Chapter 1: Collecting and Analyzing Data

Vocabulary

Data: Information in all forms.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

Introduction: The goal of collecting and analyzing data is to understand the world around us. How data is collected is very important. The goal of collecting data is to get “unbiased” data that represents the population. Analyzing biased data may result in incorrect conclusions and lead to a misguided view of the world around us. It is also important to have a goal in mind when you collect data. Are we trying to find a population percentage from categorical data or a population average from quantitative data? Are we trying to show if two variables are related or are we trying to show cause and effect? Data needs to be collected differently depending on what goal you have in mind.

Section 1A – Two Types of Data – Categorical and Quantitative

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the type of data.

There are two general types of data, categorical and quantitative.

Categorical Data

Categorical data (or qualitative data) are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person’s occupation, or what kind of pet they have?

Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used as a category. For example, a zip code can be used instead of the place a person lives. The numbers “1” and “2” may be used instead of yes and no.

Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people’s heights in inches, or their weights in kilograms, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of dogs in animal shelters in Los Angeles.

Numbers used as categories

Remember, not all numeric data is quantitative. Ask yourself if the numbers are measuring or counting something and if an average would make sense. For example, a list of people’s zip codes are numbers but an average zip code would not really tell us anything. In addition, identity numbers like hospital ID numbers, student ID numbers or social security numbers are not measuring anything and an average would not make sense in the context so they are not quantitative.
Practice Problems Section 1A

1. The following spreadsheet can be found on the website www.matt-teachout.org. Just click on the “statistics” tab and then “data sets”. This data was taken from bears. Use the bear data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data is categorical, list the different labels (variables) in that category.

