (p. 90)
$\mathrm{P}(\mathrm{x} > 2)$

Chapter 5

Continuous Random Variables

5.1 Continuous Random Variables[1]

5.1.1 Student Learning Outcomes

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

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately.

5.1.2 Introduction

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

This chapter gives an introduction to continuous random variables and the many continuous distributions. We will be studying these continuous distributions for several chapters.

> NOTE: The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. How the random variable is defined is very important.

5.1.3 Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated: **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

[1] This content is available online at <http://cnx.org/content/m16808/1.12/>.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated: **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to 1.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is 0. The area below the curve, above the x-axis, and between x=c and x=c has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also 0.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to best model and fit the particular situation.

In this chapter and the next chapter, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

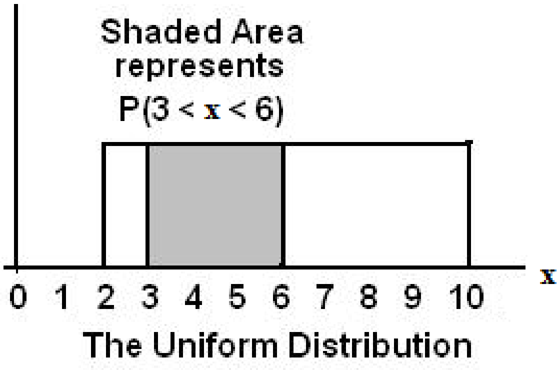


Figure 5.1: The graph shows a Uniform Distribution with the area between x=3 and x=6 shaded to represent the probability that the value of the random variable X is in the interval between 3 and 6.

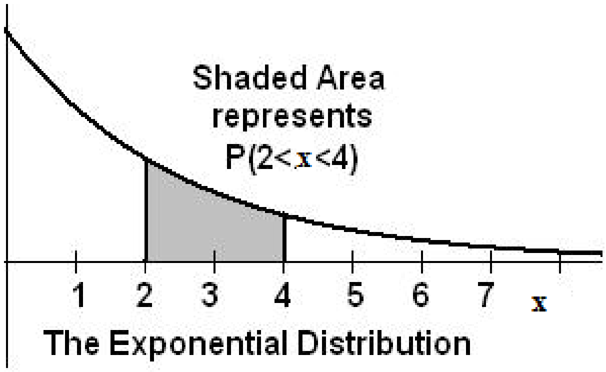


Figure 5.2: The graph shows an Exponential Distribution with the area between x=2 and x=4 shaded to represent the probability that the value of the random variable X is in the interval between 2 and 4.

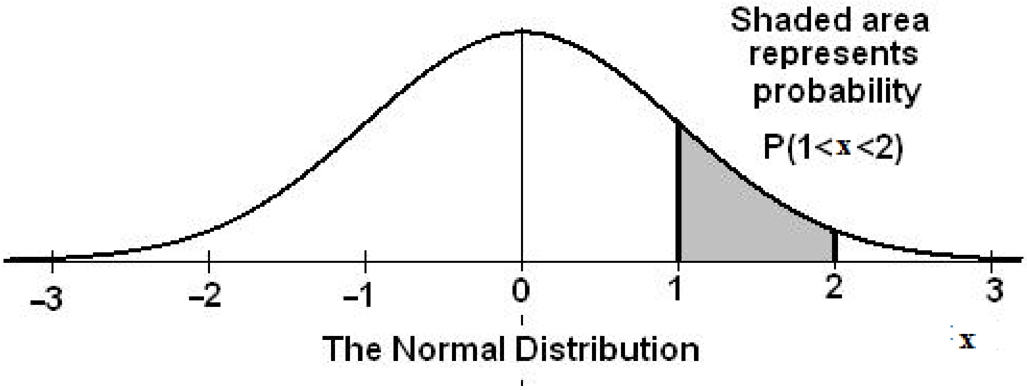


Figure 5.3: The graph shows the Standard Normal Distribution with the area between x=1 and x=2 shaded to represent the probability that the value of the random variable X is in the interval between 1 and 2.

**With contributions from Roberta Bloom

5.2 Continuous Probability Functions[2]

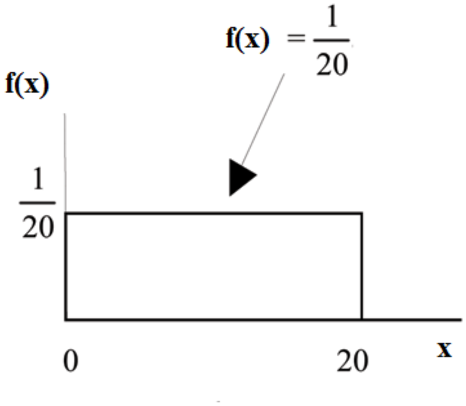
We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one.

For continuous probability distributions, PROBABILITY = AREA.

[2] This content is available online at <http://cnx.org/content/m16805/1.9/>.

Example 5.1

Consider the function $f\left(x\right)=\frac{1}{20}$ for $0\le x\le 20$. x = a real number. The graph of $f\left(x\right)=\frac{1}{20}$ is a horizontal line. However, since $0\le x\le 20$, $f\left(x\right)$ is restricted to the portion between $x=0$ and $x=20$, inclusive .



$f\left(x\right)=\frac{1}{20}$ **for** $0\le x\le 20$.

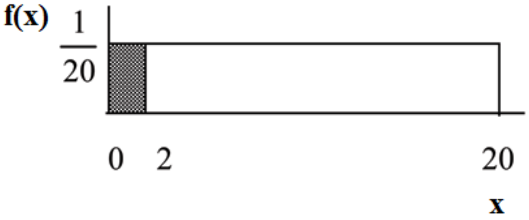
The graph of $f\left(x\right)=\frac{1}{20}$ is a horizontal line segment when $0\le x\le 20$.

The area between $f\left(x\right)=\frac{1}{20}$ where $0\le x\le 20$ and the x-axis is the area of a rectangle with base = 20 and height =$\frac{1}{20}$.

$\mathrm{AREA}=20\cdot\frac{1}{20}=1$

This particular function, where we have restricted x so that the area between the function and the x-axis is 1, is an example of a continuous probability density function. It is used as a tool to calculate probabilities.

Suppose we want to find the area between $f\left(x\right)=\frac{1}{20}$ **and the x-axis where** $0<x<2$ **.**



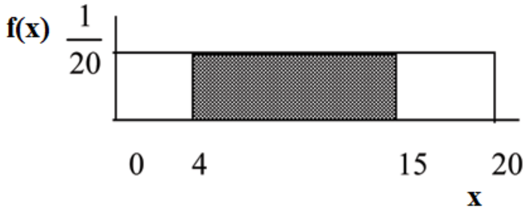
$\mathrm{AREA}=\left(2-0\right)\cdot\frac{1}{20}=0.1$

$\left(2-0\right)=2=$ base of a rectangle

$\frac{1}{20}$ = the height.

The area corresponds to a probability. The probability that x is between 0 and 2 is 0.1, which can be written mathematically as P(0<x<2) $=$ P(x<2) $=0.1$.

Suppose we want to find the area between $f\left(x\right)=\frac{1}{20}$ **and the x-axis where** $4<x<15$ **.**



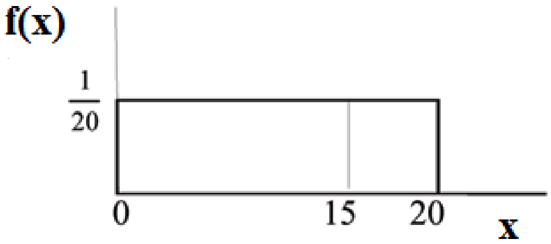
AREA $= (15-4)\cdot\frac{1}{20} = 0.55$

$(15-4) = 11 =$ the base of a rectangle

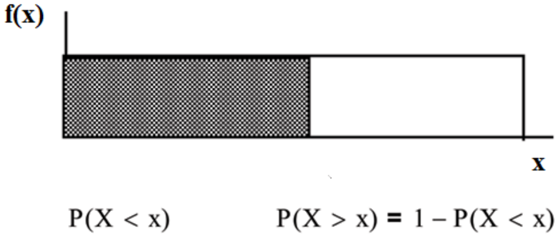
$\frac{1}{20}$ = the height.

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

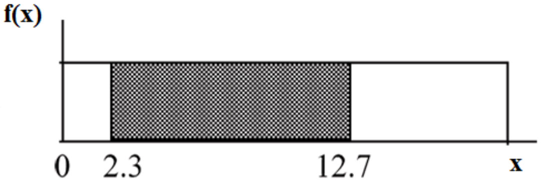
Suppose we want to find $P(x = 15)$**.** On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or 0 width). Therefore, $P(x = 15) =$ (base)(height) $= (0)\left(\frac{1}{20}\right) = 0$.



$P(X \leq x)$ (can be written as $P(X < x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right." We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.



Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}, 0 \leq x \leq 20$.



$P(2.3 < x < 12.7) =$ (base) (height) $= (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$

5.3 The Uniform Distribution[3]

Example 5.2
The previous problem is an example of the **uniform probability distribution**.

Illustrate the uniform distribution. The data that follows are 55 smiling times, in seconds, of an eight-week old baby.

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

Table 5.1

sample mean = 11.49 and sample standard deviation = 6.23

We will assume that the smiling times, in seconds, follow a uniform distribution between 0 and 23 seconds, inclusive. This means that any smiling time from 0 to and including 23 seconds is **equally likely**. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week old baby's smile.

The notation for the uniform distribution is

$X \sim U\left(a,b\right)$ where a = the lowest value of x and b = the highest value of x.

The probability density function is $f\left(x\right)=\frac{1}{b-a}$ for $a\leq x\leq b$.

For this example, $x\sim U\left(0,23\right)$ and $f\left(x\right)=\frac{1}{23-0}$ for $0\leq x\leq 23$.

Formulas for the theoretical mean and standard deviation are

$\mu=\frac{a+b}{2}$ and $\sigma=\sqrt{\frac{(b-a)^2}{12}}$

For this problem, the theoretical mean and standard deviation are

$\mu=\frac{0+23}{2}=11.50$ seconds and $\sigma=\sqrt{\frac{(23-0)^2}{12}}=6.64$ seconds

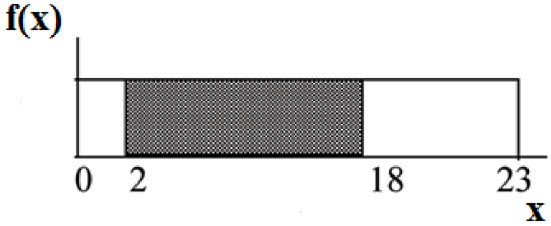
Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation.

Example 5.3
Problem 1
What is the probability that a randomly chosen eight-week old baby smiles between 2 and 18 seconds?

Solution
Find $P\left(2<x<18\right)$.

[3] This content is available online at <http://cnx.org/content/m16819/1.17/>.

$P(2 < x < 18) = (\text{base})(\text{height}) = (18 - 2) \cdot \frac{1}{23} = \frac{16}{23}$.



Problem 2
Find the 90th percentile for an eight week old baby's smiling time.

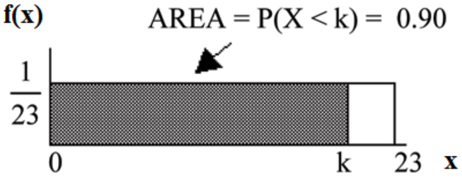
Solution
Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x < k) = 0.90$

$P(x < k) = 0.90$

$(\text{base})(\text{height}) = 0.90$

$(k - 0) \cdot \frac{1}{23} = 0.90$

$k = 23 \cdot 0.90 = 20.7$



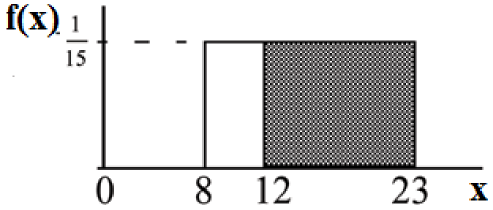
Problem 3
Find the probability that a random eight week old baby smiles more than 12 seconds **KNOWING** that the baby smiles **MORE THAN 8 SECONDS**.

Solution
Find $P(x > 12|x > 8)$ There are two ways to do the problem. **For the first way**, use the fact that this is a **conditional** and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than 8 seconds.

Write a new $f(x)$**:** $f(x) = \frac{1}{23-8} = \frac{1}{15}$

for $8 < x < 23$

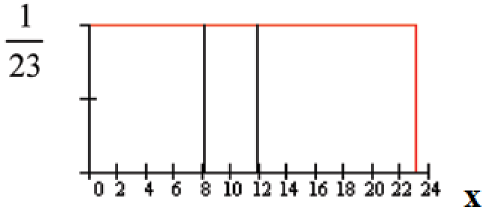
$P(x > 12|x > 8) = (23 - 12) \cdot \frac{1}{15} = \frac{11}{15}$



For the second way, use the conditional formula from **Probability Topics** with the original distribution $X \sim U\,(0,23)$:

$P\,(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ For this problem, A is $(x > 12)$ and B is $(x > 8)$.

So, $P\,(x > 12|x > 8) = \frac{(x>12 \text{ AND } x>8)}{P(x>8)} = \frac{P(x>12)}{P(x>8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = 0.733$



Example 5.4
Uniform: The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between 0 and 15 minutes, inclusive.

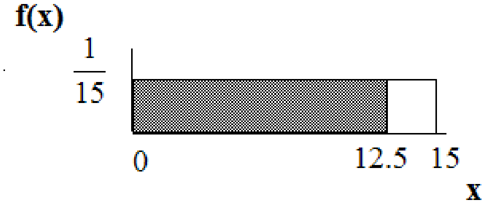
Problem 1
What is the probability that a person waits fewer than 12.5 minutes?

Solution
Let X = the number of minutes a person must wait for a bus. a = 0 and b = 15. $x \sim U\,(0,15)$. Write the probability density function. $f\,(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P\,(x < 12.5)$. Draw a graph.

$P\,(x < k) = (\text{base})\,(\text{height}) = (12.5 - 0) \cdot \frac{1}{15} = 0.8333$

The probability a person waits less than 12.5 minutes is 0.8333.



Problem 2
On the average, how long must a person wait?

Find the mean, μ, and the standard deviation, σ.

Solution
$\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7.5$. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}} = 4.3$. The Standard deviation is 4.3 minutes.

Problem 3
Ninety percent of the time, the time a person must wait falls below what value?

NOTE: This asks for the 90th percentile.

Solution
Find the 90th percentile. Draw a graph. Let k = the 90th percentile.

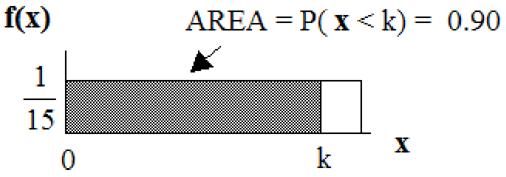
$P(x < k) = (\text{base})(\text{height}) = (k - 0) \cdot \left(\frac{1}{15}\right)$

$0.90 = k \cdot \frac{1}{15}$

$k = (0.90)(15) = 13.5$

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.



Example 5.5
Uniform: Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let X = the time, in minutes, it takes a nine-year old child to eat a donut. Then $X \sim U(0.5, 4)$.

Problem 1 ***(Solution on p. 116.)***
The probability that a randomly selected nine-year old child eats a donut in at least two minutes is ________.

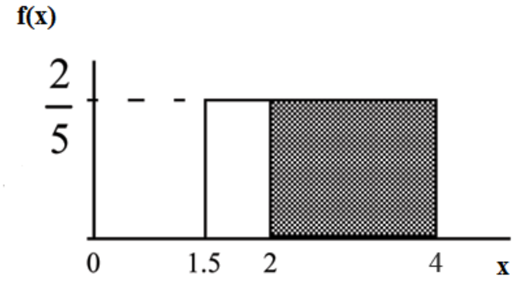
Problem 2 ***(Solution on p. 116.)***
Find the probability that a different nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes.

The second probability question has a **conditional** (refer to "Probability Topics (Section 3.1)"). You are asked to find the probability that a nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways (see the first example (Example 5.2)). You must reduce the sample space. **First way**: Since you already know the child has already been eating the donut for more than 1.5 minutes, you are no longer starting at $a = 0.5$ minutes. Your starting point is 1.5 minutes.

Write a new f(x):

$f(x) = \frac{1}{4-1.5} = \frac{2}{5}$ for $1.5 \leq x \leq 4$.

Find $P(x > 2|x > 1.5)$. Draw a graph.



$P(x > 2|x > 1.5) = \text{(base)}\ \text{(new height)} = (4-2)(2/5) = ?$

The probability that a nine-year old child eats a donut in more than 2 minutes given that the child has already been eating the donut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $x \sim U(0.5, 4)$. Use the conditional formula

$P(x > 2|x > 1.5) = \frac{P(x>2 \text{ AND } x>1.5)}{P(x>1.5)} = \frac{P(x>2)}{P(x>1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$

NOTE: See "Summary of the Uniform and Exponential Probability Distributions (Section 5.5)" for a full summary.

Example 5.6
Uniform: Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and 4 hours. Let x = the time needed to fix a furnace. Then $x \sim U(1.5, 4)$.

1. Find the problem that a randomly selected furnace repair requires more than 2 hours.
2. Find the probability that a randomly selected furnace repair requires less than 3 hours.
3. Find the 30th percentile of furnace repair times.
4. The longest 25% of repair furnace repairs take at least how long? (In other words: Find the minimum time for the longest 25% of repair times.) What percentile does this represent?
5. Find the mean and standard deviation

Problem 1
Find the probability that a randomly selected furnace repair requires longer than 2 hours.

Solution
To find $f(x)$: $f(x) = \frac{1}{4-1.5} = \frac{1}{2.5}$ so $f(x) = 0.4$

P(x>2) = (base)(height) = (4 − 2)(0.4) = 0.8

Example 4 Figure 1

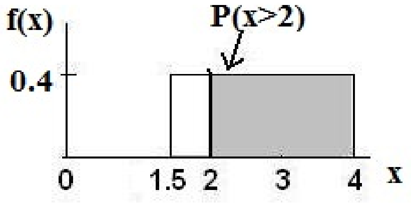


Figure 5.4: Uniform Distribution between 1.5 and 4 with shaded area between 2 and 4 representing the probability that the repair time x is greater than 2

Problem 2
Find the probability that a randomly selected furnace repair requires less than 3 hours. Describe how the graph differs from the graph in the first part of this example.

Solution
$P(x < 3)$ = (base)(height) = (3 − 1.5)(0.4) = 0.6

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between x=1.5 and x=3. Note that the shaded area starts at x=1.5 rather than at x=0; since X∼U(1.5,4), x can not be less than 1.5.

Example 4 Figure 2

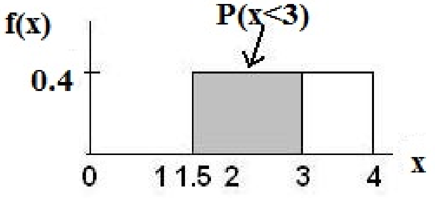


Figure 5.5: Uniform Distribution between 1.5 and 4 with shaded area between 1.5 and 3 representing the probability that the repair time x is less than 3

Problem 3
Find the 30th percentile of furnace repair times.

Solution

Example 4 Figure 3

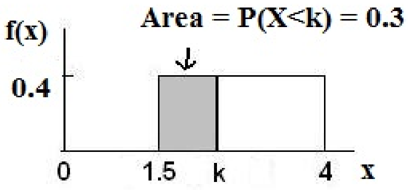


Figure 5.6: Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

$P\left(x<k\right)=0.30$

$P\left(x<k\right)=\left(\text{base}\right)\left(\text{height}\right)=\left(k-1.5\right)\cdot\left(0.4\right)$

0.3 = (k − 1.5) (0.4) ; Solve to find k:
0.75 = k − 1.5 , obtained by dividing both sides by 0.4
k = 2.25 , obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.5 hours or less.

Problem 4
The **longest 25%** of furnace repair times take **at least** how long? (Find the minimum time for the longest 25% of repairs.)

Solution

Example 4 Figure 4

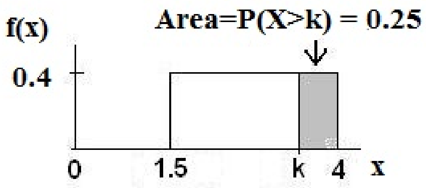


Figure 5.7: Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.

$P\left(x>k\right)=0.25$

$P\left(x>k\right)=\left(\text{base}\right)\left(\text{height}\right)=\left(4-k\right)\cdot\left(0.4\right)$

0.25 = (4 − k)(0.4) ; Solve for k:

0.625 = 4 − k , obtained by dividing both sides by 0.4
−3.375 = −k , obtained by subtracting 4 from both sides
k=3.375

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the **75th percentile** of furnace repair times.