<table>
<thead>
<tr>
<th>AGE (months)</th>
<th>Month Data Taken</th>
<th>Gender</th>
<th>Head Length (in)</th>
<th>Head Width (in)</th>
<th>Neck Circum (in)</th>
<th>Length (in)</th>
<th>Chest (in)</th>
<th>Weight (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>July</td>
<td>male</td>
<td>11</td>
<td>5.5</td>
<td>16</td>
<td>53</td>
<td>26</td>
<td>80</td>
</tr>
<tr>
<td>55</td>
<td>July</td>
<td>male</td>
<td>16.5</td>
<td>9</td>
<td>28</td>
<td>67.5</td>
<td>45</td>
<td>544</td>
</tr>
<tr>
<td>81</td>
<td>September</td>
<td>male</td>
<td>15.5</td>
<td>8</td>
<td>31</td>
<td>72</td>
<td>54</td>
<td>416</td>
</tr>
<tr>
<td>115</td>
<td>July</td>
<td>male</td>
<td>17</td>
<td>10</td>
<td>31.5</td>
<td>72</td>
<td>49</td>
<td>348</td>
</tr>
<tr>
<td>104</td>
<td>August</td>
<td>female</td>
<td>15.5</td>
<td>6.5</td>
<td>22</td>
<td>62</td>
<td>35</td>
<td>166</td>
</tr>
<tr>
<td>100</td>
<td>April</td>
<td>female</td>
<td>13</td>
<td>7</td>
<td>21</td>
<td>70</td>
<td>41</td>
<td>220</td>
</tr>
<tr>
<td>56</td>
<td>July</td>
<td>male</td>
<td>15</td>
<td>7.5</td>
<td>25.5</td>
<td>75.5</td>
<td>41</td>
<td>262</td>
</tr>
<tr>
<td>51</td>
<td>April</td>
<td>male</td>
<td>13.5</td>
<td>6</td>
<td>27</td>
<td>68.5</td>
<td>49</td>
<td>360</td>
</tr>
<tr>
<td>57</td>
<td>September</td>
<td>female</td>
<td>15.5</td>
<td>7</td>
<td>20</td>
<td>64</td>
<td>58</td>
<td>204</td>
</tr>
<tr>
<td>53</td>
<td>May</td>
<td>female</td>
<td>12.5</td>
<td>6</td>
<td>18</td>
<td>58</td>
<td>51</td>
<td>144</td>
</tr>
<tr>
<td>68</td>
<td>August</td>
<td>male</td>
<td>16</td>
<td>9</td>
<td>29</td>
<td>73</td>
<td>44</td>
<td>392</td>
</tr>
<tr>
<td>8</td>
<td>August</td>
<td>male</td>
<td>9</td>
<td>4.5</td>
<td>13</td>
<td>37</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>44</td>
<td>August</td>
<td>female</td>
<td>12.5</td>
<td>4.5</td>
<td>10.5</td>
<td>83</td>
<td>32</td>
<td>140</td>
</tr>
<tr>
<td>52</td>
<td>August</td>
<td>male</td>
<td>14</td>
<td>5</td>
<td>21.5</td>
<td>67</td>
<td>57</td>
<td>160</td>
</tr>
<tr>
<td>20</td>
<td>August</td>
<td>female</td>
<td>11.5</td>
<td>5</td>
<td>17.5</td>
<td>52</td>
<td>29</td>
<td>105</td>
</tr>
<tr>
<td>22</td>
<td>August</td>
<td>male</td>
<td>13</td>
<td>6</td>
<td>21.5</td>
<td>59</td>
<td>33</td>
<td>166</td>
</tr>
<tr>
<td>45</td>
<td>September</td>
<td>male</td>
<td>13.5</td>
<td>7</td>
<td>24</td>
<td>64</td>
<td>39</td>
<td>204</td>
</tr>
<tr>
<td>9</td>
<td>September</td>
<td>female</td>
<td>9</td>
<td>4.5</td>
<td>12</td>
<td>36</td>
<td>19</td>
<td>26</td>
</tr>
<tr>
<td>21</td>
<td>September</td>
<td>male</td>
<td>13</td>
<td>6</td>
<td>19</td>
<td>59</td>
<td>30</td>
<td>120</td>
</tr>
<tr>
<td>177</td>
<td>September</td>
<td>male</td>
<td>16</td>
<td>9.5</td>
<td>30</td>
<td>72</td>
<td>48</td>
<td>436</td>
</tr>
<tr>
<td>57</td>
<td>September</td>
<td>female</td>
<td>12.5</td>
<td>5</td>
<td>19</td>
<td>57.5</td>
<td>52</td>
<td>125</td>
</tr>
<tr>
<td>61</td>
<td>September</td>
<td>female</td>
<td>13</td>
<td>5</td>
<td>20</td>
<td>61</td>
<td>33</td>
<td>132</td>
</tr>
<tr>
<td>21</td>
<td>September</td>
<td>male</td>
<td>13</td>
<td>5</td>
<td>17</td>
<td>54</td>
<td>28</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>September</td>
<td>male</td>
<td>10</td>
<td>4</td>
<td>13</td>
<td>40</td>
<td>23</td>
<td>40</td>
</tr>
<tr>
<td>45</td>
<td>September</td>
<td>male</td>
<td>16</td>
<td>6</td>
<td>24</td>
<td>63</td>
<td>42</td>
<td>220</td>
</tr>
<tr>
<td>9</td>
<td>September</td>
<td>male</td>
<td>10</td>
<td>4</td>
<td>13.5</td>
<td>43</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>53</td>
<td>September</td>
<td>male</td>
<td>12.5</td>
<td>6</td>
<td>22</td>
<td>66.5</td>
<td>54</td>
<td>154</td>
</tr>
<tr>
<td>57</td>
<td>September</td>
<td>female</td>
<td>13</td>
<td>5.5</td>
<td>17.5</td>
<td>80.5</td>
<td>31</td>
<td>136</td>
</tr>
<tr>
<td>45</td>
<td>September</td>
<td>female</td>
<td>13</td>
<td>6.5</td>
<td>21</td>
<td>60</td>
<td>34</td>
<td>182</td>
</tr>
<tr>
<td>21</td>
<td>September</td>
<td>female</td>
<td>14.5</td>
<td>5.5</td>
<td>20</td>
<td>61</td>
<td>34</td>
<td>150</td>
</tr>
<tr>
<td>10</td>
<td>October</td>
<td>male</td>
<td>9.5</td>
<td>4.5</td>
<td>16</td>
<td>40</td>
<td>26</td>
<td>65</td>
</tr>
<tr>
<td>62</td>
<td>October</td>
<td>female</td>
<td>15.5</td>
<td>6.5</td>
<td>28</td>
<td>64</td>
<td>48</td>
<td>356</td>
</tr>
<tr>
<td>70</td>
<td>October</td>
<td>female</td>
<td>14.5</td>
<td>6.5</td>
<td>26</td>
<td>65</td>
<td>46</td>
<td>316</td>
</tr>
<tr>
<td>10</td>
<td>October</td>
<td>male</td>
<td>11</td>
<td>5</td>
<td>17</td>
<td>49</td>
<td>28</td>
<td>94</td>
</tr>
<tr>
<td>10</td>
<td>October</td>
<td>male</td>
<td>11.5</td>
<td>5</td>
<td>17</td>
<td>47</td>
<td>29.5</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>October</td>
<td>male</td>
<td>13</td>
<td>7</td>
<td>21</td>
<td>59</td>
<td>35</td>
<td>150</td>
</tr>
<tr>
<td>54</td>
<td>October</td>
<td>male</td>
<td>16.5</td>
<td>6.5</td>
<td>27</td>
<td>72</td>
<td>44.5</td>
<td>270</td>
</tr>
<tr>
<td>84</td>
<td>October</td>
<td>male</td>
<td>14</td>
<td>5.5</td>
<td>24</td>
<td>85</td>
<td>39</td>
<td>202</td>
</tr>
<tr>
<td>58</td>
<td>October</td>
<td>female</td>
<td>15.5</td>
<td>6.5</td>
<td>21.5</td>
<td>63</td>
<td>40</td>
<td>202</td>
</tr>
<tr>
<td>58</td>
<td>October</td>
<td>male</td>
<td>15.5</td>
<td>7</td>
<td>28</td>
<td>70.5</td>
<td>50</td>
<td>265</td>
</tr>
<tr>
<td>11</td>
<td>November</td>
<td>male</td>
<td>11.5</td>
<td>6</td>
<td>15.5</td>
<td>48</td>
<td>31</td>
<td>79</td>
</tr>
<tr>
<td>23</td>
<td>November</td>
<td>male</td>
<td>12</td>
<td>6.5</td>
<td>19</td>
<td>50</td>
<td>38</td>
<td>148</td>
</tr>
<tr>
<td>70</td>
<td>October</td>
<td>male</td>
<td>15.5</td>
<td>7</td>
<td>28</td>
<td>76.5</td>
<td>55</td>
<td>446</td>
</tr>
<tr>
<td>11</td>
<td>November</td>
<td>female</td>
<td>9</td>
<td>5</td>
<td>15</td>
<td>45</td>
<td>27</td>
<td>62</td>
</tr>
<tr>
<td>83</td>
<td>November</td>
<td>female</td>
<td>14.5</td>
<td>7</td>
<td>23</td>
<td>61.5</td>
<td>44</td>
<td>236</td>
</tr>
<tr>
<td>35</td>
<td>November</td>
<td>male</td>
<td>13.5</td>
<td>8.5</td>
<td>23</td>
<td>63.5</td>
<td>44</td>
<td>212</td>
</tr>
<tr>
<td>16</td>
<td>April</td>
<td>male</td>
<td>10</td>
<td>4</td>
<td>15.5</td>
<td>48</td>
<td>26</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>April</td>
<td>male</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>41</td>
<td>26</td>
<td>64</td>
</tr>
<tr>
<td>17</td>
<td>May</td>
<td>male</td>
<td>11.5</td>
<td>5</td>
<td>17</td>
<td>53</td>
<td>30.5</td>
<td>114</td>
</tr>
<tr>
<td>17</td>
<td>May</td>
<td>female</td>
<td>11.5</td>
<td>5</td>
<td>15</td>
<td>52.5</td>
<td>28</td>
<td>76</td>
</tr>
<tr>
<td>17</td>
<td>May</td>
<td>female</td>
<td>11</td>
<td>4.5</td>
<td>13</td>
<td>46</td>
<td>23</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>August</td>
<td>female</td>
<td>10</td>
<td>4.5</td>
<td>10</td>
<td>43.5</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>83</td>
<td>November</td>
<td>male</td>
<td>15.5</td>
<td>8</td>
<td>30.5</td>
<td>75</td>
<td>54</td>
<td>514</td>
</tr>
<tr>
<td>18</td>
<td>June</td>
<td>male</td>
<td>12.5</td>
<td>8.5</td>
<td>18</td>
<td>57.3</td>
<td>32.8</td>
<td>140</td>
</tr>
</tbody>
</table>
2. The following spreadsheet can be found on the website www.matt-teachout.org. Just click on the “statistics” tab and then “data sets”. This data was taken from various cereals. Use the cereal data to classify each column of data as categorical or quantitative. If the data is quantitative, what are the units? If the data is categorical, list the different labels (variables) in that category.

![Spreadsheet Image]

Section 1B – Collecting Data

Vocabulary

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

One of the most important goals in data science is to learn about the world around us (populations). It is very difficult to understand populations sometimes because data may be biased and not reflect the population very well. Bias can occur in many different ways, but certain ways people collect data have more bias than others do.

Using a method for collecting data that increases bias is sometimes called “sampling bias”.

It is important to be aware of various methods used to collect data, the good and the bad.

Method 1: Census

A census is the best way to collect data if it is possible. If our goal is to learn about the population, it makes sense to collect data from everyone in the population. There are ways for a census to be biased, but in terms of the collecting method, a census is the best. Unfortunately, it is almost impossible to collect a census if your population is large. Most statisticians and data scientists are only able to collect a sample, data collected from a small subgroup of the population.
Method 2: Simple Random Sample

If a statistician or data scientist cannot collect a census, the preferred method is to collect a random sample. A random sample is one where everyone in the population has a chance to be in the sample, so it tends to represent the population better than other non-random samples. It is nowhere near as good as a census, but as I said, a census is usually not possible.

A simple random sample is one where individuals in the population are selected randomly. This can be a difficult process. The usual method is to assign everyone in a population a number and then use a random number generator in a computer program to pick random numbers. Computer programs have many built in randomization functions for this purpose. If you have a spreadsheet of the entire population, a computer can also randomly select individuals from the list. The key with a “simple random sample” is that you are selecting people or objects one at a time. Collecting data randomly and one at a time gives greater flexibility to your sample. Almost any grouping is possible with a simple random sample, so it tends to represent populations better than other samples.

There are many examples of a simple random sample. Many statistics companies use a random phone number generator that randomly gives phone numbers. They then call the phone numbers randomly chosen and try to get information from people that answer the phone. The U.S. government may have a computer randomly select social security numbers to select individuals for a sample. A company may have a computer randomly select employee ID numbers to select individuals for a sample.

Method 3: Convenience Sample

People often find collecting a census or a simple random sample difficult, so they chose to collect data in whatever way seems easiest. A sample collected this way is often called a “convenience sample” and is popular with people not trained in statistics. A convenience sample usually has much more bias than a random sample and may not represent the population very well.