Problem 5
Find the mean and standard deviation

Solution
$\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$\mu = \frac{1.5+4}{2} = 2.75$ hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

NOTE: See "Summary of the Uniform and Exponential Probability Distributions (Section 5.5)" for a full summary.

**Example 5 contributed by Roberta Bloom

5.4 The Exponential Distribution[4]

The **exponential** distribution is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people that spend less money and fewer people that spend large amounts of money.

The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

Example 5.7
Illustrates the exponential distribution: Let X = amount of time (in minutes) a postal clerk spends with his/her customer. The time is known to have an exponential distribution with the average amount of time equal to 4 minutes.

X is a **continuous random variable** since time is measured. It is given that μ = 4 minutes. To do any calculations, you must know m, the decay parameter.

$m = \frac{1}{\mu}$. Therefore, $m = \frac{1}{4} = 0.25$

[4] This content is available online at <http://cnx.org/content/m16816/1.15/>.

The standard deviation, σ, is the same as the mean. $\mu = \sigma$

The distribution notation is X~Exp (m). Therefore, X~Exp (0.25).

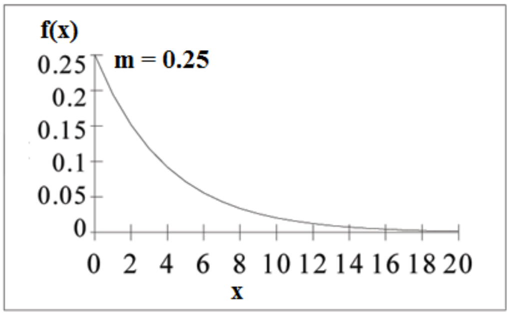
The probability density function is $f\left(x\right) = m \cdot e^{-m \cdot x}$ The number e = 2.71828182846... It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter 1 for x, the calculator will display the value e.

The curve is:

$f\left(x\right) = 0.25 \cdot e^{-\ 0.25 \cdot x}$ where x is at least 0 and m = 0.25.

For example, $f\left(5\right) = 0.25 \cdot e^{-\ 0.25 \cdot 5} = 0.072$

The graph is as follows:



μ = 4

Notice the graph is a declining curve. When x = 0,

$f\left(x\right) = 0.25 \cdot e^{-\ 0.25 \cdot 0} = 0.25 \cdot 1 = 0.25 = m$

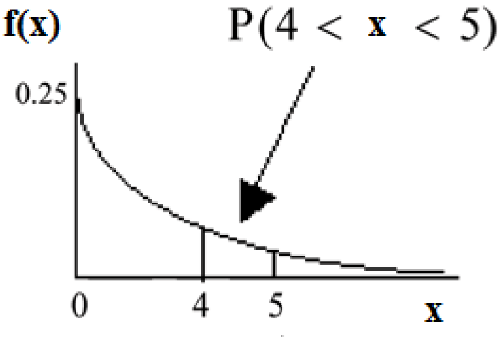
Example 5.8
Problem 1
Find the probability that a clerk spends four to five minutes with a randomly selected customer.

Solution
Find $P\left(4 < x < 5\right)$.

The **cumulative distribution function (CDF)** gives the area to the left.

$P\left(x < x\right) = 1 - e^{-m \cdot x}$

$P\left(x < 5\right) = 1 - e^{-0.25 \cdot 5} = 0.7135$ and $P\left(x < 4\right) = 1 - e^{-0.25 \cdot 4} = 0.6321$



NOTE: You can do these calculations easily on a calculator.

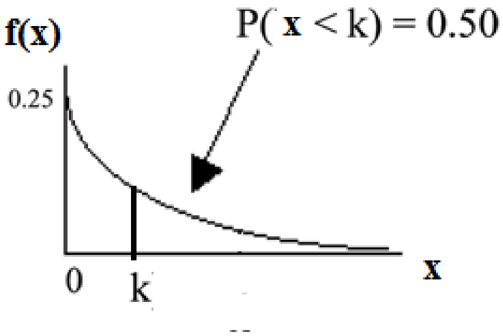
The probability that a postal clerk spends four to five minutes with a randomly selected customer is

$P(4 < x < 5) = P(x < 5) - P(x < 4) = 0.7135 - 0.6321 = 0.0814$

NOTE: TI-83+ and TI-84: On the home screen, enter (1-e^(-.25*5))-(1-e^(-.25*4)) or enter e^(-.25*4)-e^(-.25*5).

Problem 2
Half of all customers are finished within how long? (Find the 50th percentile)

Solution
Find the 50th percentile.



$P(x < k) = 0.50$, k = 2.8 minutes (calculator or computer)

Half of all customers are finished within 2.8 minutes.

You can also do the calculation as follows:

$P(x < k) = 0.50$ and $P(x < k) = 1 - e^{-0.25 \cdot k}$

Therefore, $0.50 = 1 - e^{-0.25 \cdot k}$ and $e^{-0.25 \cdot k} = 1 - 0.50 = 0.5$

Take natural logs: $ln\left(e^{-0.25 \cdot k}\right) = ln\,(0.50)$. So, $-0.25 \cdot k = ln\,(0.50)$

Solve for k: $k = \frac{ln(.50)}{-0.25} = 2.8$ minutes

NOTE: A formula for the percentile k is $k = \frac{LN(1-AreaToTheLeft)}{-m}$ where LN is the natural log.

NOTE: TI-83+ and TI-84: On the home screen, enter LN(1-.50)/-.25. Press the (-) for the negative.

Problem 3
Which is larger, the mean or the median?

Solution
Is the mean or median larger?

From part b, the median or 50th percentile is 2.8 minutes. The theoretical mean is 4 minutes. The mean is larger.

5.4.1 Optional Collaborative Classroom Activity

Have each class member count the change he/she has in his/her pocket or purse. Your instructor will record the amounts in dollars and cents. Construct a histogram of the data taken by the class. Use 5 intervals. Draw a smooth curve through the bars. The graph should look approximately exponential. Then calculate the mean.

Let X = the amount of money a student in your class has in his/her pocket or purse.

The distribution for X is approximately exponential with mean, μ = _______ and m = _______. The standard deviation, σ = ________.

Draw the appropriate exponential graph. You should label the x and y axes, the decay rate, and the mean. Shade the area that represents the probability that one student has less than $.40 in his/her pocket or purse. (Shade $P(x < 0.40)$).

Example 5.9
On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.

Problem 1
What is the probability that a computer part lasts more than 7 years?

Solution
Let x = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

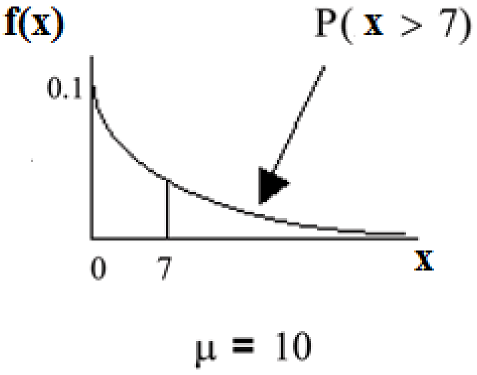
Find $P(x > 7)$. Draw a graph.

$P(x > 7) = 1 - P(x < 7)$.

Since $P(X < x) = 1 - e^{-mx}$ then $P(X > x) = 1 - (1 - e^{-m \cdot x}) = e^{-m \cdot x}$

$P(x > 7) = e^{-0.1 \cdot 7} = 0.4966$. The probability that a computer part lasts more than 7 years is 0.4966.

NOTE: TI-83+ and TI-84: On the home screen, enter e^(-.1*7).



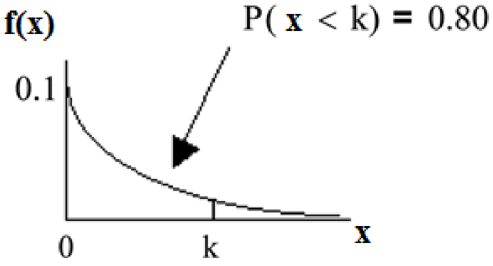
Problem 2
On the average, how long would 5 computer parts last if they are used one after another?

Solution
On the average, 1 computer part lasts 10 years. Therefore, 5 computer parts, if they are used one right after the other would last, on the average,

$(5)\,(10) = 50$ years.

Problem 3
Eighty percent of computer parts last at most how long?

Solution
Find the 80th percentile. Draw a graph. Let k = the 80th percentile.



Solve for k: $k = \frac{ln(1-.80)}{-0.1} = 16.1$ years

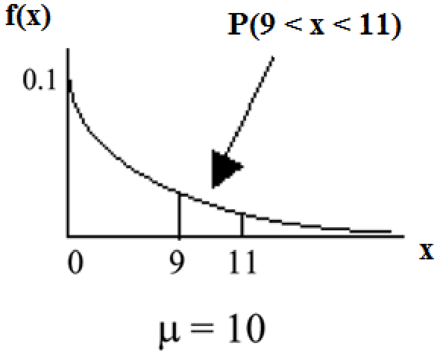
Eighty percent of the computer parts last at most 16.1 years.

NOTE: TI-83+ and TI-84: On the home screen, enter LN(1 - .80)/-.1

Problem 4
What is the probability that a computer part lasts between 9 and 11 years?

Solution

Find $P(9 < x < 11)$. Draw a graph.



$P(9 < x < 11) = P(x < 11) - P(x < 9) = \left(1 - e^{-0.1 \cdot 11}\right) - \left(1 - e^{-0.1 \cdot 9}\right) = 0.6671 - 0.5934 =$ 0.0737. (calculator or computer)

The probability that a computer part lasts between 9 and 11 years is 0.0737.

NOTE: TI-83+ and TI-84: On the home screen, enter e^(-.1*9) - e^(-.1*11).

Example 5.10

Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter = $\frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than 5 minutes. Let X = the length of a phone call, in minutes.

Problem *(Solution on p. 116.)*

What is m, μ, and σ? The probability that you must wait more than 5 minutes is _______ .

NOTE: A summary for exponential distribution is available in "Summary of The Uniform and Exponential Probability Distributions (Section 5.5)".

5.5 Summary of the Uniform and Exponential Probability Distributions[5]

Formula 5.1: Uniform
X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X ; b = largest X

$X \sim U\left(a{,}b\right)$

The mean is $\mu = \frac{a+b}{2}$

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f\left(X\right) = \frac{1}{b-a}$ for $a \le X \le b$

Area to the Left of x: $P\left(X < x\right) =$ (base)(height)

Area to the Right of x: $P\left(X > x\right) =$ (base)(height)

Area Between c and d: $P\left(c < X < d\right) =$ (base) (height) $= \left(d - c\right)$ (height).

Formula 5.2: Exponential
$X \sim Exp\left(m\right)$

X = a real number, 0 or larger. m = the parameter that controls the rate of decay or decline

The mean and standard deviation **are the same.**

$\mu = \sigma = \frac{1}{m}$ and $m = \frac{1}{\mu} = \frac{1}{\sigma}$

The probability density function: $f\left(X\right) = m \cdot e^{-m \cdot X}$, $X \ge 0$

Area to the Left of x: $P\left(X < x\right) = 1 - e^{-m \cdot x}$

Area to the Right of x: $P\left(X > x\right) = e^{-m \cdot x}$

Area Between c and d: $P\left(c < X < d\right) = P\left(X < d\right) - P\left(X < c\right) = \left(1 - e^{-\ m \cdot d}\right) - \left(1 - e^{-\ m \cdot c}\right) = e^{-\ m \cdot c} - e^{-\ m \cdot d}$

Percentile, k: $k = \frac{\text{LN(1-AreaToTheLeft)}}{-m}$

[5] This content is available online at <http://cnx.org/content/m16813/1.10/>.

Solutions to Exercises in Chapter 5

Solution to Example 5.5, Problem 1 (p. 105)
0.5714

Solution to Example 5.5, Problem 2 (p. 105)
$\frac{4}{5}$

Solution to Example 5.10, Problem (p. 114)

- $m = \frac{1}{12}$
- $\mu = 12$
- $\sigma = 12$

$P\left(x > 5\right) = 0.6592$

Chapter 6

The Normal Distribution

6.1 The Normal Distribution[1]

6.1.1 Student Learning Outcomes

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.

6.1.2 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 fit 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.

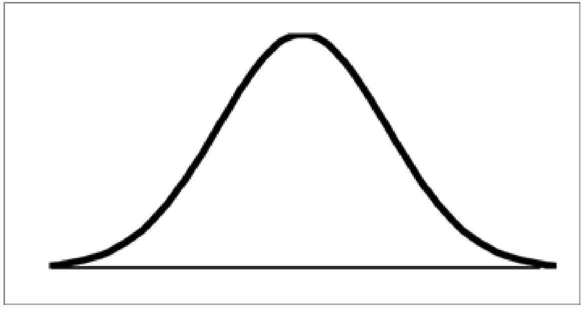
6.1.3 Optional Collaborative Classroom Activity

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?

[1] This content is available online at <http://cnx.org/content/m16979/1.12/>.

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:$X{\sim}N\left(\mu,\ \sigma\right)$



The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

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

The cumulative distribution function is $P\left(X < x\right)$. 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 infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

6.2 The Standard Normal Distribution[2]

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:

$$x = \mu + \left(z\right)\sigma = 5 + \left(3\right)\left(2\right) = 11 \tag{6.1}$$

The z-score is 3.

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

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

[2] This content is available online at <http://cnx.org/content/m16986/1.7/>.

6.3 Z-scores[3]

If X is a normally distributed random variable and $X{\sim}N\,(\mu,\ \sigma)$, then the z-score is:

$$z = \frac{x-\mu}{\sigma} \tag{6.2}$$

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 6.1
Suppose $X \sim N\,(5,\ 6)$. This says that X is a normally distributed random variable with mean $\mu\ =\ 5$ and standard deviation $\sigma\ =\ 6$. Suppose $x\ =\ 17$. Then:

$$z = \frac{x-\mu}{\sigma} = \frac{17-5}{6} = 2 \tag{6.3}$$

This means that $x\ =\ 17$ is **2 standard deviations** (2σ) above or to the right of the mean $\mu\ =\ 5$. The standard deviation is $\sigma\ =\ 6$.

Notice that:

$$5 + 2 \cdot 6 = 17 \qquad \text{(The pattern is } \ \mu + z\sigma = x.\text{)} \tag{6.4}$$

Now suppose $x = 1$. Then:

$$z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67 \qquad \text{(rounded to two decimal places)} \tag{6.5}$$

This means that $x\ =\ 1$ is 0.67 standard deviations $(-\ 0.67\sigma)$ **below or to the left of the mean $\mu\ =\ 5$. Notice that:**

$5 + (-0.67)\,(6)$ is approximately equal to 1 (This has the pattern $\mu + (-0.67)\,\sigma = 1$)

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 6.2
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. $X{\sim}N\,(5,\ 2)$. Fill in the blanks.

Problem 1 ***(Solution on p. 126.)***
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 ***(Solution on p. 126.)***
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: $X \sim N\,(5,\ 6)$ and $Y\ \sim\ N\,(2,\ 1)$. If $x\ =\ 17$, then $z\ =\ 2$. (This was previously shown.) If $y\ =\ 4$, what is z?

$$z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2 \qquad \text{where } \mu\text{=2 and } \sigma\text{=1.} \tag{6.6}$$

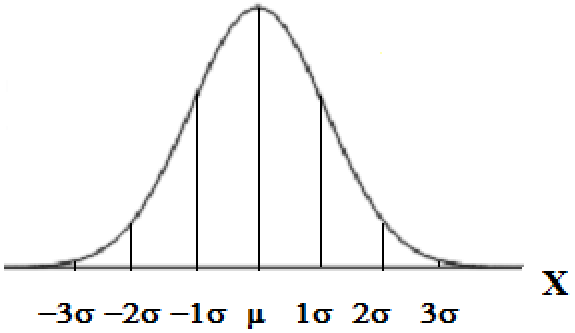
[3] This content is available online at <http://cnx.org/content/m16991/1.10/>.

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 $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a 6 week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each 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 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.



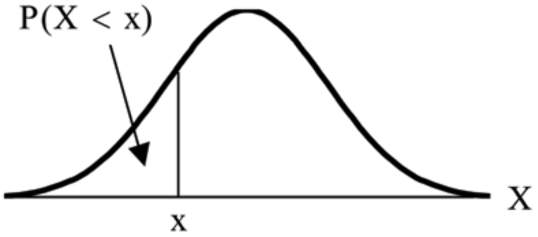
The Empirical Rule is also known as the 68-95-99.7 Rule.

Example 6.3
Suppose X has a normal distribution with mean 50 and standard deviation 6.

- About 68.27% of the x values lie between -1σ = (-1)(6) = -6 and 1σ = (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 -2σ = (-2)(6) = -12 and 2σ = (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 -3σ = (-3)(6) = -18 and 3σ = (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.

6.4 Areas to the Left and Right of x[4]

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 $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

6.5 Calculations of Probabilities[5]

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 6.4
If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Example 6.5
The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

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 $\mu = 63$ and $\sigma = 5$

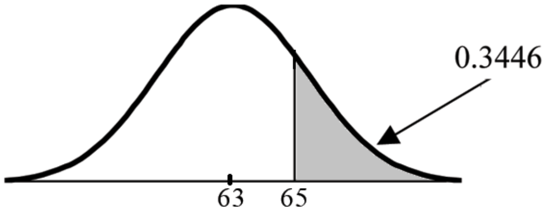
Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$ (calculator or computer)

[4] This content is available online at <http://cnx.org/content/m16976/1.5/>.

[5] This content is available online at <http://cnx.org/content/m16977/1.12/>.



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: normalcdf(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 10^{99} 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:

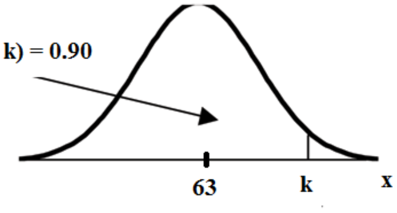
normalcdf(0,85,63,5) = 1 (rounds to 1)

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

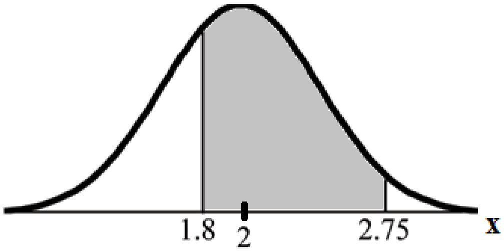
Example 6.6
A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

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 $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

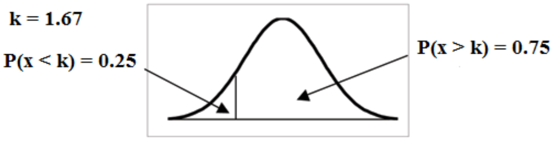


normalcdf(1.8,2.75,2,0.5) = 0.5886

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.

6.6 Summary of Formulas[6]

Formula 6.1: Normal Probability Distribution
$X{\sim}N\left(\mu,\sigma\right)$

μ = the mean σ = the standard deviation

Formula 6.2: Standard Normal Probability Distribution
$Z{\sim}N\left(0,1\right)$

z = a standardized value (z-score)

mean = 0 standard deviation = 1

Formula 6.3: Finding the kth Percentile
To find the **kth** percentile when the z-score is known: $k=\mu+\left(z\right)\sigma$

Formula 6.4: z-score
$z=\frac{x-\mu}{\sigma}$

Formula 6.5: Finding the area to the left
The area to the left: $P\left(X<x\right)$

Formula 6.6: Finding the area to the right
The area to the right: $P\left(X>x\right)=1-P\left(X<x\right)$

[6] This content is available online at <http://cnx.org/content/m16987/1.5/>.

Solutions to Exercises in Chapter 6

Solution to Example 6.2, Problem 1 (p. 119)
This z-score tells you that $x = 10$ is **2.5** standard deviations to the **right** of the mean **5**.
Solution to Example 6.2, Problem 2 (p. 119)
z = **-4**. This z-score tells you that $x = -3$ is **4** standard deviations to the **left** of the mean.

Chapter 7

The Central Limit Theorem

7.1 The Central Limit Theorem[1]

7.1.1 Student Learning Outcomes

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

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Means.
- Apply and interpret the Central Limit Theorem for Sums.

7.1.2 Introduction

Why are we so concerned with means? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study means and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n and n is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, n, that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more observations are needed for the sample means or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

[1] This content is available online at <http://cnx.org/content/m16953/1.17/>.

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times.

Each time a person rolls more than one die, he/she calculates the sample **mean** of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one mean when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more means for a total of 10 means.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "mean" when you roll one die, is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using 2 dice. Do the sample means show any kind of pattern?