An example of a convenience sample is collecting data from your friends and family. This is fine if your population of interest is your friends and family, but will by no means represent a large population. Another example might be standing outside of a store or post office and collecting data from people that leave the store. Beginning statistics students may walk into a mall and collect data from whomever they bump into. They mistakenly think that these are random samples, but they are not. A random sample means everyone in the population has a chance to be included in the sample. Not everyone in the population has a chance to bump into you at a mall or come out of a store at 2:30 pm on a Tuesday afternoon. These are convenience samples and generally do not reflect the population very well.

Method 4: Voluntary Response Sample

Some say that all surveys are bad, but that is not the case. A survey is just a form to collect data from people. When a company takes a census of all its employees, it may require all of the employees to fill out a survey. That is a census. As long as no other forms of bias creep into the data, a census will probably be a very good representation of the population. The point is that giving a survey is not the issue. The issue is who you give the survey to and who is allowed to fill out the survey.

A voluntary response sample puts a survey out into the world and allow anyone to fill it out. The usual method used today is to put a survey on a website and allow anyone that comes across the survey to fill it out. The survey can also be mailed to every address in a given population. Again, those that fill it out self-select themselves to be in our data.
On the surface, a voluntary response sample may seem like a good way of collecting data. It usually gives a large amount of data. Also, wouldn't everyone in the population have a chance to fill it out? It turns out the answer is no. Ask yourself the following question. When you are surfing the web and a survey pops up, do you fill it out? I have been asking my statistics classes that question for years and rarely have anyone that says that they do fill out surveys. The key problem is that only certain types of people will fill out a survey voluntarily. It may be a person who is bored and has nothing better to do. It is certainly not a person with three children, working a full time job and going to college full time. It may also be a person who is upset by or feels very passionate about the topic in the voluntary response survey. They are so upset by the lack of pay for teachers that they are willing to fill out a survey to tell you what they think. The point is that voluntary response surveys tend to over-sample people that are bored or upset and under-sample everyone else. For this reason, voluntary response samples can be very biased and may not represent the population very well.

I have had many students ask me if sample size is important. Isn’t a voluntary response sample of five thousand people better than a random sample of fifty people?” I would tell them that though sample size is important, method is important also. The voluntary response sample of five thousand would tend to over-represent people that are bored or upset about the topic. It does not represent typical people in the population. The random sample of fifty people, while a small sample size, at least does not have that bias.

Method 5: Cluster Sample

A cluster sample is one where data is collected from groups of people in a population instead of one at a time. For example, a company that has 250 stores worldwide might have a computer randomly select ten stores and get data from those people that work at those ten stores. Notice this would be a random sample since every employee has a chance to be in the data. If their store was chosen, then they will be included in the sample. This is not a simple random sample however, since they are not choosing one at a time. This example is sometimes called a “random cluster sample”. While it is a good method for collecting data, it has less flexibility than a simple random sample. Think of it this way. In a simple random sample, any grouping is possible, but in this random cluster example, only groups of people that work at the same store can be chosen. It is still a random sample though, and would tend to be more representative of the population than non-random samples like convenience or voluntary response.

It is good to note that the goal of a cluster sample should be to choose the groups of people randomly. If we choose groups of people that are convenient to collect data from, our cluster sample will have more sampling bias and will not represent the population nearly as well.

Method 6: Stratified Sample

One of the most common studies done in statistics is to compare groups. We may compare data from 2016 to data from this year. We may compare people living in Canada to people living in Australia. To compare groups, you need to collect a stratified sample.

Some people in statistics explain a stratified sample as comparing two or more groups in one population. I like to think of it as comparing two or more populations. Whether you explain a stratified sample as comparing groups in one population or comparing populations, the key is that you are comparing.

For example, we may want to compare the percentage of adults in the U.S. with diabetes to the percentage of children in the U.S. with diabetes. Some statistics authors think of this as comparing adults and children in the one population of all people in the U.S. I like to think of it as comparing the population of U.S. adults to the population of U.S. children.
Another example may be to compare the mean average salary of people working in London, England to the mean average salary of people working in Toronto, Canada. Again, a stratified sample is needed because we are comparing.

To do a stratified sample, we often take a simple random sample from each group. I like to think of it as taking a simple random sample from each population you want to compare. In the previous example, we may collect a simple random sample of adults in the U.S. and another simple random sample of children in the U.S. We then can calculate the sample percentages that are diabetic from each sample and use statistical methods to compare them. For the salary example, we can collect a simple random sample of salaries for people working in London, and another simple random sample for people working in Toronto. The goal is then to use statistical methods to compare the mean average salaries.

It should be noted that when taking a stratified sample, we should use randomization. Again, if we just take a convenience sample from each group or voluntary response sample from each group, we will likely have a lot more bias and the data will not reflect the population (or populations) as well as we would like.

Many people confuse a cluster sample with a stratified sample because they both involve groups. The goal of a cluster is to get data on and analyze one population, not to compare. You are just collecting data from groups of people from that one population instead of one at a time. The goal of a stratified sample is to compare two or more populations so we need to collect data from each population.

Method 7: Systematic Sample

A systematic sample is one where we use a system to collect the sample. Usually it involves collecting data from every fifth person that comes in your store or every twentieth person on a list.