Draw the graph for the means using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the means using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

7.2 The Central Limit Theorem for Sample Means (Averages)[2]

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:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

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

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

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.

[2] This content is available online at <http://cnx.org/content/m16947/1.23/>.

To put it more formally, if you draw random samples of size n,the distribution of the random variable $\overline{X}$, 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 $\overline{X}$ has a different z-score associated with it than the random variable X. $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \tag{7.1}$$

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

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

Example 7.1
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.

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**.

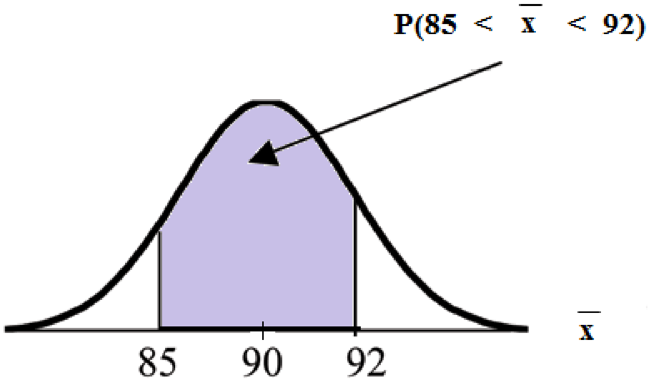
Let $\overline{X}$ = the mean of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$;

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

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

$P(85 < \overline{x} < 92) = 0.6997$

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, $\frac{\sigma}{\sqrt{n}}$)

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

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

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

Example 7.2
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.

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 $\overline{X}$ = the **mean** time, in hours, it takes to play one soccer match.

If μ_X = ________, σ_X = _________, and n = __________, then $\overline{X} \sim N$(______, ______) by the Central Limit Theorem for Means.

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

Find $P\left(1.8 < \overline{x} < 2.3\right)$. Draw a graph.

$P\left(1.8 < \overline{x} < 2.3\right) = 0.9977$

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

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

7.3 The Central Limit Theorem for Sums[3]

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

a. $\mu_X =$ the mean of X
b. $\sigma_X =$ the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

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

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution) which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.
b. $z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$

a. $n \cdot \mu_X =$ the mean of ΣX
b. $\sqrt{n} \cdot \sigma_X =$ standard deviation of ΣX

Example 7.3
An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

Problem

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
b. Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution
Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values.**

ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then

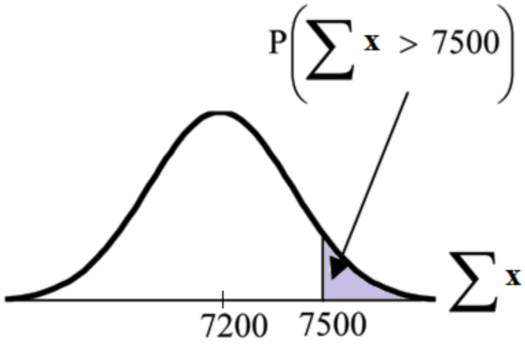
$\Sigma X \sim N\left(80 \cdot 90, \sqrt{80} \cdot 15\right)$

- . mean of the sums = $n \cdot \mu_X = (80)(90) = 7200$
- . standard deviation of the sums = $\sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
- . sum of 80 values = $\Sigma x = 7500$

a: Find $P(\Sigma x > 7500)$

[3] This content is available online at <http://cnx.org/content/m16948/1.16/>.

$P\left(\Sigma x > 7500\right) = 0.0127$



`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated (lower, upper, $n \cdot \mu_X$, $\sqrt{n} \cdot \sigma_X$)

`normalcdf`(7500,1E99, $80 \cdot 90$, $\sqrt{80} \cdot 15 = 0.0127$

Reminder: $1E99 = 10^{99}$. Press the `EE` key for E.

b: Find Σx where $z = 1.5$:

$\Sigma x = n \cdot \mu_X + z \cdot \sqrt{n} \cdot \sigma_X = (80)(90) + (1.5)(\sqrt{80})\ (15) = 7401.2$

7.4 Using the Central Limit Theorem[4]

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.**

7.4.1 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 $\overline{x}$ 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 $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ 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

[4] This content is available online at <http://cnx.org/content/m16958/1.21/>.

Example 7.4
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, $X \sim U(1,5)$ where $a = 1$ and $b = 5$ (See Continuous Random Variables (Section 5.1) for the uniform).

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

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

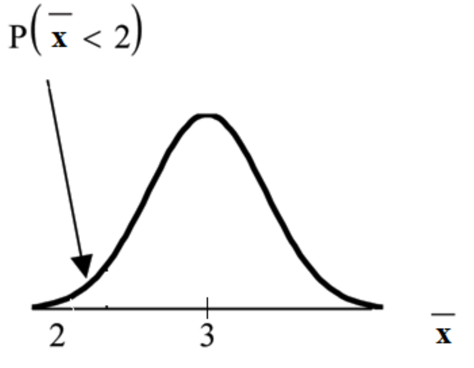
For problems 1. and 2., let $\overline{X}$ = the mean stress score for the 75 students. Then,

$\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where $n = 75$.

Problem 1
Find $P(\overline{x} < 2)$. Draw the graph.

Solution
$P(\overline{x} < 2) = 0$

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



`normalcdf` $\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

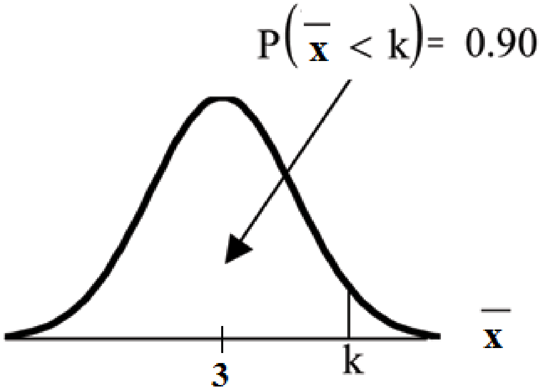
REMINDER: The smallest stress score is 1. Therefore, the smallest mean for 75 stress scores is 1.

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

$k = 3.2$



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` $\left(.90, 3, \frac{1.15}{\sqrt{75}}\right) = 3.2$

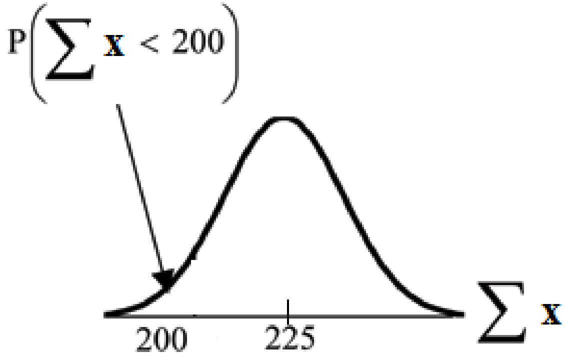
For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N\left[(75)\cdot(3), \sqrt{75}\cdot 1.15\right]$

Problem 3
Find $P\left(\Sigma x < 200\right)$. Draw the graph.

Solution
The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75} \cdot 1.15 = 9.96$

$P\left(\Sigma x < 200\right) = 0$



The probability that the total of 75 scores is less than 200 is about 0.

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

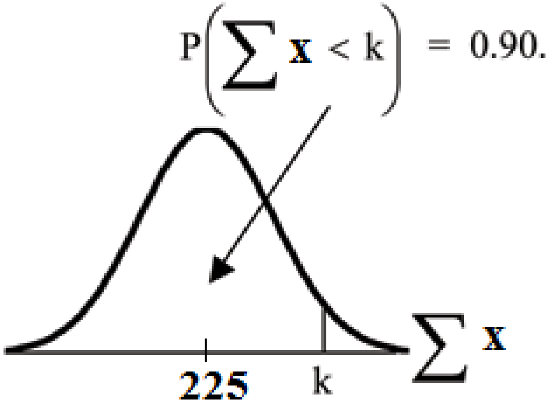
REMINDER: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

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 $P\left(\Sigma x < k\right) = 0.90$.

$k = 237.8$



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.

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

Example 7.5
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 $\mu = 22$ and $\sigma = 22$.

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

$\overline{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means

Problem 1

Using the CLT to find Probability:

a. 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.

b. 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 $P(x > 20)$

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

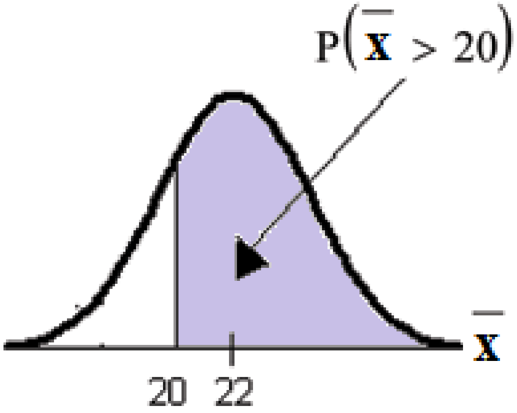
Solution

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.



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^(–(1/22)*20) or e^(–.04545*20) = 0.4029

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

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

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.

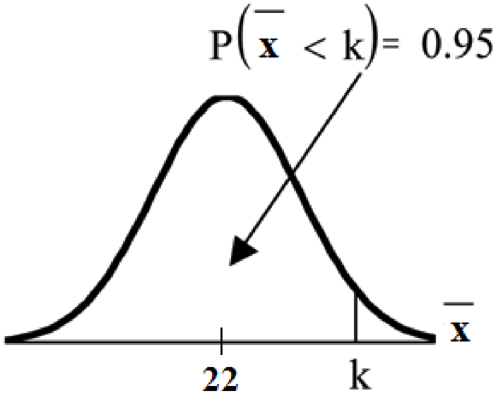
Problem 2

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 $P\left(\overline{x} < k\right) = 0.95$

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



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** simplified 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 $X{\sim}B\left(n, p\right)$. 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 $\mu = np$ 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 7.6
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. $X{\sim}B\left(n,p\right)$ 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 $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N\left(159, 8.6447\right)$. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so $P\left(x \geq 150\right)$ has normal approximation $P\left(Y \geq 149.5\right) = 0.8641$.

`normalcdf` $\left(149.5, 10\hat{\ }99, 159, 8.6447\right) = 0.8641$.

For Problem 2., you **include 160** so $P\left(x \leq 160\right)$ has normal approximation $P\left(Y \leq 160.5\right) = 0.5689$.

`normalcdf` $\left(0, 160.5, 159, 8.6447\right) = 0.5689$

For Problem 3., you **exclude 155** so $P\left(x > 155\right)$ has normal approximation $P\left(y > 155.5\right) = 0.6572$.

`normalcdf` $\left(155.5, 10\hat{\ }99, 159, 8.6447\right) = 0.6572$

For Problem 4., you **exclude 147** so $P\left(x < 147\right)$ has normal approximation $P\left(Y < 146.5\right) = 0.0741$.

`normalcdf` $\left(0, 146.5, 159, 8.6447\right) = 0.0741$

For Problem 5., $P\left(x = 175\right)$ has normal approximation $P\left(174.5 < y < 175.5\right) = 0.0083$.

`normalcdf` $\left(174.5, 175.5, 159, 8.6447\right) = 0.0083$

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.

$P\left(x \geq 150\right)$: `1 - binomialcdf` $\left(300, 0.53, 149\right) = 0.8641$

$P\left(x \leq 160\right)$: `binomialcdf` $\left(300, 0.53, 160\right) = 0.5684$

$P\left(x > 155\right)$: `1 - binomialcdf` $\left(300, 0.53, 155\right) = 0.6576$

$P\left(x < 147\right)$: `binomialcdf` $\left(300, 0.53, 146\right) = 0.0742$

$P\left(x = 175\right)$: (You use the binomial pdf.) `binomialpdf` $\left(175, 0.53, 146\right) = 0.0083$

**Contributions made to Example 2 by Roberta Bloom

7.5 Summary of Formulas[5]

Formula 7.1: Central Limit Theorem for Sample Means

$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ **The Mean** $\left(\overline{X}\right)$**:** μ_X

Formula 7.2: Central Limit Theorem for Sample Means Z-Score and Standard Error of the Mean

$z = \frac{\overline{x}-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$ **Standard Error of the Mean (Standard Deviation** $\left(\overline{X}\right)$**):** $\frac{\sigma_X}{\sqrt{n}}$

Formula 7.3: Central Limit Theorem for Sums

$\Sigma X \sim N\left[(n) \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right]$ **Mean for Sums** (ΣX)**:** $n \cdot \mu_X$

Formula 7.4: Central Limit Theorem for Sums Z-Score and Standard Deviation for Sums

$z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$ **Standard Deviation for Sums** (ΣX)**:** $\sqrt{n} \cdot \sigma_X$

[5] This content is available online at <http://cnx.org/content/m16956/1.8/>.

Chapter 8

Hypothesis Testing: Single Mean and Single Proportion

8.1 Hypothesis Testing: Single Mean and Single Proportion[1]

8.1.1 Student Learning Outcomes

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

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

8.1.2 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 sufficient 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] This content is available online at <http://cnx.org/content/m16997/1.11/>.

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".

8.2 Null and Alternate Hypotheses[2]

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

H_o: **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_o and what we conclude when we reject H_o.

Example 8.1
H_o: 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 8.2
We want to test whether the mean grade point average in American colleges is different from 2.0 (out of 4.0).

H_o: $\mu = 2.0$ H_a: $\mu \neq 2.0$

Example 8.3
We want to test if college students take less than five years to graduate from college, on the average.

H_o: $\mu \geq 5$ H_a: $\mu < 5$

Example 8.4
In an issue of **U. S. News and World Report**, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U. S. students take advanced placement exams and 4.4 % pass. Test if the percentage of U. S. students who take advanced placement exams is more than 6.6%.

H_o: p= 0.066 H_a: $p > 0.066$

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 H_o" or "fail to reject H_o" if the sample information is insufficient to reject the null hypothesis.

[2] This content is available online at <http://cnx.org/content/m16998/1.14/>.

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 (>)

Table 8.1

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 symbol in the Alternate Hypothesis. This practice is acceptable because we only make the decision to reject or not reject the Null Hypothesis.

8.2.1 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.

8.3 Outcomes and the Type I and Type II Errors[3]

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_o **IS ACTUALLY**	**...**
	True	False
Do not reject H_o	Correct Outcome	Type II error
Reject H_o	Type I Error	Correct Outcome

Table 8.2

The four possible outcomes in the table are:

- The decision is to **not reject** H_o when, in fact, H_o **is true (correct decision).**
- The decision is to **reject** H_o when, in fact, H_o **is true** (incorrect decision known as a **Type I error**).
- The decision is to **not reject** H_o when, in fact, H_o **is false** (incorrect decision known as a **Type II error**).
- The decision is to **reject** H_o when, in fact, H_o **is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

[3] This content is available online at <http://cnx.org/content/m17006/1.8/>.

β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely 0.

The Power of the Test is $1 - \beta$. Ideally, we want a high power that is as close to 1 as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example 8.5
Suppose the null hypothesis, H_o, is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α **= probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β **= probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Example 8.6
Suppose the null hypothesis, H_o, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α **= probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = P(Type I error). β **= probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = P(Type II error).

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

8.4 Distribution Needed for Hypothesis Testing[4]

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a **normal distribution** or a **student's-t distribution.** (Remember, use a student's-t distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.) In this chapter we perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \qquad \text{or} \qquad t_{\text{df}}$$

[4] This content is available online at <http://cnx.org/content/m17017/1.13/>.

The population parameter is μ. The estimated value (point estimate) for μ is $\overline{x}$, the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

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

The population parameter is p. The estimated value (point estimate) for p is p'. $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

8.5 Assumption[5]

When you perform a **hypothesis test of a single population mean** μ using a **Student's-t distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{p \cdot q}{n}}$. Remember that $q = 1 - p$.

8.6 Rare Events[6]

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption** - it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

[5] This content is available online at <http://cnx.org/content/m17002/1.16/>.

[6] This content is available online at <http://cnx.org/content/m16994/1.8/>.

8.7 Using the Sample to Support One of the Hypotheses[7]

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 8.7: (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.

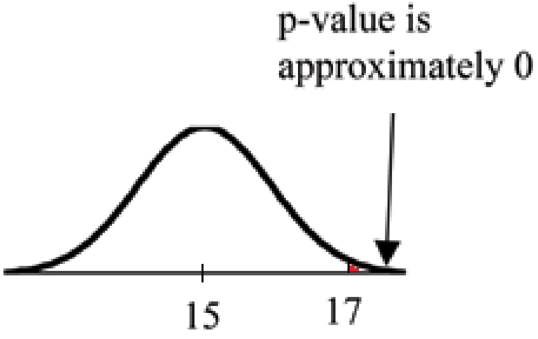
The null hypothesis could be H_o: $\mu \leq 15$ The alternate hypothesis is H_a: $\mu > 15$

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** ($\sigma = 0.5$ 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 $= P\left(\overline{x} > 17\right)$ which is approximately 0.

[7] This content is available online at <http://cnx.org/content/m16995/1.17/>.

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 sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

8.8 Decision and Conclusion[8]

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 $\alpha >$ p-value, reject H_o. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If $\alpha \leq$ p-value, do not reject H_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_0", 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.

8.9 Additional Information[9]

- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, the accepted standard is to use $\alpha = 0.05$.**
- When you calculate the **p-value** and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternate hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the p-value**: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value like 0.4, as opposed to a p-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in failing to reject the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

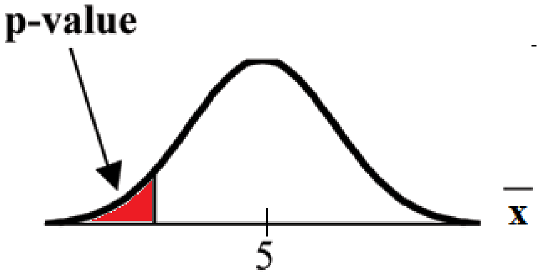
[8] This content is available online at <http://cnx.org/content/m16992/1.11/>.

[9] This content is available online at <http://cnx.org/content/m16999/1.13/>.

The following examples illustrate a left, right, and two-tailed test.

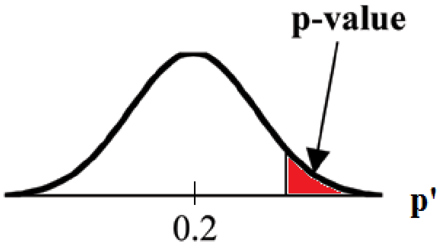
Example 8.8
H_o: $\mu = 5$ H_a: $\mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:



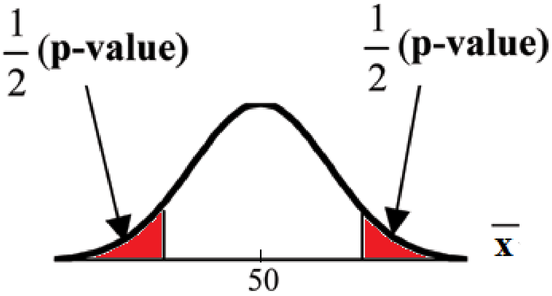
Example 8.9
H_o: $p \leq 0.2$ H_a: $p > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:



Example 8.10
H_o: $\mu = 50$ H_a: $\mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.



8.10 Summary of the Hypothesis Test[10]

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_o and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the **p-value**. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_o), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

[10] This content is available online at <http://cnx.org/content/m16993/1.6/>.

8.11 Lab: Hypothesis Testing of a Single Mean and Single Proportion[11]

Class Time:

Names:

8.11.1 Student Learning Outcomes:

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

8.11.2 Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_o:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately.Shade the actual level of significance.

   a. Graph:

[11] This content is available online at <http://cnx.org/content/m17007/1.12/>.

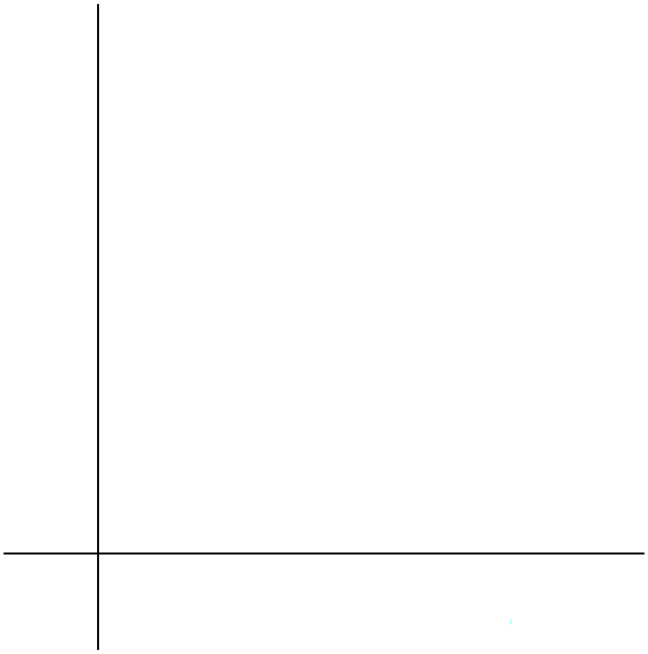


Figure 8.1

b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

8.11.3 Language Survey

About 42.3% of Californians and 19.6% of all Americans over age 5 speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percent of the students at your school that speak a language other than English at home is different from 42.3%. *(Source: http://www.census.gov/hhes/socdemo/language/* [12])

1. H_0:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

[12] http://cnx.org/content/m17007/latest/ http://www.census.gov/hhes/socdemo/language/

a. Graph:

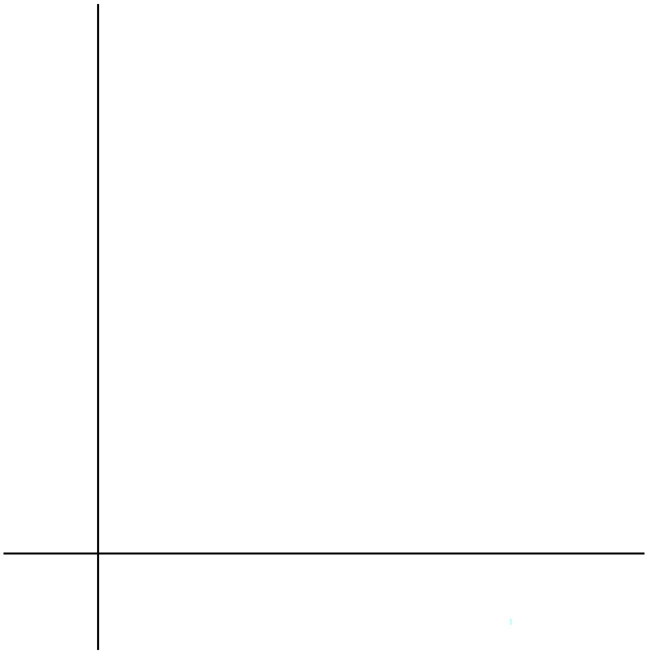


Figure 8.2

b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

8.11.4 Jeans Survey

Suppose that young adults own an average of 3 pairs of jeans. Survey 8 people from your class to determine if the average is higher than 3.

1. H_o:
2. H_a:
3. In words, define the random variable. __________ =
4. The distribution to use for the test is:
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph:

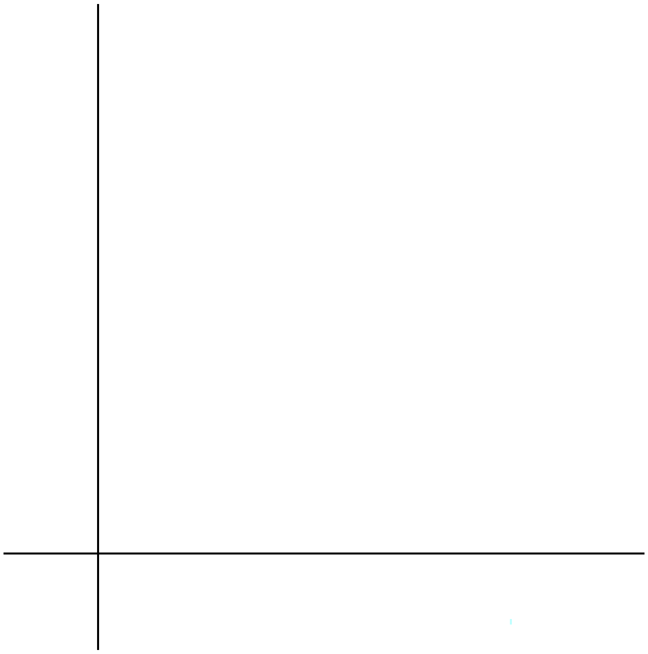


Figure 8.3

b. Determine the p-value:

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Chapter 9

Hypothesis Testing: Two Means, Paired Data, Two Proportions

9.1 Hypothesis Testing: Two Population Means and Two Population Proportions[1]

9.1.1 Student Learning Outcomes

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

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

9.1.2 Introduction

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported about various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

In the previous chapter, you learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

[1] This content is available online at <http://cnx.org/content/m17029/1.9/>.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** mean that the two samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

> NOTE: This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and p-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using the TI-83+/TI-84 calculators, we do not need to separate two population means, independent groups, population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Becomes a test of one population mean.

9.2 Comparing Two Independent Population Means with Unknown Population Standard Deviations[2]

1. The two independent samples are simple random samples from two distinct populations.
2. Both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30. In that case, the populations need not be normally distributed.

> NOTE: The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\overline{X_1}$ - $\overline{X_2}$, and divide by the standard error (shown below) in order to standardize the difference. The result is a t-score test statistic (shown below).

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\overline{X_1}$ - $\overline{X_2}$.

The standard error is:

$$\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}} \tag{9.1}$$

The test statistic (t-score) is calculated as follows:

[2] This content is available online at <http://cnx.org/content/m17025/1.18/>.

t-score

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}} \tag{9.2}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\overline{x_1}$ and $\overline{x_2}$ are the sample means. μ_1 and μ_2 are the population means.

The **degrees of freedom (df)** is a somewhat complicated calculation. However, a computer or calculator calculates it easily. The dfs are not always a whole number. The test statistic calculated above is approximated by the student's-t distribution with dfs as follows:

Degrees of freedom

$$df = \frac{\left[\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right]^2}{\frac{1}{n_1-1} \cdot \left[\frac{(s_1)^2}{n_1}\right]^2 + \frac{1}{n_2-1} \cdot \left[\frac{(s_2)^2}{n_2}\right]^2} \tag{9.3}$$

When both sample sizes n_1 and n_2 are five or larger, the student's-t approximation is very good. Notice that the sample variances ${s_1}^2$ and ${s_2}^2$ are not pooled. (If the question comes up, do not pool the variances.)

NOTE: It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example 9.1: Independent groups
The average amount of time boys and girls ages 7 through 11 spend playing sports each day is believed to be the same. An experiment is done, data is collected, resulting in the table below. Both populations have a normal distribution.

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2 hours	$\sqrt{0.75}$
Boys	16	3.2 hours	1.00

Table 9.1

Problem
Is there a difference in the mean amount of time boys and girls ages 7 through 11 play sports each day? Test at the 5% level of significance.

Solution
The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{X_g} - \overline{X_b}$ = difference in the sample mean amount of time girls and boys play sports each day.

H_o: $\mu_g = \mu_b$ $\qquad$ $\mu_g - \mu_b = 0$

H_a: $\mu_g \neq \mu_b$ $\qquad \mu_g - \mu_b \neq 0$

The words **"the same"** tell you H_o has an "=". Since there are no other words to indicate H_a, then assume **"is different."** This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the p-value using a student's-t distribution: p-value = 0.0054

Graph:

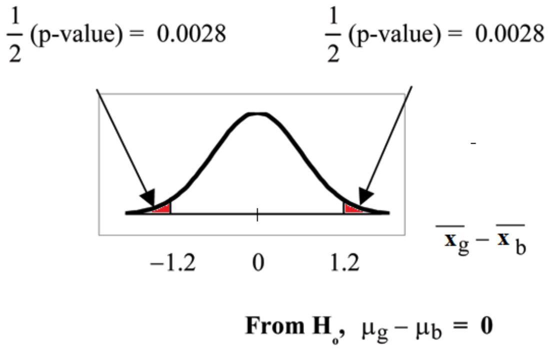


Figure 9.1

$s_g = \sqrt{0.75}$

$s_b = 1$

So, $\overline{x_g} - \overline{x_b} = 2 - 3.2 = -1.2$

Half the p-value is below -1.2 and half is above 1.2.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means you reject $\mu_g = \mu_b$. The means are different.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged 7 through 11 play sports per day is different (mean number of hours boys aged 7 through 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged 7 through 11 play sports per day is greater than the mean number of hours played by boys).

NOTE: TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `4:2-SampTTest`. Arrow over to Stats and press `ENTER`. Arrow down and enter `2` for the first sample mean, $\sqrt{0.75}$ for Sx1, `9` for n1, `3.2` for the second sample mean, `1` for Sx2, and `16` for n2. Arrow down to μ1: and arrow to `does not equal` μ2. Press `ENTER`. Arrow down to Pooled: and `No`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.0054, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Example 9.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples 9 graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from college A **has taken more math classes,** on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

Problem 1 ***(Solution on p. 169.)***

Is this a test of two means or two proportions?

Problem 2 ***(Solution on p. 169.)***

Are the populations standard deviations known or unknown?

Problem 3 ***(Solution on p. 169.)***

Which distribution do you use to perform the test?

Problem 4 ***(Solution on p. 169.)***

What is the random variable?

Problem 5 ***(Solution on p. 169.)***

What are the null and alternate hypothesis?

Problem 6 ***(Solution on p. 169.)***

Is this test right, left, or two tailed?

Problem 7 ***(Solution on p. 169.)***

What is the p-value?

Problem 8 ***(Solution on p. 169.)***

Do you reject or not reject the null hypothesis?

Conclusion:

At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

9.3 Comparing Two Independent Population Means with Known Population Standard Deviations[3]

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal and both populations must be normal. The random variable is $\overline{X_1} - \overline{X_2}$. The normal distribution has the following format:

Normal distribution

$$\overline{X_1} - \overline{X_2} \sim N\left[u_1 - u_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right] \tag{9.4}$$

[3] This content is available online at <http://cnx.org/content/m17042/1.10/>.

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}} \tag{9.5}$$

The test statistic (z-score) is:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}} \tag{9.6}$$

Example 9.3
independent groups, population standard deviations known: The mean lasting time of 2 competing floor waxes is to be compared. **Twenty floors** are randomly assigned **to test each wax**. Both populations have a normal distribution. The following table is the result.

Wax	Sample Mean Number of Months Floor Wax Last	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Table 9.2

Problem
Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

Solution
This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\overline{X_1} - \overline{X_2}$ = difference in the mean number of months the competing floor waxes last.

$H_o : \mu_1 \leq \mu_2$

$H_a : \mu_1 > \mu_2$

The words **"is more effective"** says that **wax 1 lasts longer than wax 2**, on the average. "Longer" is a $'' > ''$ symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula above, the distribution is:

$\overline{X_1} - \overline{X_2} \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right)$

Since $\mu_1 \leq \mu_2$ then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is 0.

Calculate the p-value using the normal distribution: p-value = 0.1799

Graph:

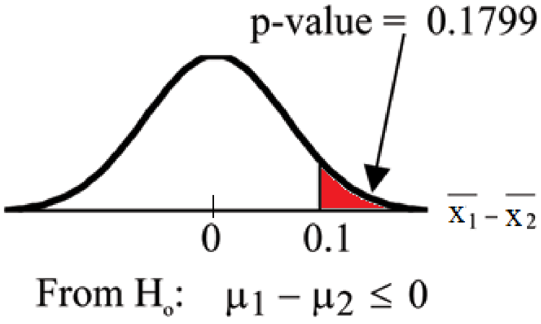


Figure 9.2

$\overline{x_1} - \overline{x_2} = 3 - 2.9 = 0.1$

Compare α and the p-value: $\alpha = 0.05$ and p-value = 0.1799. Therefore, $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject H_o.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

NOTE: TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `3:2-SampZTest`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter `.33` for sigma1, `.36` for sigma2, `3` for the first sample mean, `20` for n1, `2.9` for the second sample mean, and `20` for n2. Arrow down to μ1: and arrow to $>$ μ2. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.1799 and the test statistic is 0.9157. Do the procedure again but instead of `Calculate` do `Draw`.

9.4 Comparing Two Independent Population Proportions[4]

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five and the number of failures is at least five for each of the samples.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions $(P_A - P_B)$ reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_o : p_A = p_B$. To conduct the test, we use a pooled proportion, p_c.

[4] This content is available online at <http://cnx.org/content/m17043/1.12/>.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B} \tag{9.7}$$

The distribution for the differences is:

$$P'_A - P'_B \sim N\left[0, \sqrt{p_c \cdot (1 - p_c) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right] \tag{9.8}$$

The test statistic (z-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c \cdot (1 - p_c) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{9.9}$$

Example 9.4: Two population proportions
Two types of medication for hives are being tested to determine if there is a **difference in the proportions of adult patient reactions. Twenty** out of a random **sample of 200** adults given medication A still had hives 30 minutes after taking the medication. **Twelve** out of another **random sample of 200 adults** given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

9.4.1 Determining the solution

This is a test of 2 population proportions.

Problem ***(Solution on p. 169.)***
How do you know?

Let A and B be the subscripts for medication A and medication B. Then p_A and p_B are the desired population proportions.

Random Variable:
$P'_A - P'_B$ = difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

$H_o : p_A = p_B \qquad p_A - p_B = 0$

$H_a : p_A \neq p_B \qquad p_A - p_B \neq 0$

The words **"is a difference"** tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 0.08 \quad 1 - p_c = 0.92$

Therefore, $P'_A - P'_B \sim N\left[0, \sqrt{(0.08) \cdot (0.92) \cdot \left(\frac{1}{200} + \frac{1}{200}\right)}\right]$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution: p-value = 0.1404.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

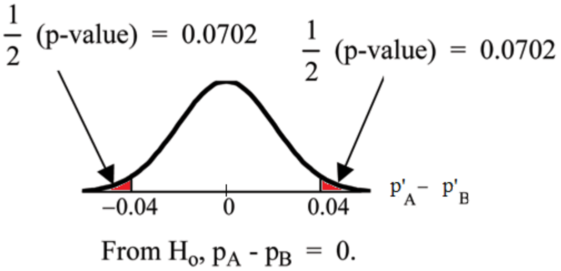


Figure 9.3

$P'_A - P'_B = 0.1 - 0.06 = 0.04.$

Half the p-value is below -0.04 and half is above 0.04.

Compare α and the p-value: $\alpha = 0.01$ and the p-value $= 0.1404$. $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject H_o.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

NOTE: TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `6:2-PropZTest`. Arrow down and enter `20` for $x1$, `200` for $n1$, `12` for $x2$, and `200` for $n2$. Arrow down to `p1:` and arrow to `not equal p2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is $p = 0.1404$ and the test statistic is 1.47. Do the procedure again but instead of `Calculate` do `Draw`.

9.5 Matched or Paired Samples[5]

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.

[5] This content is available online at <http://cnx.org/content/m17033/1.15/>.

6. The matched pairs have differences that either come from a population that is normal or the number of differences is sufficiently large so the distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student-t test for a single population mean with $n-1$ degrees of freedom where n is the number of differences.

The test statistic (t-score) is:

$$t = \frac{\overline{x_d} - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)} \tag{9.10}$$

Example 9.5: Matched or paired samples

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution.

Subject:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Table 9.3

Problem

Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Solution

Corresponding "before" and "after" values form matched pairs. (Calculate "sfter" - "before").

After Data	**Before Data**	**Difference**
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

Table 9.4

The data **for the test** are the differences: {0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6}

The sample mean and sample standard deviation of the differences are: $\overline{x_d} = -3.13$ and $s_d = 2.91$ Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random Variable: $\overline{X_d}$ = the mean difference of the sensory measurements

$$H_o : \mu_d \geq 0 \tag{9.11}$$

There is no improvement. (μ_d is the population mean of the differences.)

$$H_a : \mu_d < 0 \tag{9.12}$$

There is improvement. The score should be lower after hypnotism so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a student-t with $df = n - 1 = 8 - 1 = 7$. Use t_7. **(Notice that the test is for a single population mean.)**

Calculate the p-value using the Student-t distribution: p-value $= 0.0095$

Graph:

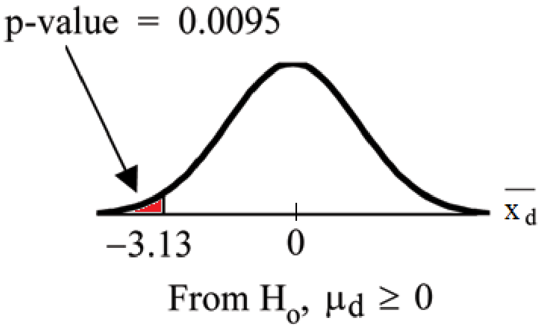


Figure 9.4

$\overline{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$\overline{x}_d = -3.13$

$\overline{s}_d = 2.91$

Compare α and the p-value: $\alpha = 0.05$ and p-value $= 0.0095$. $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

NOTE: For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction and you will have the differences in the third list.

NOTE: TI-83+ and TI-84: Use your list of differences as the data. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to μ: and arrow over to $< \mu_0$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is 0.0094 and the test statistic is -3.04. Do these instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`.

Example 9.6
A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked 4 of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	**Player 1**	**Player 2**	**Player 3**	**Player 4**
Amount of weighted lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

Table 9.5

The coach wants to know if the strength development class makes his players stronger, on average.

Problem ***(Solution on p. 169.)***
Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$\overline{x}_d = 21.3$ $\qquad s_d = 46.7$

Using the difference data, this becomes a test of a single __________ (fill in the blank).

Define the random variable: $\overline{X}_d$ = mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

$H_o : \mu_d \leq 0$ $\qquad H_a : \mu_d > 0$

Graph:

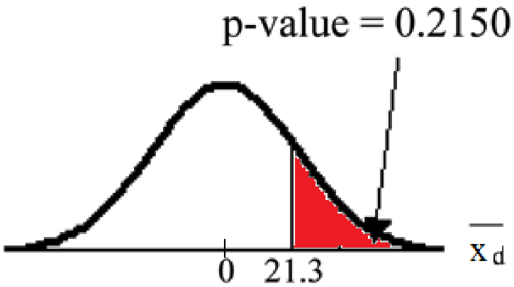


Figure 9.5

Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is to not reject the null hypothesis because $\alpha <$ p-value.

What is the conclusion?

Example 9.7
Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The following data was collected.

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Table 9.6

Problem ***(Solution on p. 169.)***
Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

HINT: use a t-test on the difference data. Assume the differences have a normal distribution. The random variable is the mean difference.

CHECK: The test statistic is 2.18 and the p-value is 0.0716.

What is your conclusion?

9.6 Summary of Types of Hypothesis Tests[6]

Two Population Means

- Populations are independent and population standard deviations are unknown.
- Populations are independent and population standard deviations are known (not likely).

Matched or Paired Samples

- Two samples are drawn from the same set of objects.
- Samples are dependent.

Two Population Proportions

- Populations are independent.

[6] This content is available online at <http://cnx.org/content/m17044/1.5/>.

Solutions to Exercises in Chapter 9

Solution to Example 9.2, Problem 1 (p. 159)
two means
Solution to Example 9.2, Problem 2 (p. 159)
unknown
Solution to Example 9.2, Problem 3 (p. 159)
student's-t
Solution to Example 9.2, Problem 4 (p. 159)
$\overline{X_A} - \overline{X_B}$
Solution to Example 9.2, Problem 5 (p. 159)

- $H_o : \mu_A \leq \mu_B$
- $H_a : \mu_A > \mu_B$

Solution to Example 9.2, Problem 6 (p. 159)
right
Solution to Example 9.2, Problem 7 (p. 159)
0.1928
Solution to Example 9.2, Problem 8 (p. 159)
Do not reject.
Solution to Example 9.4, Problem (p. 162)
The problem asks for a difference in proportions.
Solution to Example 9.6, Problem (p. 166)
means; At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.
Solution to Example 9.7, Problem (p. 167)
H_0: μ_d equals 0; H_a: μ_d does not equal 0; Do not reject the null; At a 5% significance level, from the sample data, there is not sufficient evidence to conclude that the mean difference in distances between the children's dominant versus weaker hands is significant (there is not sufficient evidence to show that the children could push the shot-put further with their dominant hand). Alpha and the p-value are close so the test is not strong.

Chapter 10

Confidence Intervals

10.1 Confidence Intervals[1]

10.1.1 Student Learning Outcomes

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.

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

[1] This content is available online at <http://cnx.org/content/m16967/1.16/>.

Each of $\overline{x}$ 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 $\sigma = 1$ and our sample size is 100. Then by the Central Limit Theorem, the standard deviation for the sample mean is

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

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, 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 $\overline{x}$ is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ 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 $\overline{x} - 0.2$ and $\overline{x} + 0.2$ in 95% of all the samples.

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

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

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 $\overline{x}$ 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).

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

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

10.2 Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal[2]

10.2.1 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 $\overline{x}$ 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 $\overline{x}$ 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,$(\overline{x} - EBM, \overline{x} + EBM)$

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 10.1

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.