For example, let us suppose we want to collect a sample of students from our college. We could look at an alphabetical list of the names of all students that attend our college and then chose every 50th person on the list. Is this a random data set? Ask yourself this question. Does everyone on this list have a chance to be chosen? No. Only the 50th, 100th, 150th, 200th and so forth have a chance. People from 1-49 have no chance. People from 51-99 have no chance. Therefore, it is not a random sample. This may not be random, but we may make the argument that it is representative of the population. This method would have less bias than convenience or voluntary response samples. There is a way to incorporate randomization into the method. Many data scientists have a computer chose a random number between 1 and 50. Suppose it is 33. Then they collect data from the 33rd person on the alphabetical list. Now, from there use the system of choosing every 50th person. Therefore, they would choose the 33rd person, then the 83rd person, then the 133rd person and so on. Making the first choice random, makes the whole data set random, because everyone on the list now has a chance to be chosen.

Summary

So let's summarize the various methods.

- An unbiased census is the best way to collect data to represent a population, because we are collecting data from everyone in the population.
- If you cannot do a census, then use a random sample of some sort. It may be a simple random sample, random cluster, or a random systematic sample. The main thing is that if you are collecting a sample, randomization needs to be involved.
- Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.
Practice Problems Section 1B

Directions: For each of the following, identify the population of interest. Then identify the method used to collect the data (census, systematic, convenience, voluntary response, cluster, stratified, or simple random). Explain why you chose your answer and if the method will represent the population of interest or not?

1. The admissions department at a college wants to see how many of their students would be in favor of using a new program to register for classes. They put a link on their website so that any students that want to try out the program can. The students can then take a survey and say how well they like the new system.

2. Rick works for a sports equipment manufacturing company. He wants to compare the opinion of his older employees to the new employees. To do this, he separates all the employees into two groups, employees that have been with the company five or more years and those that have been with the company less than five years. He then chooses 12 of his most trusted older employees and 16 new employees that have proven themselves and ask what they think about changing the medical insurance coverage.

3. Michelle, a teacher at a local high school, wants to see how many students at her high school will be attending community college. She gives the students in her one section of advanced placement U.S. History a questionnaire to fill out that asks where they will be attending college.

4. Jamie is working at the Republican recruiting committee in her city. She is curious how many people that live in her city will vote for the Republican candidate in the next election. She uses a computer to randomly select phone numbers in her city. She then calls those phone numbers to ask people about their voting preferences.

5. Rachael works at the Democrat recruiting center in her hometown. To determine what percent of people will vote for the Democratic candidate, she obtains a list of all residents in her town and decides to ask every 40th person on the list.

6. Laya is passionate about bringing an NFL football team to her city. She needs to take an opinion poll about how people in her city would feel about raising taxes in order to build a stadium for a professional football team. She randomly selects 75 streets in her city and asks every person living on those streets.

7. Micah is the CEO of a large software development company. He wants to see if his employees have any ideas about areas of software development that the company should pursue. He has every single employee in his company fill out a questionnaire outlining his or her ideas. He gives the employees a stipend on their paycheck to pay them for their time it took to fill out the questionnaire.

Section 1C – Bias

Vocabulary

Population: The collection of all people or objects you want to study.

Bias: When data does not reflect the population.
The purpose of collecting data is to learn about the world around us, to learn about populations. The problem is that many people that collect data may not have had any training in Statistics or Data Science. The result is that many data sets collected do not reflect the population very well. When this happens, we say that the data is biased.

Many people think that if you collect a random sample or a census, it will guarantee that you will have an unbiased data set. This is not true. There are many types of bias and it is possible to have a census or a random sample that does not reflect the population very well. It is critical that we be aware of these other forms of bias and to try our best to make sure they are not incorporated into our data sets.

**Sampling Bias**

In the last section, we said that the best way to collect data is a census. This means that we collected data from everyone in the population. If we cannot collect a census then we should try to collect a random sample or at least a sample that represents the population. We said that convenience samples or voluntary response samples are inherently biased and usually do not reflect populations very well. Using a bad data collecting method like convenience or voluntary response is called sampling bias. This means the technique for collecting the data is poor.

**Question Bias**

It has been said that there are lies, bad lies, and then there is statistics. There is some truth in this. People with specific agendas may twist data and statistical analysis to suit their purpose. One way to do this is question bias.

A question bias occurs when someone phrases a question in a specific way to force people to answer the way they want.

For example, suppose a politician wants to show that most people in her city agree with her policy on raising taxes to improve health care. She may collect a great simple random sample, but ask the question this way.

“Health care in our city is extremely bad. Hospitals and urgent cares are in bad need of renovation and need better supplies. The elderly need to know that we have not forgotten them. We need to improve the quality of care for our children. Will you support my policy for improving health care across our city?”

Phrasing the question this way, no one would guess that the real issue was whether to raise taxes. People, hearing this question, think about helping the children and elderly, not about taxes. When a large percentage of people answer that they support her plan, she now has data to support her agenda.

When you collect data, you want to ask questions in a neutral way that does not attempt to sway people in one direction or another. It also should not leave out key information like what the real question is. If the politician had simply asked people in the simple random sample if they would be in favor of raising taxes to improve health care, she likely would have gotten a much smaller percentage of people to agree.