$\overline{x} = 7$ and EBM $= 2.5$.

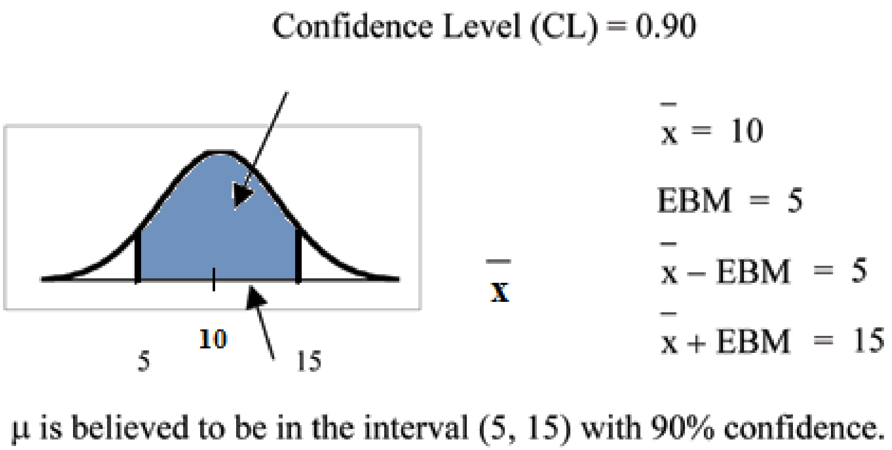
The confidence interval is $(7 - 2.5, 7 + 2.5)$; calculating the values gives $(4.5, 9.5)$.

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 $\overline{x} = 10$ 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.

[2] This content is available online at <http://cnx.org/content/m16962/1.23/>.



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:

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

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

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

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

Finding 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. $CL = 1 - \alpha$. So α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

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

For example, when $CL = 0.95$ then $\alpha = 0.05$ 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: $\texttt{invNorm}(0.975, 0, 1) = 1.96$

CALCULATOR NOTE: Remember to use area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs 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

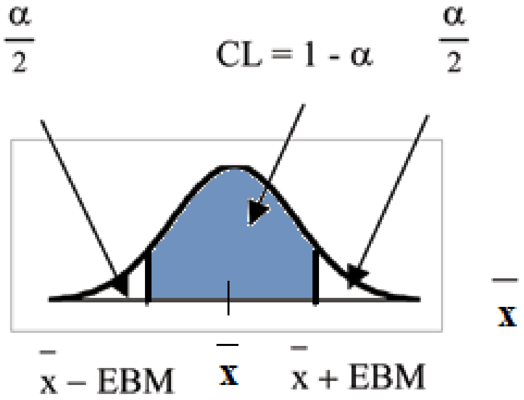
- $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Constructing the Confidence Interval

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

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 10.2

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, $\overline{x}$, and the EBM.

$\overline{x} = 68$
$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
$\sigma = 3$; $n = 36$; The confidence level is 90% (CL=0.90)

$CL \ = \ 0.90$ so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$\frac{\alpha}{2} = 0.05 \qquad z_{\frac{\alpha}{2}} = z_{.05}$

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

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

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.

$EBM = 1.645 \cdot \left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$\overline{x} - EBM = 68 - 0.8225 = 67.1775$

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

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 $\overline{x}$, 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.

10.2.2 Changing the Confidence Level or Sample Size

Example 10.3: 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, $\overline{x}$, and the EBM.

$\overline{x} = 68$
$EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
$\sigma = 3$; $n = 36$; The confidence level is 95% (CL=0.95)

$CL = 0.95$ so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \qquad z_{\frac{\alpha}{2}} = z_{.025}$

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.)

$EBM = 1.96 \cdot \left(\frac{3}{\sqrt{36}}\right) = 0.98$

$\overline{x} - EBM = 68 - 0.98 = 67.02$

$\overline{x} + EBM = 68 + 0.98 = 68.98$

Interpretation

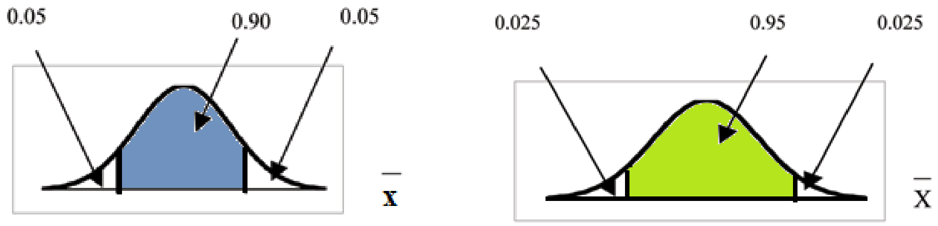
We estimate with 95 % confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

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.



(a)

(b)

Figure 10.1

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 10.4: Changing the Sample Size:
Suppose we change the original problem to see what happens to the error bound if the sample size is changed.

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?

- $\overline{x} = 68$
- $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; 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.

10.2.3 Working Backwards to Find the Error Bound or Sample Mean

Working Bacwards 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 difference 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 10.5
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: $\overline{x} = 68.82 - 0.82 = 68$
- If we don't know the error bound: $\overline{x} = \frac{(67.18+68.82)}{2} = 68$

10.2.4 Calculating the Sample Size n

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

The error bound formula for a population mean when the population standard deviation is known is $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$

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 wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 10.6
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 $\sigma = 15$ and EBM=2
$z = z_{.025} = 1.96$, because the confidence level is 95%.

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.96^2 15^2}{2^2}$ =216.09 using the sample size equation.
Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

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

**With contributions from Roberta Bloom

10.3 Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T[3]

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 $t = \frac{\overline{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, 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 $\overline{x}$ 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 ($x - \overline{x}$ **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.

[3] This content is available online at <http://cnx.org/content/m16959/1.24/>.

- 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 specified.

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 (See the Table of Contents **15. Tables**) 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

- $T \sim t_{\text{df}}$ where $\text{df} = n - 1$.
- 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 $\text{df} = n - 1$ degrees of freedom
- s = sample standard deviation

The format for the confidence interval is:

$(\overline{x} - EBM, \overline{x} + EBM)$.

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 10.7
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, $\overline{x}$, and the EBM.

$\overline{x} = 8.2267 \qquad s = 1.6722 \qquad n = 15$

$\text{df} = 15 - 1 = 14$

$\text{CL} = 0.95 \;\; \text{so} \;\; \alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \qquad t_{\frac{\alpha}{2}} = t_{.025}$

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.

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

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

$\overline{x} - \text{EBM} = 8.2267 - 0.9240 = 7.3$

$\overline{x} + \text{EBM} = 8.2267 + 0.9240 = 9.15$

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

10.4 Confidence Interval for a Population Proportion[4]

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 $X \sim B(n, p)$ 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 : $\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$

[4] This content is available online at <http://cnx.org/content/m16963/1.20/>.

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 $(p' - \textbf{EBP}, p' + \textbf{EBP})$.

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

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

$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} \qquad \textit{whereq}' = 1 - p'$

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 10.8

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 $\text{CL} = 0.95$, then $\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$ $\frac{\alpha}{2} = 0.025$.

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

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 $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

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

$p' - \text{EBP} = 0.842 - 0.032 = 0.81$

$p' + \text{EBP} = 0.842 + 0.032 = 0.874$

The confidence interval for the true binomial population proportion is $(p' - \text{EBP}, p' + \text{EBP}) = (0.810, 0.874)$.

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 to `A: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 10.9
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 $\mathrm{CL} = 0.90$, then $\alpha = 1 - \mathrm{CL} = 1 - 0.90 = 0.10$ $\frac{\alpha}{2} = 0.05$.

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

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.

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

$p' - \mathrm{EBP} = 0.60 - 0.036 = 0.564$

$p' + \mathrm{EBP} = 0.60 + 0.036 = 0.636$

The confidence interval for the true binomial population proportion is $(p' - \mathrm{EBP}, p' + \mathrm{EBP}) = (0.564, 0.636)$.

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).

10.4.1 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 10.10
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.

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

10.5 Summary of Formulas[5]

Formula 10.1: General form of a confidence interval
(lower value, upper value) = (point estimate − error bound, point estimate + error bound)

Formula 10.2: To find the error bound when you know the confidence interval
error bound = upper value − point estimate OR error bound = $\frac{\text{upper value}-\text{lower value}}{2}$

Formula 10.3: Single Population Mean, Known Standard Deviation, Normal Distribution
Use the Normal Distribution for Means (Section 7.2) EBM = $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

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

Formula 10.4: Single Population Mean, Unknown Standard Deviation, Student's-t Distribution
Use the Student's-t Distribution with degrees of freedom df = $n - 1$. EBM = $t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

Formula 10.5: Single Population Proportion, Normal Distribution
Use the Normal Distribution for a single population proportion $p' = \frac{x}{n}$

EBP = $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}}$ $\quad p' + q' = 1$

The confidence interval has the format $(p' - \text{EBP}, p' + \text{EBP})$.

Formula 10.6: Point Estimates
$\overline{x}$ is a point estimate for μ
p' is a point estimate for ρ

s is a point estimate for σ

[5] This content is available online at <http://cnx.org/content/m16973/1.8/>.

Chapter 11

F Distribution and ANOVA

11.1 F Distribution and ANOVA[1]

11.1.1 Student Learning Outcomes

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

- Interpret the F probability distribution as the number of groups and the sample size change.
- Discuss two uses for the F distribution: One-Way ANOVA and the test of two variances.
- Conduct and interpret One-Way ANOVA.
- Conduct and interpret hypothesis tests of two variances.

11.1.2 Introduction

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests involving more than two averages, statisticians have developed a method called Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or One-Way ANOVA. You will also study the F distribution, used for One-Way ANOVA, and the test of two variances. This is just a very brief overview of One-Way ANOVA. You will study this topic in much greater detail in future statistics courses.

- One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.
- For further information about One-Way ANOVA, use the online link ANOVA[2] . Use the back button to return here. (The url is http://en.wikipedia.org/wiki/Analysis_of_variance.)

[1] This content is available online at <http://cnx.org/content/m17065/1.11/>.

[2] http://en.wikipedia.org/wiki/Analysis_of_variance

11.2 ANOVA[3]

11.2.1 F Distribution and One-Way ANOVA: Purpose and Basic Assumptions of One-Way ANOVA

The purpose of a **One-Way ANOVA** test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to help determine if the means are equal or not.

In order to perform a One-Way ANOVA test, there are five basic **assumptions** to be fulfilled:

- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.
- The populations are assumed to have **equal standard deviations (or variances).**
- The factor is the categorical variable.
- The response is the numerical variable.

11.2.2 The Null and Alternate Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternate hypothesis is that at least one pair of means is different. For example, if there are k groups:

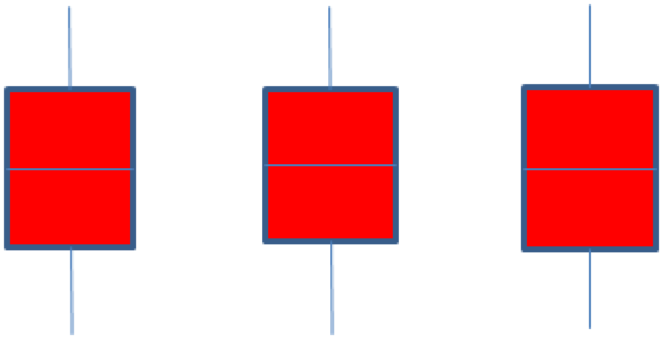
$H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_a : At least two of the group means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal.

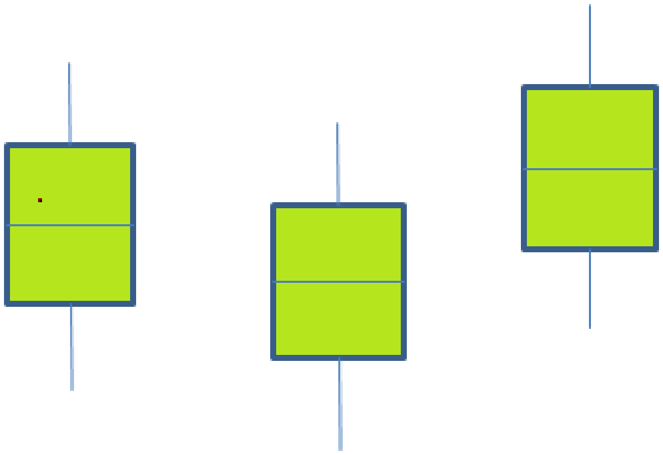
The graphs help in the understanding of the hypothesis test. In the first graph (red box plots), $H_o : \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

[3] This content is available online at <http://cnx.org/content/m17068/1.10/>.



Ho is true. All means are the same.



Ho is false. All means are not the same.

11.3 The F Distribution and the F Ratio[4]

The distribution used for the hypothesis test is a new one. It is called the *F* distribution, named after Sir Ronald Fisher, an English statistician. The *F* statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

[4] This content is available online at <http://cnx.org/content/m17076/1.14/>.

For example, if F follows an F distribution and the degrees of freedom for the numerator are 4 and the degrees of freedom for the denominator are 10, then $F \sim F_{4,10}$.

> NOTE: The F distribution is derived from the Student's-t distribution. One-Way ANOVA expands the t-test for comparing more than two groups. The scope of that derivation is beyond the level of this course.

To calculate the F ratio, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when there is equal n). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- SS_{between} = the sum of squares that represents the variation among the different samples.
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities which, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in **Descriptive Statistics**.

MS means "mean square." MS_{between} is the variance between groups and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the jth group
- s_j= the sum of the values in the jth group
- n = total number of all the values combined. (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $\text{SS}_{\text{total}} = \sum x^2 - \frac{(\sum x)^2}{n}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{n}$
- Explained variation- sum of squares representing variation among the different samples $\text{SS}_{\text{between}} = \sum \left[\frac{(\text{sj})^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$
- Unexplained variation- sum of squares representing variation within samples due to chance: $\text{SS}_{\text{within}} = \text{SS}_{\text{total}} - \text{SS}_{\text{between}}$
- df's for different groups (df's for the numerator): $\text{df}_{\text{between}} = k - 1$
- Equation for errors within samples (df's for the denominator): $\text{df}_{\text{within}} = n - k$
- Mean square (variance estimate) explained by the different groups: $\text{MS}_{\text{between}} = \frac{\text{SS}_{\text{between}}}{\text{df}_{\text{between}}}$
- Mean square (variance estimate) that is due to chance (unexplained): $\text{MS}_{\text{within}} = \frac{\text{SS}_{\text{within}}}{\text{df}_{\text{within}}}$

MS_{between} and MS_{within} can be written as follows:

- $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$
- $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{SS_{\text{within}}}{n-k}$

The One-Way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

> NOTE: The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{11.1}$$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_o is true), then the F-ratio should be approximately equal to 1. Mostly just sampling errors would contribute to variations away from 1. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}. Then the F-ratio will be larger than 1. However, if the population effect size is small it is not unlikely that MS_{within} will be larger in a give sample.

The above calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot {s_{\bar{x}}}^2}{s^2{}_{\text{pooled}}} \tag{11.2}$$

where ...

- n =the sample size
- $\text{df}_{\text{numerator}} = k - 1$
- $\text{df}_{\text{denominator}} = n - k$
- $s^2{}_{pooled}$ = the mean of the sample variances (pooled variance)
- ${s_{\bar{x}}}^2$ = the variance of the sample means

The data is typically put into a table for easy viewing. One-Way ANOVA results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F

continued on next page

Factor (Between)	SS(Factor)	k - 1	MS(Factor) = SS(Factor)/(k-1)	F = MS(Factor)/MS(Error)
Error (Within)	SS(Error)	n - k	MS(Error) = SS(Error)/(n-k)	
Total	SS(Total)	n - 1		

Table 11.1

Example 11.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The One-Way ANOVA table is shown below.

Plan 1	**Plan 2**	**Plan 3**
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

Table 11.2

One-Way ANOVA Table: The formulas for SS(Total), SS(Factor) = SS(Between) and SS(Error) = SS(Within) are shown above. This same information is provided by the TI calculator hypothesis test function ANOVA in STAT TESTS (syntax is ANOVA(L1, L2, L3) where L1, L2, L3 have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	**Sum of Squares (SS)**	**Degrees of Freedom (df)**	**Mean Square (MS)**	**F**
Factor (Between)	SS(Factor) = SS(Between) =2.2458	k - 1 = 3 groups - 1 = 2	MS(Factor) = SS(Factor)/(k-1) = 2.2458/2 = 1.1229	F = MS(Factor)/MS(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	SS(Error) = SS(Within) = 20.8542	n - k = 10 total data - 3 groups = 7	MS(Error) = SS(Error)/(n-k) = 20.8542/7 = 2.9792	
Total	SS(Total) = 2.9792 + 20.8542 =23.1	n - 1 = 10 total data - 1 = 9		

Table 11.3

The One-Way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_o.

11.3.1 Notation

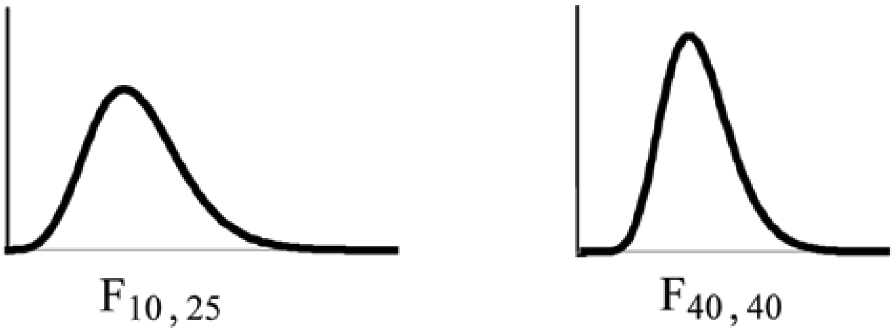
The notation for the F distribution is $F \sim F_{\text{df(num)},\text{df(denom)}}$

where $\text{df(num)} = df_{\text{between}}$ and $\text{df(denom)} = df_{\text{within}}$

The mean for the F distribution is $\mu = \frac{df(num)}{df(denom)-1}$

11.4 Facts About the F Distribution[5]

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of dfs.
3. The F statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
5. Other uses for the F distribution include comparing two variances and Two-Way Analysis of Variance. Comparing two variances is discussed at the end of the chapter. Two-Way Analysis is mentioned for your information only.



(a)

(b)

Figure 11.1

Example 11.2
One-Way ANOVA: Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown below:

[5] This content is available online at <http://cnx.org/content/m17062/1.14/>.

MEAN GRADES FOR FOUR SORORITIES			
Sorority 1	**Sorority 2**	**Sorority 3**	**Sorority 4**
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Table 11.4

Problem

Using a significance level of 1%, is there a difference in mean grades among the sororities?

Solution

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each size 5.

NOTE: This is an example of a **balanced design**, since each factor (i.e. Sorority) has the same number of observations.

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means $\mu_1, \mu_2, \mu_3, \mu_4$ are equal.

Distribution for the test: $F_{3,16}$

where $k = 4$ groups and $n = 20$ samples in total

$df\ (num) = k - 1 = 4 - 1 = 3$

$df\ (denom) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph:

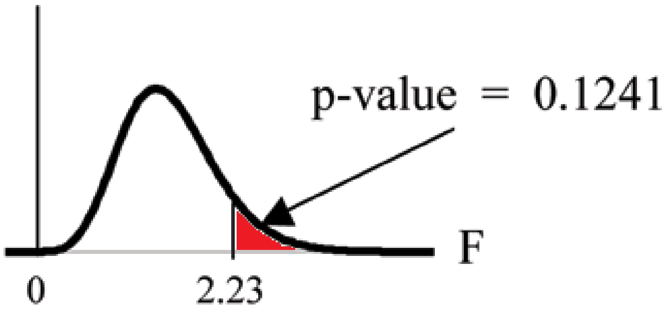


Figure 11.2

Probability statement: p-value $= P(F > 2.23) = 0.1241$

Compare α **and the** $p - value$**:** $\alpha = 0.01$ p-value $= 0.1241$ $\alpha <$ p-value

Make a decision: Since $\alpha <$ p-value, you cannot reject H_o.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

TI-83+ or TI 84: Put the data into lists L1, L2, L3, and L4. Press `STAT` and arrow over to `TESTS`. Arrow down to `F:ANOVA`. Press `ENTER` and Enter (`L1,L2,L3,L4`). The F statistic is 2.2303 and the p-value is 0.1241. df(numerator) = 3 (under "`Factor`") and df(denominator) = 16 (under `Error`).

Example 11.3
A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew 5 plants. At the end of the growing period, each plant was measured, producing the following data (in inches):

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Table 11.5

Problem 1

Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 3% level of significance.

Solution

This time, we will perform the calculations that lead to the F′ statistic. Notice that each group has the same number of plants so we will use the formula $F' = \frac{n \cdot s_{\overline{x}}{}^2}{s^2{}_{\text{pooled}}}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Table 11.6

Next, calculate the variance of the three group means (Calculate the variance of 24.2, 25.4, and 24.4). **Variance of the group means = 0.413** = $s_{\overline{x}}{}^2$

Then $MS_{\text{between}} = ns_{\overline{x}}{}^2 = (5)\,(0.413)$ where $n = 5$ is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (Calculate the mean of 11.7, 18.3, and 16.3). **Mean of the sample variances = 15.433** $= s^2{}_{\text{pooled}}$

Then $MS_{\text{within}} = s^2{}_{\text{pooled}} = 15.433$.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{n \cdot s_{\overline{x}}{}^2}{s^2{}_{\text{pooled}}} = \frac{(5)\cdot(0.413)}{15.433} = 0.134$

The dfs for the numerator = the number of groups $- 1 = 3 - 1 = 2$

The dfs for the denominator = the total number of samples $-$ the number of groups $= 15 - 3 = 12$

The distribution for the test is $F_{2,12}$ and the F statistic is $F = 0.134$

The p-value is $P\,(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value $= 0.8759$, do not reject H_o. (Why?)

Conclusion: With a 3% the level of significance, from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

(This experiment was actually done by three classmates of the son of one of the authors.)

Another fourth grader also grew bean plants but this time in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32.

Problem 2 ***(Solution on p. 203.)***

Do a One-Way ANOVA test on the 4 groups. You may use your calculator or computer to perform the test. Are the heights of the bean plants different? Use a solution sheet[6].

[6] "Collaborative Statistics: Solution Sheets: F Distribution and One-Way ANOVA" <http://cnx.org/content/m17135/latest/>

11.4.1 Optional Classroom Activity

From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1% level of significance. Use one of the solution sheets[7] at the end of the chapter (after the homework).

11.5 Test of Two Variances[8]

Another of the uses of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

In order to perform a F test of two variances, it is important that the following are true:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}$$

F has the distribution $F \sim F\left(n_1 - 1, n_2 - 1\right)$

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F-Ratio becomes $F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}$.

> NOTE: The F ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to 1. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than 1. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

[7] "Collaborative Statistics: Solution Sheets: F Distribution and One-Way ANOVA" <http://cnx.org/content/m17135/latest/>

[8] This content is available online at <http://cnx.org/content/m17075/1.8/>.

Therefore, if F is close to 1, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than 1, then the evidence is against the null hypothesis.

A test of two variances may be left, right, or two-tailed.

Example 11.4
Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9.

Problem
Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

Solution
Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30.$

H_o: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

Calculate the test statistic: By the null hypothesis $\left(\sigma_1^2 = \sigma_2^2\right)$, the F statistic is

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818$$

Distribution for the test: $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph: This test is left tailed.

Draw the graph labeling and shading appropriately.

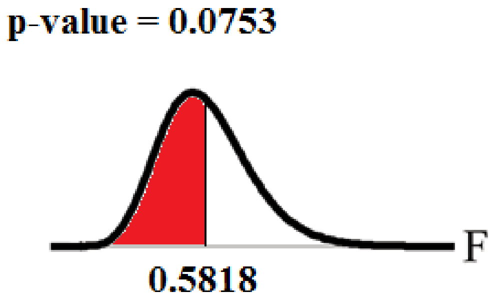


Figure 11.3

Probability statement: p-value $= P\,(F < 0.5818) = 0.0753$

Compare α and the p-value: $\alpha = 0.10$ $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o.

Conclusion: With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

TI-83+ and TI-84: Press `STAT` and arrow over to `TESTS`. Arrow down to `D:2-SampFTest.` Press `ENTER`. Arrow to `Stats` and press `ENTER`. For `Sx1`, `n1`, `Sx2`, and `n2`, enter $\sqrt{(52.3)}$, `30`, $\sqrt{(89.9)}$, and `30`. Press `ENTER` after each. Arrow to σ`1:` and $<\sigma 2$. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. $F = 0.5818$ and p-value $= 0.0753$. Do the procedure again and try `Draw` instead of `Calculate`.

11.6 Summary[9]

- A **One-Way ANOVA** hypothesis test determines if several population means are equal. The distribution for the test is the F distribution with 2 different degrees of freedom.

  Assumptions:

  a. Each population from which a sample is taken is assumed to be normal.
  b. Each sample is randomly selected and independent.
  c. The populations are assumed to have equal standard deviations (or variances)

- A **Test of Two Variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the F distribution with 2 different degrees of freedom.

  Assumptions:

  a. The populations from which the two samples are drawn are normally distributed.
  b. The two populations are independent of each other.

[9] This content is available online at <http://cnx.org/content/m17072/1.4/>.

Solutions to Exercises in Chapter 11

Solution to Example 11.3, Problem 2 (p. 198)

- $F = 0.9496$
- $p - value = 0.4402$

From the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Chapter 12

The Chi-Square Distribution

12.1 The Chi-Square Distribution[1]

12.1.1 Student Learning Outcomes

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

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests.

12.1.2 Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to the above examples. This distribution is called the Chi-square distribution.

In this chapter, you will learn the three major applications of the Chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as with the lottery example
- The test of independence, which determines if events are independent, such as with the movie example
- The test of a single variance, which tests variability, such as with the coffee example

> NOTE: Though the Chi-square calculations depend on calculators or computers for most of the calculations, there is a table available (see the Table of Contents **15. Tables**). TI-83+ and TI-84 calculator instructions are included in the text.

[1] This content is available online at <http://cnx.org/content/m17048/1.9/>.

12.1.3 Optional Collaborative Classroom Activity

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, etc.). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

12.2 Notation[2]

The notation for the chi-square distribution is:

$\chi^2 \sim \chi^2_{\mathrm{df}}$

where df = degrees of freedom depend on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2 \cdot df}$.

The random variable is shown as χ^2 but may be any upper case letter.

The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

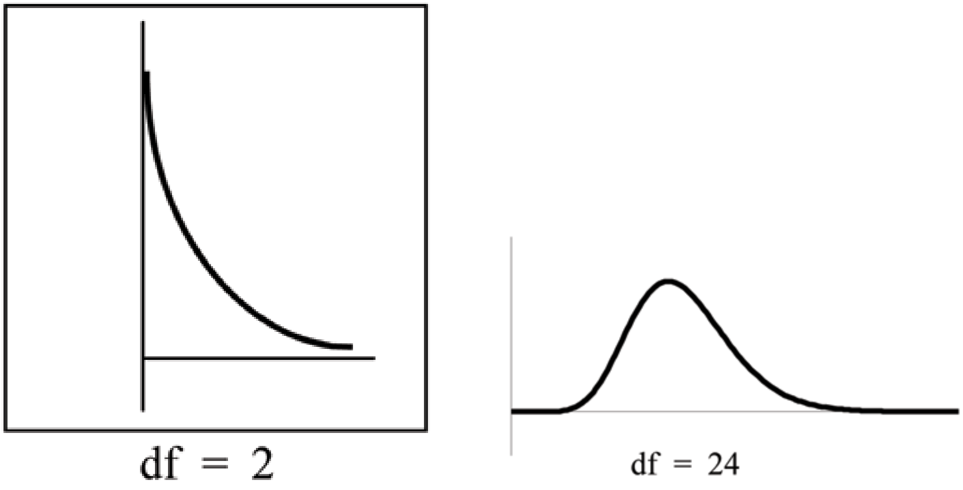
$\chi^2 = (Z_1)^2 + (Z_2)^2 + ... + (Z_k)^2$

12.3 Facts About the Chi-Square Distribution[3]

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.

[2] This content is available online at <http://cnx.org/content/m17052/1.6/>.

[3] This content is available online at <http://cnx.org/content/m17045/1.6/>.



(a)

(b)

Figure 12.1

3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal. For $X \sim \chi^2_{1000}$ the mean, $\mu = df = 1000$ and the standard deviation, $\sigma = \sqrt{2 \cdot 1000} = 44.7$. Therefore, $X \sim N\,(1000, 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.

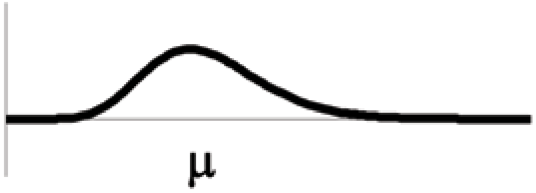


Figure 12.2

In the next sections, you will learn about four different applications of the Chi-Square Distribution. These hypothesis tests are almost always right-tailed tests. In order to understand why the tests are mostly right-tailed, you will need to look carefully at the actual definition of the test statistic. Think about the following while you study the next four sections. If the expected and observed values are "far" apart, then the test statistic will be "large" and we will reject in the right tail. The only way to obtain a test statistic very close to zero, would be if the observed and expected values are very, very close to each other. A left-tailed test could be used to determine if the fit were "too good." A "too good" fit might occur if data had been manipulated or invented. Think about the implications of right-tailed versus left-tailed hypothesis tests as you learn the applications of the Chi-Square Distribution.

12.4 Goodness-of-Fit Test[4]

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:

$$\sum_{k} \frac{(O-E)^2}{E} \tag{12.1}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The degrees of freedom are df = (number of categories - 1).

The goodness-of-fit test is almost always right tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

NOTE: The expected value for each cell needs to be at least 5 in order to use this test.

Example 12.1
Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism follows faculty perception. The faculty expected that a group of 100 students would miss class according to the following chart.

Number absences per term	Expected number of students
0 - 2	50
3 - 5	30
6 - 8	12
9 - 11	6
12+	2

Table 12.1

A random survey across all mathematics courses was then done to determine the actual number **(observed)** of absences in a course. The next chart displays the result of that survey.

[4] This content is available online at <http://cnx.org/content/m17192/1.8/>.

Number absences per term	Actual number of students
0 - 2	35
3 - 5	40
6 - 8	20
9 - 11	1
12+	4

Table 12.2

Determine the null and alternate hypotheses needed to conduct a goodness-of-fit test.

H_o**:** Student absenteeism **fits** faculty perception.

The alternate hypothesis is the opposite of the null hypothesis.

H_a**:** Student absenteeism **does not fit** faculty perception.

Problem 1

Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution

No. Notice that the expected number of absences for the "12+" entry is less than 5 (it is 2). Combine that group with the "9 - 11" group to create new tables where the number of students for each entry are at least 5. The new tables are below.

Number absences per term	Expected number of students
0 - 2	50
3 - 5	30
6 - 8	12
9+	8

Table 12.3

Number absences per term	Actual number of students
0 - 2	35
3 - 5	40
6 - 8	20
9+	5

Table 12.4

Problem 2

What are the degrees of freedom (df)?

Solution

There are 4 "cells" or categories in each of the new tables.

$$df = number\ of\ cells - 1 = 4 - 1 = 3$$

Example 12.2

Employers particularly want to know which days of the week employees are absent in a five day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week did they have the highest number of employee absences. The results were distributed as follows:

Day of the Week Employees were most Absent

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Number of Absences	15	12	9	9	15

Table 12.5

Problem

For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five day work week? Test at a 5% significance level.

Solution

The null and alternate hypotheses are:

- H_o: The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- $(O - E)$
- $(O - E)^2$
- $\frac{(O - E)^2}{E}$

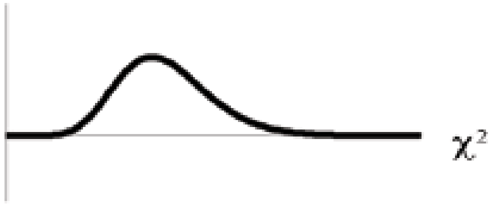
The last column ($\frac{(O - E)^2}{E}$) should have 0.75, 0, 0.75, 0.75, 0.75.
Now add (sum) the last column. Verify that the sum is 3. This is the χ^2 test statistic.

To find the p-value, calculate $P\left(\chi^2 > 3\right)$. This test is right-tailed.
(Use a computer or calculator to find the p-value. You should get p-value $= 0.5578$.)

The dfs are the number of cells $- 1 = 5 - 1 = 4$.

TI-83+ and TI-84: Press `2nd DISTR`. Arrow down to χ^2`cdf`. Press `ENTER`. Enter `(3,10^99,4)`. Rounded to 4 decimal places, you should see 0.5578 which is the p-value.

Next, complete a graph like the one below with the proper labeling and shading. (You should shade the right tail.)



The decision is to not reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

NOTE: TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example (Example 11-3) has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start. See below.

NOTE: **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Or, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER`.

Example 12.3
One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Number of Televisions	**Percent**
0	10
1	16
2	55
3	11
over 3	8

Table 12.6

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the following data:

Number of Televisions	**Frequency**
0	66
1	119
2	340
3	60
over 3	15
	Total = 600

Table 12.7

The table contains observed (O) frequency values.

Problem

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are:

Number of Televisions	**Percent**	**Expected Frequency**
0	10	$(0.10)\cdot(600)=60$
1	16	$(0.16)\cdot(600)=96$
2	55	$(0.55)\cdot(600)=330$
3	11	$(0.11)\cdot(600)=66$
over 3	8	$(0.08)\cdot(600)=48$

Table 12.8

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter .10*600.

H_o: The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

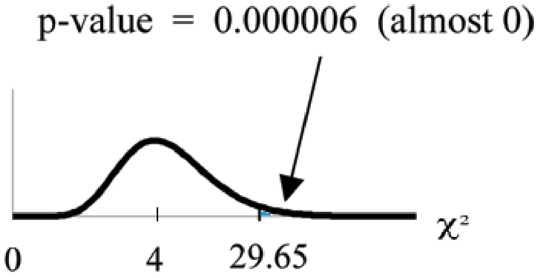
H_a: The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where $df=$ (the number of cells) $-1=5-1=4$.

NOTE: $df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:



Probability statement: p-value $= P\left(\chi^2 > 29.65\right) = 0.000006$.

Compare α **and the p-value:**

- $\alpha = 0.01$
- p-value $= 0.000006$

So, $\alpha\ >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

NOTE: TI-83+ and some TI-84 calculators: Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of Example 11-2). Into `L1`, put the observed frequencies `66, 119, 349, 60, 15`. Into `L2`, put the expected frequencies `.10*600, .16*600, .55*600, .11*600, .08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:`χ`2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to 4 places, you should see `5.77E-6 = .000006` (rounded to 6 decimal places) which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Example 12.4

Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the value 0, 1, 2. (There are 0, 1, or 2 heads in the flip of 2 coins.) Therefore, the **number of cells is 3**. Since X = the number of heads, the observed frequencies are 20 (for 2 heads), 57 (for 1 head), and 23 (for 0 heads or both tails). The expected frequencies are 25 (for 2 heads), 50 (for 1 head), and 25 (for 0 heads or both tails). This test is right-tailed.

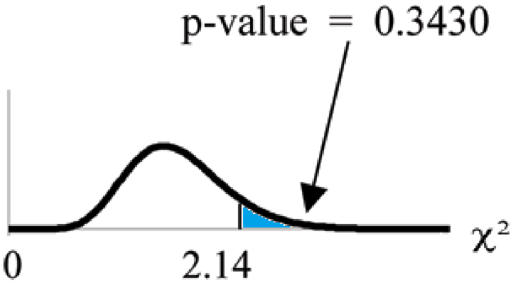
H_o: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



Probability statement: p-value $= P\left(\chi^2 > 2.14\right) = 0.3430$

Compare α and the p-value:

- $\alpha = 0.05$
- p-value $= 0.3430$

So, $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject H_o.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

NOTE: TI-83+ and some TI- 84 calculators: Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`.`Enter L3`. Rounded to 2 decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:`χ`2cdf` (or press 7). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to 4 places, you should see `.3430` which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

12.5 Test of Independence[5]

Tests of independence involve using a **contingency table** of observed (data) values. You first saw a contingency table when you studied probability in the Probability Topics (Section 3.1) chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E} \tag{12.2}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \ \cdot \ j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

NOTE: The expected value for each cell needs to be at least 5 in order to use this test.

Example 12.5

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A \ AND \ B) = P(A)\,P(B)$. $A \ AND \ B$ is the event that a driver received a speeding violation last year and is also a cell phone user while driving. Suppose, in a study of drivers who received speeding violations in the last year and who uses cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were cell phone users while driving and 450 were not.

Let y = expected number of drivers that use a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \ AND \ B) = P(A)\,P(B)$. By substitution,

$\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$

Solve for $y : y = \frac{70 \cdot 305}{755} = 28.3$

[5] This content is available online at <http://cnx.org/content/m17191/1.12/>.

About 28 people from the sample are expected to be cell phone users while driving and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**. If we do a test of independence using the example above, then the null hypothesis is:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be cell phone users while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

The degrees of freedom for the test of independence are:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the **expected number** (E):

$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}}$

Example 12.6

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Number of Hours Worked Per Week by Volunteer Type (Observed)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Table 12.9: The table contains **observed (O)** values (data).

Problem

Are the number of hours volunteered **independent** of the type of volunteer?

Solution

The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_0: The number of hours volunteered is **independent** of the type of volunteer.

H_a: The number of hours volunteered is **dependent** on the type of volunteer.

The expected table is:

Number of Hours Worked Per Week by Volunteer Type (Expected)

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Table 12.10: The table contains **expected** (E) values (data).

For example, the calculation for the expected frequency for the top left cell is

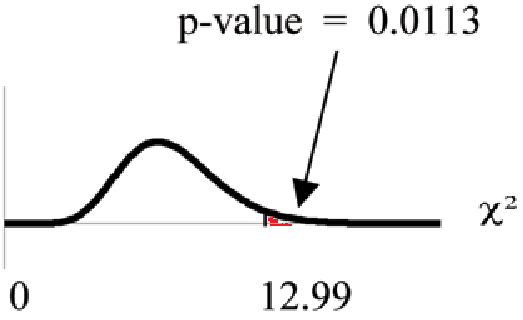
$$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}} = \frac{255 \cdot 298}{839} = 90.57$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

df $= (3 \text{ columns} - 1)\,(3 \text{ rows} - 1) = (2)\,(2) = 4$

Graph:



Probability statement: p-value $= P\left(\chi^2 > 12.99\right) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value $= 0.0113$. $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, reject H_o. This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

NOTE: Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from Example 11-6. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:`χ`2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the p-value = 0.0113. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Example 12.7
De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School vs. Anxiety Level

Need to Succeed in School	**High Anxiety**	**Med-high Anxiety**	**Medium Anxiety**	**Med-low Anxiety**	**Low Anxiety**	**Row Total**
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Table 12.11

Problem 1
How many high anxiety level students are expected to have a high need to succeed in school?

Solution
The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Problem 2
If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution
The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Problem 3

a. $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}} =$

b. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

12.6 Test of a Single Variance (Optional)[6]

A test of a single variance assumes that the underlying distribution is **normal**. The null and alternate hypotheses are stated in terms of the **population variance** (or population standard deviation). The test statistic is:

$$\frac{(n-1)\cdot s^2}{\sigma^2} \tag{12.3}$$

where:

- n = the total number of data
- s^2 = sample variance
- σ^2 = population variance

You may think of s as the random variable in this test. The degrees of freedom are $\text{df} = n - 1$.

A test of a single variance may be right-tailed, left-tailed, or two-tailed.

The following example will show you how to set up the null and alternate hypotheses. The null and alternate hypotheses contain statements about the population variance.

Example 12.8
Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is 5 points. One of his best students thinks otherwise. The student claims that the standard deviation is more than 5 points. If the student were to conduct a hypothesis test, what would the null and alternate hypotheses be?

Solution
Even though we are given the population standard deviation, we can set the test up using the population variance as follows.

- H_o: $\sigma^2 = 5^2$
- H_a: $\sigma^2 > 5^2$

Example 12.9
With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

[6] This content is available online at <http://cnx.org/content/m17059/1.7/>.

With a significance level of 5%, test the claim that **a single line causes lower variation among waiting times (shorter waiting times) for customers**.

Solution
Since the claim is that a single line causes lower variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random Variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

- H_o: $\sigma^2 = 7.2^2$
- H_a: $\sigma^2 < 7.2^2$

The word **"lower"** tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where:

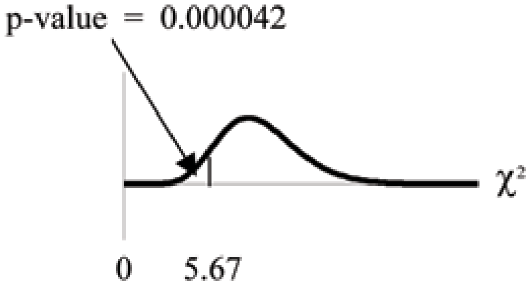
- n = the number of customers sampled
- df $= n - 1 = 25 - 1 = 24$

Calculate the test statistic:

$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2} = \frac{(25-1)\cdot 3.5^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph:



Probability statement: p-value $= P\left(\chi^2 < 5.67\right) = 0.000042$

Compare α and the p-value: $\alpha = 0.05$ p-value $= 0.000042$ $\alpha >$ p-value

Make a decision: Since $\alpha >$ p-value, reject H_o.