Notice that in this example, the data was a simple random sample. This is a good data collection method, as methods go. However, the incorporation of a question bias into the data makes the data very bad. This simple random sample does not reflect the population at all. The data has been manipulated to support an agenda.
Response Bias

Many topics are very difficult to get data on because people do not feel comfortable answering truthfully. If you ask people if they are addicted to alcohol or drugs, they are likely to deny it even if they do struggle with substance addiction. People may lie about their age, weight, or salary. When a large percentage of people in your data lie, you have a response bias in your data.

Suppose a church wants to collect data on how many hours per week their congregation spends helping the homeless. They decide to have every person in their congregation fill out a survey listing how many hours per week they help the homeless. Remember a census is usually the best way to collect data about a population, but this census has a problem. It is a topic that people are likely to lie about. People may put a higher number of hours on the survey than they really do so that they will not look bad to the church leaders. The average number of hours calculated from this data will likely be larger than the population average number of hours. Even though this is a census, it probably does not reflect the population very well.

When dealing with topics that people are likely to lie about, the data scientist needs to have a plan to deal with the response bias. Instead of asking people their weights, maybe they weigh them on a scale. Instead of asking people about their salary, maybe they look at paycheck stubs. Instead of asking people about substance abuse, they may collect data from agencies that support people with addiction.

Deliberate Bias

We have stated already that people may misuse statistics and data in order to support their agenda. Deliberate bias is another example of this. Deliberate bias can take on a variety of forms. It could be someone deliberately leaving out groups from the data. The most common is collecting data and then leaving out the data of people that disagreed with you. It can also be deliberately lying about the results of the data report. Maybe the data makes your restaurant or hospital or school look bad, so people just falsify their records and deliberately lie about the results of the study. The data may be census or a random sample put the conclusions have been falsified and the data distorted.

Deliberate bias is a major problem in statistics. It is also a good reason to have an independent statistics company collect the data and do the analysis. Use a statistics company that is not tied to the government, business, hospital, restaurant or politician in question. An independent statistics company is less likely to lie about the results or to falsify the data, though it is naive to think that it never happens.

I tend to be suspicious about internal statistics reports that come out where the company, government or politician refuses to share the data. We are supposed to take their word for it and agree with the findings. There are good reasons why companies do not share data, but I always wonder if they are they afraid that someone analyzing that data would come to a very different conclusion?

There is large worldwide discussion of ethics for people that work in the fields of statistics or data science. Statistical analysis is a powerful tool and is a vital discipline to understand and improve the world around us, but falsifying records or manipulating data should never be an option. It is not only unethical, but also makes people question the integrity of our science.

Sometimes specific groups in the population may not be represented very well in the data. This also falls under the umbrella of deliberate bias. For example, a person may wish to collect data on adults living in a city, but only collected data from people living in the wealthier areas of that city. It may not have been done deliberately. It could just be that the person collecting the data did not think about certain groups in the population that are not being represented. In large cities, the homeless are often difficult to get data on. A person collecting data has to have a plan for getting data that will represent all the groups in their population, including the homeless.
Non-response Bias

Non-response bias is becoming a huge problem for all people that collect data. A computer may randomly select people to collect data from, but more often than not, the person does not want to participate. They may fear identity theft or are just too busy to participate. It is a huge problem. We need data. We need to understand the world around us, but it now becoming increasing difficult to get unbiased data. Many people that collect data report that sometimes only one in every five randomly selected people will participate and give data. The problem of non-response bias continues to get worse. This makes us consider what type of person gives data and if that person is truly reflective of all people in the population.

To combat the problem of non-response bias, many people that collect data offer a reward system for people that will participate and give data. This may help a little, but then offering a reward may incorporate its own bias into the data.

Summary

There are many reasons why data may not reflect a population. It is a mistake to think that a random sample or a census will always be devoid of bias. It is increasingly important to be aware of possible sources of bias and to strive to keep them out of our data as much as possible. The goal of data collecting is to collect unbiased data that reflects the population. Always phrase questions in a neutral way that avoids question bias. Have a plan for collecting data about topics where people are likely to lie. We have to have a good plan on how we will collect data. It should be a census or a random sample, but we should also think about groups that may not be represented. We need to avoid deliberate bias and never falsify reports or distort data to support someone’s agenda.

Practice Problems Section 1C

1. Define each of the following types of bias and give an example of each.
   A. Question Bias
   B. Response Bias
   C. Sampling Bias
   D. Deliberate Bias
   E. Non-response Bias

Directions for #2-7: For each of the following scenarios, describe the population of interest and all of the types of bias that the data may have (Question, Response, Sampling, Deliberate or Non-response). There may be more than one type of bias involved. Explain your answers and if there is bias, what groups of people were not represented.

2. We are interested in calculating the percent of children in LA County that are up to date with their vaccines. To figure this out, a person put a survey up on the yahoo webpage asking the following question: “Is your child up to date with their vaccines?” The computer will keep track of the number of people that answer yes or no.
3. We are interested in finding what percent of people in the U.S. agree or disagree with vaccinating children. To figure this out, we randomly selected 350 people in the U.S. and asked them the following question: “In order to save children from devastating diseases, do you agree that all children should be vaccinated?”

4. We are interested in finding out how many people in the U.S. have had whooping cough this year. To figure this out, we called every major hospital in the United States and asked how many people at their hospital were diagnosed with whooping cough this year.

5. We are interested in finding out what percent of Americans use Cocaine. We randomly chose 400 Americans and asked them if they use Cocaine or not.

6. What is the average age of college students in Canada? Since my cousin lives in Canada, I asked him to drive to two colleges near his house and ask people he bumps into what their age is.

7. Julie is interested in calculating the yearly income of adults in Palmdale. She drives around Palmdale, stops at certain streets, and then asks people that live on that street what their yearly income is? She skips streets that look “sketchy” as she is worried about her safety.

---

Section 1D – Experimental Design

Vocabulary

Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable causes the effect.

Response Variable: The dependent variable. In an experiment this is the variable that measures the effect.

Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

In statistics, we often want to determine if there is a relationship or association between two variables. We also may want to measure the strength of the relationship. For example, we may want to know if there is a relationship between blood pressure and heart rate. We may want to see if living in tropical climates is associated with having nut allergies.
In order to show that two variables are related or associated we use an observational study. We would collect data and use statistical methods to analyze and measure the strength of the relationship. However, showing that two variables are related does not prove that one causes the other.

**Association ≠ Causation!!!!**

Why?

Let us suppose that we have shown that there is a strong relationship between drinking alcohol and getting into a car accident. This tells us that alcohol consumption is an important factor to be considered when studying car accidents. However, this does not prove that drinking alcohol causes car accidents. There are many factors that go into having a car accident besides how much alcohol they consume. Can you name a few?

Other factors that may influence having a car accident besides alcohol: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like texting, eating or changing a radio station), using drugs, ...

These are called “confounding variables”. Confounding variables are factors that might influence your response variable other than the explanatory variable you are studying. In this case, factors that might influence having a car accident other than how much alcohol the driver consumed. Some statistics books call these “confounding variables” or “lurking variables”.

Note: The explanatory variable (alcohol consumption) is not a confounding variable. Alcohol is the explanatory variable we were studying. Confounding variables are factors other than alcohol that might influence the response (car accident).

Here is the point. If many variables were involved in having a car accident, it would be wrong to say that the alcohol was solely responsible for the car accident. Alcohol is just one of many factors involved. We have shown that drinking alcohol is related but we have not proven cause and effect. To prove cause and effect we need to deal with the confounding variables.

**Experimental Design**

So how do we prove cause and effect? It is difficult. You would need to prove that none of the confounding variables are involved and it is only the explanatory variable that is causing the response. The key is controlling the confounding variables. Thankfully, scientists have put a great deal of thought into this process of controlling confounding variables and proving cause and effect. We call this process “experimental design”.

Experimental design is a scientific method for controlling confounding variables and proving cause and effect. A key component to experimental design is the creation of similar groups through random assignment.

To control confounding variables, we will need to create two or more groups of people or objects that are very alike. One way to do this is by “random assignment”. Random assignment is a process where you take a group of people or objects and randomly split them into two or more groups. The randomly assigned groups tend to be very similar. If we do not think the groups are similar enough, we can use techniques like blocking or direct control to make the groups even more alike.

Another way to make alike groups is to use the same group of people twice. Think about it. The two groups would be perfectly alike. They would have the same ages, same amount of stress, same genetics, same blood pressures and the same jobs.
Example

Let's look at the previous example. How do we prove that drinking alcohol does cause car accidents?

Explanatory (treatment) Variable: Drinking alcohol or not

Response Variable (what we will measure): Did the person get into a car accident or not?

So how do we set up an experiment to prove that drinking alcohol causes car accidents? The first thing is to list out your possible confounding variables.

Possible Confounding Variables: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like passengers, texting, eating or changing a radio station), other drugs, gender, race, genetics

To control the confounding variables, we need to create two groups of people. The two groups should be the same (or at least as similar as possible) in all areas that the confounding variables address. Therefore, the groups should have similar ages, similar driving experience, similar cars and car condition, similar road conditions and similar distractions, similar genders, similar race and ethnicity, similar genetics and reflexes.

There are two ways to go about this. Let's suppose we have a group of 80 adult paid volunteers to conduct this experiment. One option would be to randomly put the volunteers into two groups and try to make the groups as similar as possible. A better option in this case would be to use the same people twice.

We had the people in the experiment drive an obstacle course sober. They must have no alcohol or other drugs in their system. They all used the same car on the same track with the same weather. The course was designed with cones and we will monitor how many cones the people hit. They all were not allowed to have any other person in the car. There was no other distractions as radios and phones were not allowed. We will monitor how many car accidents they had by checking how many cones they hit.

Now we will have all the people drink a certain amount of alcohol and then drive the course again. It is important to see that the alcohol (treatment) group was made up of exactly the same people as the sober (control) group. The response variable we measured was the number of cones they hit.

Conclusion

The results found that the alcohol group hit significantly more cones (significantly more car accidents) than the sober group. We have now proven that drinking alcohol causes car accidents.

Think about it. It cannot be the ages of the drivers or driving experience. The two groups had the exact same ages and the exact same driving experience. It cannot be gender, race, genetics, or reflexes. The two groups had the exact same genders, race, genetics, and reflexes. It cannot be drugs or other distractions like phones or radios. Neither group had drugs or any other distractions. If you notice, every one of the confounding variables is the same in the two groups. The only difference was that one group had alcohol and the other did not. So the only reason why the alcohol group had significantly more accidents is the alcohol. The experiment has proven that drinking alcohol causes car accidents.