This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes, but lower.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times **or** with a single line, the customer waiting times vary less than 7.2 minutes.

TI-83+ and TI-84 calculators: In `2nd DISTR`, use `7:`χ`2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example 11-9, χ`2cdf(-1E99,5.67,24)`. The p-value = 0.000042.

12.7 Summary of Formulas[7]

The Chi-Square Probability Distribution
$\mu = \text{df}$ and $\sigma = \sqrt{2 \cdot \text{df}}$

Goodness-of-Fit Hypothesis Test

- Use goodness-of-fit to test whether a data set fits a particular probability distribution.
- The degrees of freedom are number of cells or categories - 1.
- The test statistic is $\sum\limits_{k} \frac{(O-E)^2}{E}$, where O = observed values (data), E = expected values (from theory), and k = the number of different data cells or categories.
- The test is right-tailed.

Test of Independence

- Use the test of independence to test whether two factors are independent or not.
- The degrees of freedom are equal to (number of columns - 1)(number of rows - 1).
- The test statistic is $\sum\limits_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

Test of Homogeneity

- Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other.
- The degrees of freedom are equal to number of columns - 1.
- The test statistic is $\sum\limits_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

NOTE: The expected value for each cell needs to be at least 5 in order to use the Goodness-of-Fit, Independence and Homogeneity tests.

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom are the number of samples - 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left, right, or two-tailed.

[7] This content is available online at <http://cnx.org/content/m17058/1.8/>.

Chapter 13

Linear Regression and Correlation

13.1 Linear Regression and Correlation[1]

13.1.1 Student Learning Outcomes

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.

13.1.2 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.

13.2 Linear Equations[2]

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

$$y = a + \mathrm{bx} \tag{13.1}$$

[1] This content is available online at <http://cnx.org/content/m17089/1.6/>.

[2] This content is available online at <http://cnx.org/content/m17086/1.4/>.

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 13.1
The following examples are linear equations.

$$y = 3 + 2\text{x} \tag{13.2}$$

$$y = -0.01 + 1.2\text{x} \tag{13.3}$$

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

Example 13.2

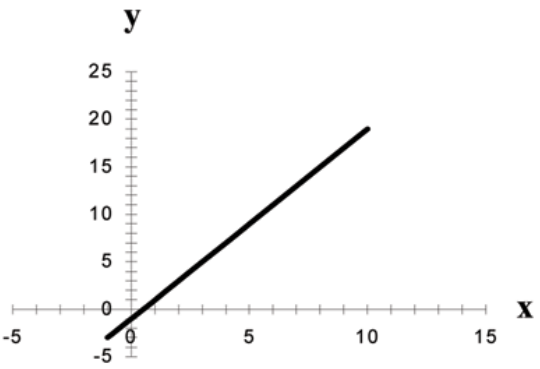


Figure 13.1: Graph of the equation $y = -1 + 2\text{x}$.

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

Example 13.3
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.

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:

$y = 31.50 + 32\text{x}$

13.3 Slope and Y-Intercept of a Linear Equation[3]

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

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

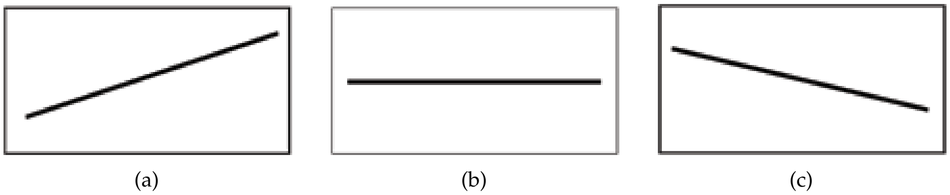


Figure 13.2: Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 13.4
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 $y = 25 + 15\text{x}$.

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.

[3] This content is available online at <http://cnx.org/content/m17083/1.5/>.

13.4 Scatter Plots[4]

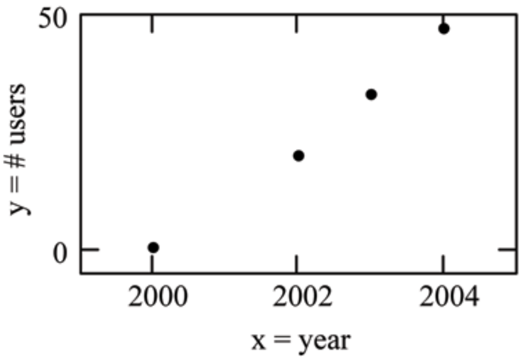
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example 13.5

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

(a)



(b)

Figure 13.3: (a) Table showing the number of m-commerce users (in millions) by year. (b) Scatter plot showing the number of m-commerce users (in millions) by year.

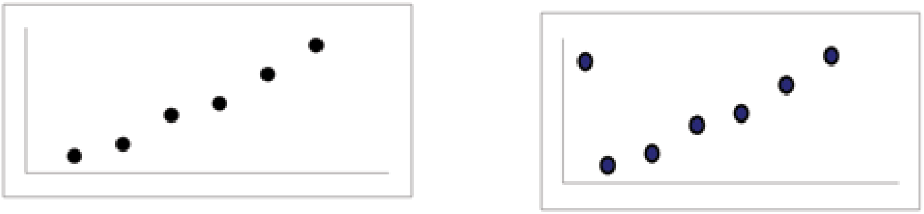
A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

- 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.

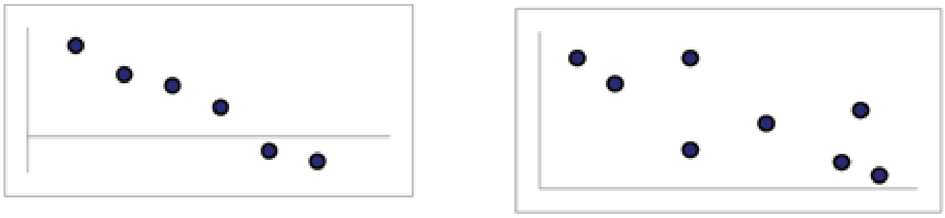
[4] This content is available online at <http://cnx.org/content/m17082/1.8/>.



(a) Positive Linear Pattern (Strong)

(b) Linear Pattern w/ One Deviation

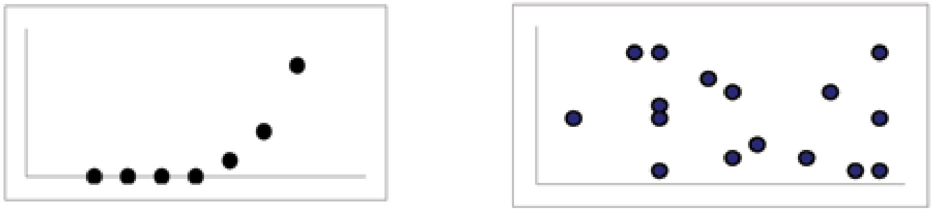
Figure 13.4



(a) Negative Linear Pattern (Strong)

(b) Negative Linear Pattern (Weak)

Figure 13.5



(a) Exponential Growth Pattern

(b) No Pattern

Figure 13.6

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.

13.5 The Regression Equation[5]

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**.

13.5.1 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?

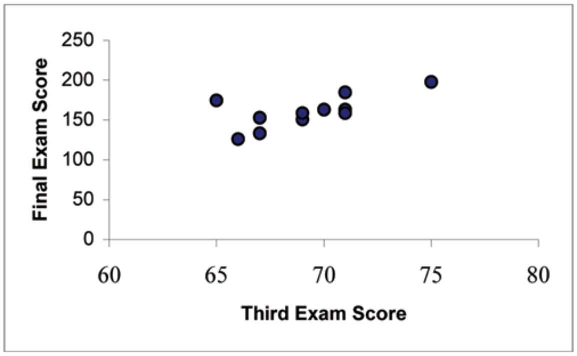
Example 13.6

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?

[5] This content is available online at <http://cnx.org/content/m17090/1.15/>.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

(a)



(b)

Figure 13.7: (a) Table showing the scores on the final exam based on scores from the third exam. (b) Scatter plot showing the scores on the final 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 the form $(x, y)$and each point of the line of best fit using least-squares linear regression has the form $\left(x, \hat{y}\right)$.

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.

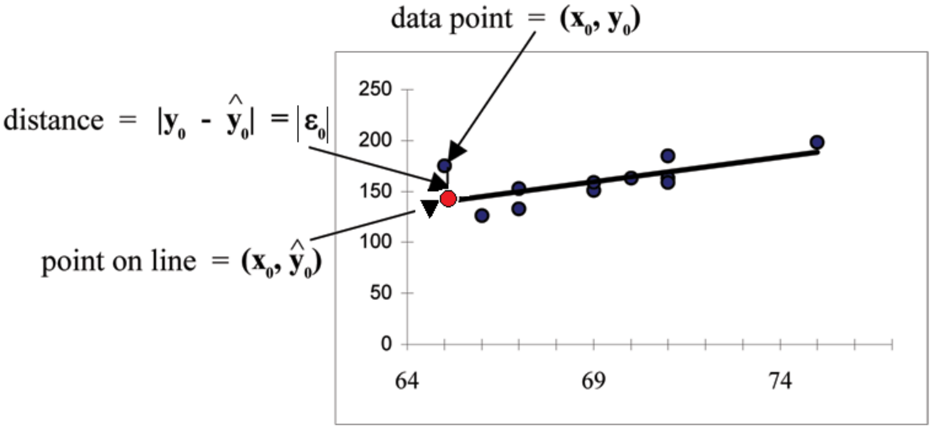


Figure 13.8

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 final 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:

$$\hat{y} = a + \text{bx} \tag{13.4}$$

where $a = \overline{y} - b \cdot \overline{x}$ and $b = \frac{\Sigma(x-\overline{x})\cdot(y-\overline{y})}{\Sigma(x-\overline{x})^2}$.

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

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:

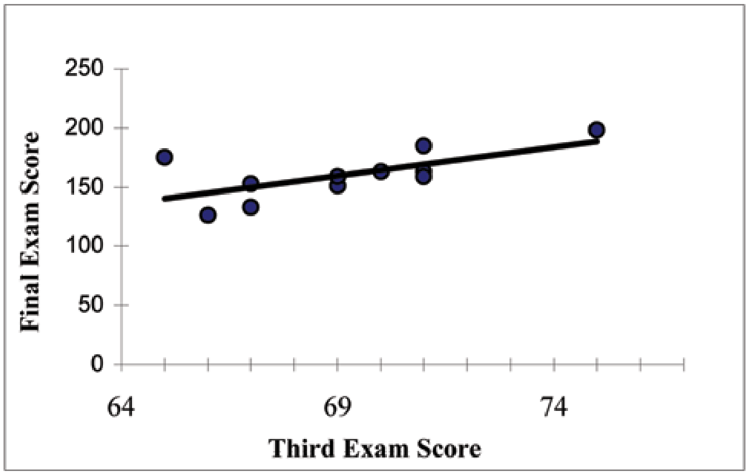


Figure 13.9

The least squares regression line (best fit line) for the third exam/final exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{13.5}$$

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.**

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 final exam score increases by 4.83 points, on average.

13.5.2 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.)

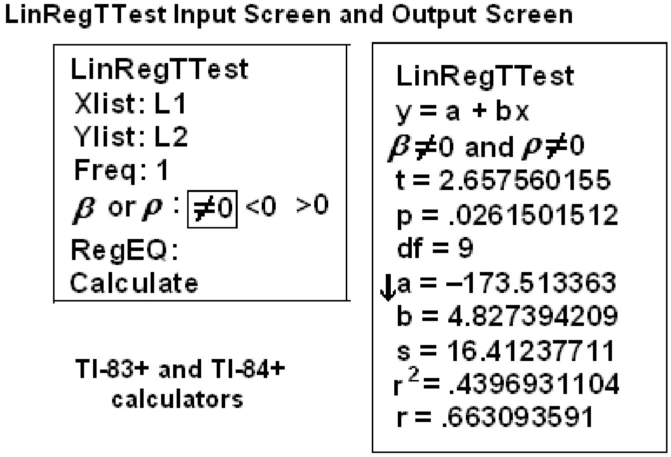
Step 2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)

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
r²= .4396931104
r = .663093591
```

Figure 13.10

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 fit the window to the data
Step 8. To graph the best fit 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

13.6 The Correlation Coefficient[6]

13.6.1 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

$$r = \frac{n \cdot \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y)}{\sqrt{\left[n \cdot \Sigma x^2 - (\Sigma x)^2\right] \cdot \left[n \cdot \Sigma y^2 - (\Sigma y)^2\right]}} \tag{13.6}$$

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:

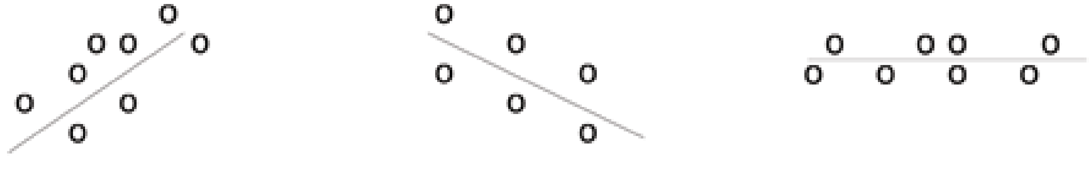
- The value of r is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to -1 or to +1 indicate a stronger linear relationship between x and y.
- If $r = 0$ there is absolutely no linear relationship 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!

[6] This content is available online at <http://cnx.org/content/m17092/1.12/>.



(a) Positive Correlation (b) Negative Correlation (c) Zero Correlation

Figure 13.11: (a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (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).

13.6.2 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 final 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.

13.7 Facts About the Correlation Coefficient for Linear Regression[7]

13.7.1 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 compute 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_o: $\rho = 0$

[7] This content is available online at <http://cnx.org/content/m17077/1.15/>.

- **Alternate Hypothesis:** H_a**:** $\rho \neq \mathbf{0}$

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** H_a**:** 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 $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. 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_o: $\rho = 0$
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 significance level of $\alpha = 0.05$
Decision: Reject the Null Hypothesis H_o
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.

METHOD 2: Using a table of Critical Values to make a decision
The 95% Critical Values of the Sample Correlation Coefficient Table (Section 13.10) at the end of this chapter (before the Summary (Section 13.11)) 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 13.7
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.

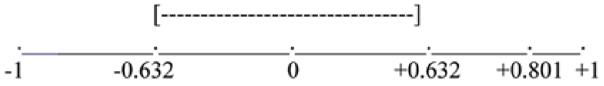


Figure 13.12: r is not significant between -0.632 and +0.632. $r = 0.801 > +0.632$. Therefore, r is significant.

Example 13.8
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

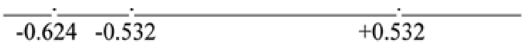


Figure 13.13: $r = -0.624 < -0.532$. Therefore, r is significant.

Example 13.9
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.

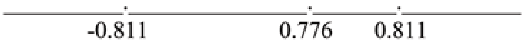


Figure 13.14: $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

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_o: $\rho = 0$
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 H_o:
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 13.10: 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. $0.708 > 0.666$ so r is significant.
3. $r = 0.134$ and the sample size, n, is 14. The df $= 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
4. $r = 0$ 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.

13.7.2 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).

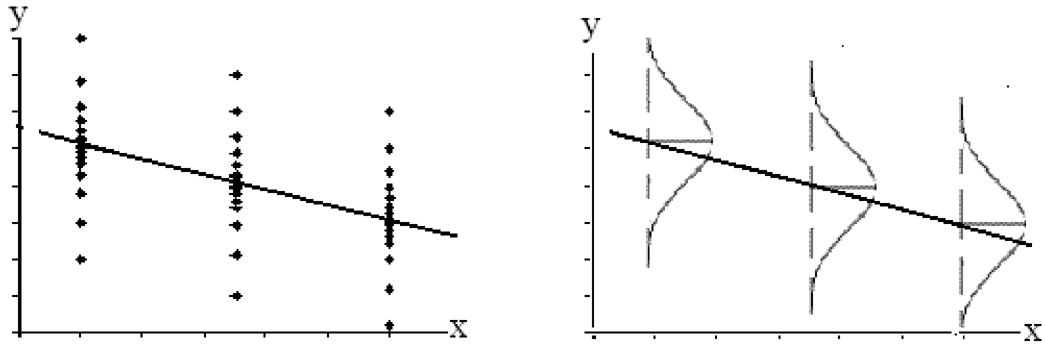


Figure 13.15: The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

**With contributions from Roberta Bloom

13.8 Prediction[8]

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:

$$\hat{y} = -173.51 + 4.83\,(73) = 179.08 \qquad (13.8)$$

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 13.11
Recall the third exam/final exam example.

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 ***(Solution on p. 252.)***
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

13.9 Outliers[9]

In some data sets, there are values **(observed data points)** called **outliers**. **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to carefully examine what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

[8] This content is available online at <http://cnx.org/content/m17095/1.8/>.
[9] This content is available online at <http://cnx.org/content/m17094/1.14/>.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers
We could guess at outliers by looking at a graph of the scatterplot and best fit line. However we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best fit line as an outlier**. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatterplot by drawing an extra pair of lines that are two standard deviations above and below the best fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally only need to use one of these methods.

Example 13.12
In the third exam/final exam example, you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Solution
Graphical Identification of Outliers
With the TI-83,83+,84+ graphing calculators, it is easy to identify the outlier graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance was equal to 2*s* or farther, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find **s=16.412**

Line Y2=-173.5+4.83x-2(16.4) and line Y3=-173.5+4.83x+2(16.4)

where $\hat{y}$=-173.5+4.83x is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatterplot with the best fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the "Y="equation editor and press ZOOM 9. You will find that the only data point that is not between lines Y2 and Y3 is the point x=65, y=175. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than 2 standard deviations away from the best fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

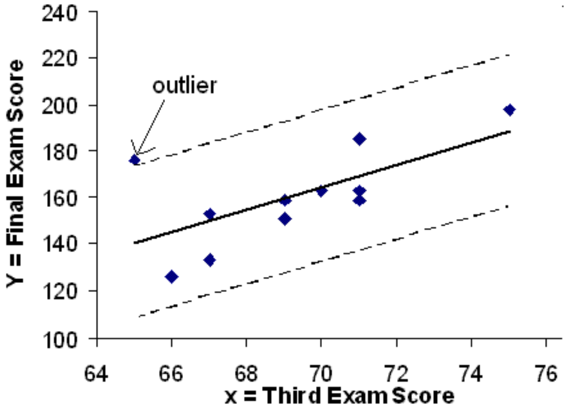


Figure 13.16

Numerical Identification of Outliers

In the table below, the first two columns are the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y}$=-173.5+4.83x. The residuals, or errors, have been calculated in the fourth column of the table: observed y value $-$ predicted y value $= y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \epsilon$ values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{\text{SSE}}{n-2}}$$

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, the calculator function LinRegTTest found $s = 16.4$ as the standard deviation of the residuals 35; -17; 16; -6; -19; 9; 3; -1; -10; -9; -1.

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	$175 - 140 = 35$
67	133	150	$133 - 150 = -17$
71	185	169	$185 - 169 = 16$
71	163	169	$163 - 169 = -6$
66	126	145	$126 - 145 = -19$
75	198	189	$198 - 189 = 9$
67	153	150	$153 - 150 = 3$
70	163	164	$163 - 164 = -1$
71	159	169	$159 - 169 = -10$
69	151	160	$151 - 160 = -9$
69	159	160	$159 - 160 = -1$

Table 13.1

We are looking for all data points for which the residual is greater than $2s$=2(16.4)=32.8 or less than -32.8. Compare these values to the residuals in column 4 of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line?
Numerically and graphically, we have identified the point (65,175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. **For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.**

Compute a new best-fit line and correlation coefficient using the 10 remaining points:
On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39\text{x}$ and $r = 0.9121$

The new line with $r = 0.9121$ is a stronger correlation than the original (r=0.6631) because $r = 0.9121$ is closer to 1. This means that the new line is a better fit to the 10 remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating s and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, **square each** $|y - \hat{y}|$ (See the TABLE above):

The squares are 35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2

Then, add (sum) all the $|y - \hat{y}|$ **squared terms** using the formula

$$\sum_{i=1}^{11}\left(|y_i - \hat{y}_i|\right)^2 = \sum_{i=1}^{11} \epsilon_i{}^2 \qquad \text{(Recall that } y_i - \hat{y}_i = \epsilon_i.\text{)}$$

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

$= 2440 =$ **SSE**. The result, **SSE** is the Sum of Squared Errors.

Next, calculate s**, the standard deviation of all the** $y - \hat{y} = \epsilon$ **values where** n **= the total number of data points.**

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$

Next, multiply s by 1.9:
$(1.9) \cdot (16.47) = 31.29$

31.29 is almost 2 standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $1.9s$, then we would consider the data point to be "too far" from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $|y - \hat{y}|$ values are **at least** 31.29, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam problem, all the $|y - \hat{y}|$'s are less than 31.29 except for the first one which is 35.

$35 > 31.29$ That is, $|y - \hat{y}| \geq (1.9) \cdot (s)$

The point which corresponds to $|y - \hat{y}| = 35$ is $(65, 175)$. **Therefore, the data point** $(65, 175)$ **is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

Example 13.13

Using this new line of best fit (based on the remaining 10 data points), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Example 13.14

(*From The Consumer Price Indexes Web site*) The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Data:

x	y
1915	10.1
1926	17.7
1935	13.7
1940	14.7
1947	24.1
1952	26.5
1964	31.0
1969	36.7
1975	49.3
1979	72.6
1980	82.4
1986	109.6
1991	130.7
1999	166.6

Table 13.2

Problem

- Make a scatterplot of the data.
- Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
- Draw the line on the scatterplot.
- Find the correlation coefficient. Is it significant?
- What is the average CPI for the year 1990?

Solution

- Scatter plot and line of best fit.
- $\hat{y}= -3204 + 1.662x$ is the equation of the line of best fit.
- $r = 0.8694$
- The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since $0.8694 > 0.532$, r is significant.
- $\hat{y}= -3204 + 1.662\,(1990) = 103.4$ CPI
- Using the calculator LinRegTTest, we find that s = 25.4 ; graphing the lines Y2=-3204+1.662X-2(25.4) and Y3=-3204+1.662X+2(25.4) shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)

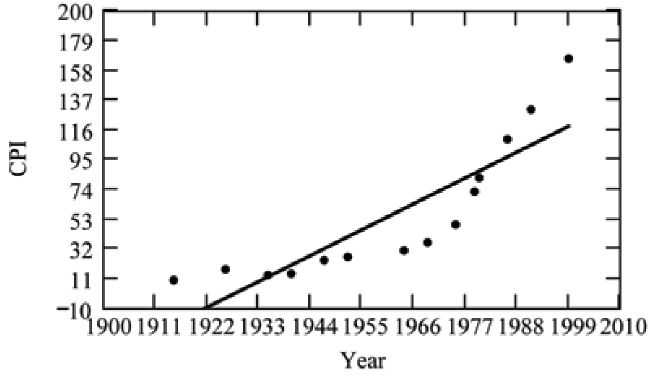


Figure 13.17

NOTE: In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt ; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years 2004 : CPI=188.9, 2008 : CPI=215.3 and 2011: CPI=224.9. See how it affects the model. (Check: $\hat{y}= -4436 + 2.295x$. $r = 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

**With contributions from Roberta Bloom

13.10 95% Critical Values of the Sample Correlation Coefficient Table[10]

[10] This content is available online at <http://cnx.org/content/m17098/1.6/>.

Degrees of Freedom: $n-2$	**Critical Values: ($+$ and $-$)**
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374

continued on next page

27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Table 13.3

13.11 Summary[11]

Bivariate Data: Each data point has two values. The form is (x, y).

Line of Best Fit or Least Squares Line (LSL): $\hat{y} = a + \text{bx}$

x = independent variable; y = dependent variable

Residual: Actual y value $-$ predicted y value $= y - \hat{y}$
Correlation Coefficient r:

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer r is to 1 or -1, the closer the original points are to a straight line.
3. If r is negative, the slope is negative. If r is positive, the slope is positive.
4. If $r = 0$, then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the **SSE**, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

[11] This content is available online at <http://cnx.org/content/m17081/1.4/>.

Solutions to Exercises in Chapter 13

Solution to Example 13.11, Problem 2 (p. 241)
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**.

Glossary

A

Analysis of Variance

Also referred to as ANOVA. A method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- All populations of interest are normally distributed.
- The populations have equal standard deviations.
- Samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the F-ratio.

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

Bernoulli Trials

An experiment with the following characteristics:

- There are only 2 possible outcomes called "success" and "failure" for each trial.
- The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X{\sim}B\left(n,p\right)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P\left(X = x\right) = \binom{n}{x} p^x q^{n-x}$.

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, $\overline{X}$, and the sample sum, ΣX. If the size n of the sample is sufficiently large, then $\overline{X}{\sim}\ N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N\left(n\mu, \sqrt{n}\sigma\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

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:

$$r = \frac{n\sum \mathrm{xy} - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}, \tag{13.7}$$

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.

Conditional Probability

The likelihood that an event will occur given that another event has already occurred.

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.

Contingency Table

The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

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 **Equally Likely**

Each outcome of an experiment has the same probability.

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.

Event

A subset in the set of all outcomes of an experiment. The set of all outcomes of an experiment is called a **sample space** and denoted usually by S. An event is any arbitrary subset in **S**. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, etc. Standard notations for events are capital letters such as A, B, C, etc.

Expected Value

Expected arithmetic average when an experiment is repeated many times. (Also called the mean). Notations: $E(x)$, μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $E(x) = \mu = \sum xP(x)$.

Experiment

A planned activity carried out under controlled conditions.

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 $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = \text{me}^{-\text{mx}}, \ x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-\text{mx}}$.

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 **Independent Events**

The occurrence of one event has no effect on the probability of the occurrence of any other event. Events A and B are independent if one of the following is true: (1). $P(A|B) = P(A)$; (2) $P(B|A) = P(B)$; (3) $P(AandB) = P(A)P(B)$.

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.

Interquartile Range (IRQ)

The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

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

Median

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

Mode

The value that appears most frequently in a set of data.

Mutually Exclusive

An observation cannot fall into more than one class (category). Being in more than one category prevents being in a mutually exclusive category.

N

Normal Distribution

A continuous random variable (RV) with pdf f(x) $= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)2}/2\sigma^{2}$, where μ is the mean of the distribution and σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

O

Outcome (observation)

A particular result of an experiment.

Outlier

An observation that does not fit the rest of the data.

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.

Percentile

A number that divides ordered data into hundredths.

Example: Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0...\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Point Estimate

A single number computed from a sample and used to estimate a population parameter.

Poisson Distribution

A discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval. Characteristics of the variable:

- The probability that the event occurs in a given interval is the same for all intervals.
- The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean μ of the event in the interval. Notation: $X\sim P(\mu)$. The mean is $\mu = \text{np}$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(X = x) = e^{-\mu}\frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution when n is "large" and p is "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to .05).

Population

The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Probability

A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1$;.
- If A and B are any two mutually exclusive events, then $P(AorB) = P(A) + P(B)$.
- $P(S) = 1$.

Probability Distribution Function (PDF)

A mathematical description of a discrete random variable (RV), given either in the form of an equation (formula) , or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Example: A biased coin with probability 0.7 for a head (in one toss of the coin) is tossed 5 times. We are interested in the number of heads (the RV X = the number of heads). X is Binomial, so $X \sim B(5, 0.7)$ and $P(X = x) = \begin{pmatrix} 5 \\ x \end{pmatrix} .7^x .3^{5-x}$or in the form of the table:

x	$P(X = x)$
0	0.0024
1	0.0284
2	0.1323
3	0.3087
4	0.3602
5	0.1681

Table 4.3

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. $P' = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, $P' \sim N\left(p, \frac{pq}{n}\right)$.

Q **Qualitative Data**

See **Data**.

Quantitative Data

Quartiles

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

R **Random Variable (RV)**

see **Variable**

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.

Sample Space

The set of all possible outcomes of an experiment.

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) $X \sim N(0, 1)$ .. When X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

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 flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Student-t Distribution

T

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Type 2 Error

The decision is to not reject the Null hypothesis when, in fact, the Null hypothesis is false.

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: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(X) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

V

Variable (Random Variable)

A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

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 $x - \overline{x}$ where x is a value of the data and $\overline{x}$ 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 $z = \frac{x-\mu}{\sigma}$. If this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$, the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ , the result is called the z-score of x. Z-scores allow us to compare data that are normally distributed but scaled differently.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Quantitative Information Analysis III*
Edited by: Jeffrey Stanton
URL: http://cnx.org/content/col11155/1.1/
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Introduction"
Used here as: "Sampling and Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16008/1.9/
Page: 1
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Statistics"
Used here as: "Statistics"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16020/1.16/
Pages: 1-4
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Probability"
Used here as: "Probability"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16015/1.11/
Page: 4
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Key Terms"
Used here as: "Key Terms"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16007/1.17/
Pages: 4-6
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Data"
Used here as: "Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16005/1.18/
Pages: 6-12
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Sampling"
Used here as: "Sampling"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16014/1.17/
Pages: 12-19
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Variation and Critical Evaluation"
Used here as: "Variation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16021/1.15/
Pages: 19-21
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Answers and Rounding Off"
Used here as: "Answers and Rounding Off"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16006/1.8/
Page: 21
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Frequency, Relative Frequency, and Cumulative Frequency"
Used here as: "Frequency"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16012/1.20/
Pages: 21-25
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Sampling and Data: Summary"
Used here as: "Summary"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16023/1.10/
Page: 26
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Introduction"
Used here as: "Descriptive Statistics"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16300/1.9/
Page: 29
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Displaying Data"
Used here as: "Displaying Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16297/1.9/
Pages: 29-30
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs"
Used here as: "Stem and Leaf Graphs (Stemplots)"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16849/1.17/
Pages: 30-33
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Histogram"
Used here as: "Histograms"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16298/1.14/
Pages: 33-37
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Box Plot"
Used here as: "Box Plots"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16296/1.13/
Pages: 38-41
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Measuring the Location of the Data"
Used here as: "Measures of the Location of the Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16314/1.18/
Pages: 41-46
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Measuring the Center of the Data"
Used here as: "Measures of the Center of the Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17102/1.13/
Pages: 46-49
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Skewness and the Mean, Median, and Mode"
Used here as: "Skewness and the Mean, Median, and Mode"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17104/1.9/
Pages: 49-50
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Measuring the Spread of the Data"
Used here as: "Measures of the Spread of the Data"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17103/1.15/
Pages: 51-58
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Descriptive Statistics: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16310/1.9/
Page: 59
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Introduction"
Used here as: "Probability Topics"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16838/1.11/
Pages: 63-64
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Terminology"
Used here as: "Terminology"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16845/1.13/
Pages: 64-66
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Independent & Mutually Exclusive Events"
Used here as: "Independent and Mutually Exclusive Events"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16837/1.14/
Pages: 66-69
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Two Basic Rules of Probability"
Used here as: "Two Basic Rules of Probability"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16847/1.11/
Pages: 69-73
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Contingency Tables"
Used here as: "Contingency Tables"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16835/1.12/
Pages: 73-76
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Probability Topics: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16843/1.5/
Page: 77
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Introduction"
Used here as: "Discrete Random Variables"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16825/1.14/
Pages: 81-82
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Probability Distribution Function (PDF) for a Discrete Random Variable"
Used here as: "Probability Distribution Function (PDF) for a Discrete Random Variable"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16831/1.14/
Pages: 82-83
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Mean or Expected Value and Standard Deviation"
Used here as: "Mean or Expected Value and Standard Deviation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16828/1.16/
Pages: 83-86
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Common Discrete Probability Distribution Functions"
Used here as: "Common Discrete Probability Distribution Functions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16821/1.6/
Page: 86
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Binomial"
Used here as: "Binomial"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16820/1.17/
Pages: 86-89
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Poisson (optional)"
Used here as: "Poisson"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16829/1.16/
Pages: 89-91
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Discrete Random Variables: Summary of the Discrete Probability Functions"
Used here as: "Summary of Functions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16833/1.11/
Pages: 92-93
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Continuous Random Variables: Introduction"
Used here as: "Continuous Random Variables"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16808/1.12/
Pages: 97-99
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Continuous Random Variables: Continuous Probability Functions"
Used here as: "Continuous Probability Functions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16805/1.9/
Pages: 99-101
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Continuous Random Variables: The Uniform Distribution"
Used here as: "The Uniform Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16819/1.17/
Pages: 102-109
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Continuous Random Variables: The Exponential Distribution"
Used here as: "The Exponential Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16816/1.15/
Pages: 109-114
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Continuous Random Variables: Summary of The Uniform and Exponential Probability Distributions"
Used here as: "Summary of the Uniform and Exponential Probability Distributions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16813/1.10/
Page: 115
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Normal Distribution: Introduction"
Used here as: "The Normal Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16979/1.12/
Pages: 117-118
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Normal Distribution: Standard Normal Distribution"
Used here as: "The Standard Normal Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16986/1.7/
Page: 118
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Normal Distribution: Z-scores"
Used here as: "Z-scores"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16991/1.10/
Pages: 119-120
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Normal Distribution: Areas to the Left and Right of x"
Used here as: "Areas to the Left and Right of x"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16976/1.5/
Page: 121
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Normal Distribution: Calculations of Probabilities"
Used here as: "Calculations of Probabilities"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16977/1.12/
Pages: 121-124
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Normal Distribution: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16987/1.5/
Page: 125
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Central Limit Theorem: Introduction"
Used here as: "The Central Limit Theorem"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16953/1.17/
Pages: 127-128
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Central Limit Theorem: Central Limit Theorem for Sample Means"
Used here as: "The Central Limit Theorem for Sample Means (Averages)"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16947/1.23/
Pages: 128-130
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Central Limit Theorem: Central Limit Theorem for Sums"
Used here as: "The Central Limit Theorem for Sums"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16948/1.16/
Pages: 131-132
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Central Limit Theorem: Using the Central Limit Theorem"
Used here as: "Using the Central Limit Theorem"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16958/1.21/
Pages: 132-138
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Central Limit Theorem: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16956/1.8/
Page: 139
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Introduction"
Used here as: "Hypothesis Testing: Single Mean and Single Proportion"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16997/1.11/
Pages: 141-142
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Null and Alternate Hypotheses"
Used here as: "Null and Alternate Hypotheses"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16998/1.14/
Pages: 142-143
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Outcomes and the Type I and Type II Errors"
Used here as: "Outcomes and the Type I and Type II Errors"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17006/1.8/
Pages: 143-144
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Distribution Needed for Hypothesis Testing"
Used here as: "Distribution Needed for Hypothesis Testing"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17017/1.13/
Pages: 144-145
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Assumptions"
Used here as: "Assumption"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17002/1.16/
Page: 145
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Rare Events"
Used here as: "Rare Events"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16994/1.8/
Page: 145
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Using the Sample to Test the Null Hypothesis"
Used here as: "Using the Sample to Support One of the Hypotheses"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16995/1.17/
Pages: 146-147
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Decision and Conclusion"
Used here as: "Decision and Conclusion"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16992/1.11/
Page: 147
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Additional Information"
Used here as: "Additional Information"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16999/1.13/
Pages: 147-148
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Summary of the Hypothesis Test"
Used here as: "Summary of the Hypothesis Test"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16993/1.6/
Page: 149
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing of Single Mean and Single Proportion: Lab"
Used here as: "Lab: Hypothesis Testing of a Single Mean and Single Proportion"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17007/1.12/
Pages: 150-153
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Introduction"
Used here as: "Hypothesis Testing: Two Population Means and Two Population Proportions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17029/1.9/
Pages: 155-156
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Means with Unknown Population Standard Deviations"
Used here as: "Comparing Two Independent Population Means with Unknown Population Standard Deviations"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17025/1.18/
Pages: 156-159
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Means with Known Population Standard Deviations"
Used here as: "Comparing Two Independent Population Means with Known Population Standard Deviations"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17042/1.10/
Pages: 159-161
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Comparing Two Independent Population Proportions"
Used here as: "Comparing Two Independent Population Proportions"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17043/1.12/
Pages: 161-163
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Matched or Paired Samples"
Used here as: "Matched or Paired Samples"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17033/1.15/
Pages: 163-167
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Hypothesis Testing: Two Population Means and Two Population Proportions: Summary of Types of Hypothesis Tests"
Used here as: "Summary of Types of Hypothesis Tests"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17044/1.5/
Page: 168
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Confidence Intervals: Introduction"
Used here as: "Confidence Intervals"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16967/1.16/
Pages: 171-173
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Confidence Intervals: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal"
Used here as: "Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16962/1.23/
Pages: 173-180
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Confidence Intervals: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's-t"
Used here as: "Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16959/1.24/
Pages: 180-183
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Confidence Intervals: Confidence Interval for a Population Proportion"
Used here as: "Confidence Interval for a Population Proportion"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16963/1.20/
Pages: 183-187
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Confidence Intervals: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m16973/1.8/
Page: 188
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: Introduction"
Used here as: "F Distribution and ANOVA"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17065/1.11/
Page: 189
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: Purpose and Basic Assumptions of One-Way ANOVA"
Used here as: "ANOVA"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17068/1.10/
Pages: 190-191
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: The F Distribution And The F Ratio"
Used here as: "The F Distribution and the F Ratio"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17076/1.14/
Pages: 191-195
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: Facts About the F Distribution"
Used here as: "Facts About the F Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17062/1.14/
Pages: 195-199
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: Test of Two Variances"
Used here as: "Test of Two Variances"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17075/1.8/
Pages: 199-201
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "F Distribution and One-Way ANOVA: Summary"
Used here as: "Summary"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17072/1.4/
Page: 202
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Introduction"
Used here as: "The Chi-Square Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17048/1.9/
Pages: 205-206
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Notation"
Used here as: "Notation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17052/1.6/
Page: 206
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Facts About The Chi-Square Distribution"
Used here as: "Facts About the Chi-Square Distribution"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17045/1.6/
Pages: 206-207
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Goodness-of-Fit Test"
Used here as: "Goodness-of-Fit Test"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17192/1.8/
Pages: 208-215
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Test of Independence"
Used here as: "Test of Independence"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17191/1.12/
Pages: 215-219
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Test of a Single Variance"
Used here as: "Test of a Single Variance (Optional)"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17059/1.7/
Pages: 219-220
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "The Chi-Square Distribution: Summary of Formulas"
Used here as: "Summary of Formulas"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17058/1.8/
Page: 221
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Introduction"
Used here as: "Linear Regression and Correlation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17089/1.6/
Page: 223
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Linear Equations"
Used here as: "Linear Equations"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17086/1.4/
Pages: 223-225
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation"
Used here as: "Slope and Y-Intercept of a Linear Equation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17083/1.5/
Page: 225
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Module: "Linear Regression and Correlation: Scatter Plots"
Used here as: "Scatter Plots"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17082/1.8/
Pages: 226-227
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: The Regression Equation"
Used here as: "The Regression Equation"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17090/1.15/
Pages: 228-233
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Correlation Coefficient and Coefficient of Determination"
Used here as: "The Correlation Coefficient"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17092/1.12/
Pages: 234-235
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient"
Used here as: "Facts About the Correlation Coefficient for Linear Regression"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17077/1.15/
Pages: 236-240
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Prediction"
Used here as: "Prediction"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17095/1.8/
Page: 241
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Outliers"
Used here as: "Outliers"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17094/1.14/
Pages: 241-247
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: 95% Critical Values of the Sample Correlation Coefficient Table"
Used here as: "95% Critical Values of the Sample Correlation Coefficient Table"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17098/1.6/
Pages: 248-250
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/3.0/

Module: "Linear Regression and Correlation: Summary"
Used here as: "Summary"
By: Susan Dean, Barbara Illowsky, Ph.D.
URL: http://cnx.org/content/m17081/1.4/
Page: 251
Copyright: Maxfield Foundation
License: http://creativecommons.org/licenses/by/2.0/

Quantitative Information Analysis III

These chapters are a subset of Collaborative Statistics, written by Barbara Illowsky and Susan Dean, faculty members at De Anza College in Cupertino, California. The textbook was developed over several years and has been used in regular and honors-level classroom settings and in distance learning classes. This textbook is intended for introductory statistics courses being taken by students at two– and four–year colleges who are majoring in fields other than math or engineering. Intermediate algebra is the only prerequisite. The book focuses on applications of statistical knowledge rather than the theory behind it.

About Connexions

Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities.

Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai. Connexions is part of an exciting new information distribution system that allows for **Print on Demand Books**. Connexions has partnered with innovative on-demand publisher QOOP to accelerate the delivery of printed course materials and textbooks into classrooms worldwide at lower prices than traditional academic publishers.