Note: It is easy to confuse the two variables in an experiment with the two groups. They are not the same thing.

In this case, the explanatory variable is having alcohol or not. The response variable is the number of cones (accidents) the drivers had. The two groups are decided by those that have explanatory variable (alcohol) and those that do not. In this case, the two groups are the exact same people measured twice.

We usually call the group that has the explanatory variable the “treatment group” and the group that does not have the explanatory variable the “control group”.

This chapter is from Introduction to Statistics for Community College Students,
1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” Creative Commons Attribution 4.0 International license – 10/1/18
**Example 2**

A very common experiment is when a pharmaceutical company needs to prove that a medicine works. In the United States, pharmaceutical companies have to prove to the Food and Drug Administration (FDA) that their medicine has the effect it is supposed to and is relatively safe with few side effects.

Suppose a company has a new blood pressure medicine on the market and needs to prove to the FDA that taking it does decrease a person’s blood pressure. The company needs to prove cause and effect.

If we have to prove cause and effect, we need an experiment. The first step is to think about the possible confounding variables. What are some reasons why a person’s blood pressure might decrease other than taking this new medicine?

Possible Confounding Variables? Stress, Diet, Exercise, Genetics, Age, Gender, Race, Genetics, taking other medicines ...

To set up the experiment we need to create two groups of people that are similar in these areas. We start with a group of volunteers with high blood pressure that want to try out this new medicine. We randomly assign the people into two groups. Amazingly when scientists randomly assign people into two groups, the groups tend to be a lot alike. The two groups would have similar numbers of people in each race, similar number of males and females, similar numbers of stressed out people, similar numbers of people that exercise a lot or do not exercise. The people running the experiment can also exercise direct control and intentionally assign people to certain groups to make the groups even more alike.

Human Brain (placebo effect)

There is a problem with our experiment. If a person believes something is true, their brain can tell the body to manifest physical responses. We call this the “placebo effect”. Think of it this way. The group that thinks they are getting blood pressure medicine will not be as stressed out about it and their blood pressure may decrease slightly because of that belief. Similarly, the group that thinks they are not getting blood pressure medicine will be more stressed and worried and their blood pressure may increase because of that belief. In a sense, the human brain is a confounding variable that we need to control.

Placebo (fake medicine)

To control the placebo effect as a confounding variable, we need the groups to believe the same thing. One group cannot think they are getting medicine, and the other group cannot believe they are not getting medicine. So we introduce a placebo or fake medicine. The treatment group gets the real blood pressure medicine and the control group gets a fake medicine (placebo). No one in the experiment knows if he or she will be receiving real medicine or a placebo. Some may ask, “Won’t that make them more stressed and increase their blood pressure?” Yes. The key is that the two groups will be equally stressed and believe the same thing. That way we control the placebo effect.

For this to work, the people in the experiment cannot know if they are getting the medicine or a placebo. This is called “single blind”. When scientists first started using placebos, they were shocked to find that the people in the experiments somehow knew if it was a placebo. This defeated the whole purpose. It turned out they could tell by the body language of the person giving the medicine. The person giving the medicine tended to act differently if they were giving the real medicine versus a placebo. So the standard for an experiment about medicines is to use a “double blind” approach. A double blind experiment means that neither the people in the experiment, nor the people giving the medicine, know if it is a placebo or not. Someone knows though. The scientists keep very careful track of who receives a placebo and who receives the medicine. The person directly giving the medicine or placebo cannot know if it is a placebo or not.
Double blind works well. The people in the experiment no longer know if they are receiving a placebo or the real medicine. The experimental design has controlled the placebo effect.

Conclusion

Since we have controlled all of the confounding variables, the experiment has the possibility of proving cause and effect. We still need to see the blood pressures of both groups and make a conclusion. If the treatment group had a significantly lower average blood pressure than the control group, this would prove that taking the medicine does cause a person to have lower blood pressure. If the treatment group and control group have relatively the same average blood pressure, then we may conclude that the medicine is not effective in lowering blood pressure. This would be bad news for the pharmaceutical company. Deciding if one group is significantly higher than another can be very difficult. We will study confidence intervals, test statistics and P-value in later chapters to address this.

Summary

Use an experiment to control confounding variables and prove cause and effect. The groups in the experiment should be either the same people measured multiple times or separated by random assignment. The main idea is that the groups should be very similar in all areas that involve confounding variables. Experiments with medicines should be double blind with a placebo to control the placebo effect.

Use an observational study to see if there is a relationship (association) between two things. Remember observational studies do not control confounding variables, so cannot prove cause and effect.

How can I tell if a study is an experiment or not? Generally, look for random assignment. An experiment usually does not have a random sample of people from the population. The people in the experiment are usually volunteer. The volunteers are then randomly assigned into two or more groups. Random assignment means that they are not trying to apply something to the population, but instead are trying to use experimental design in order to prove cause and effect. If a study takes a random sample from the population, but does not randomly assign, it is probably just an observational study and cannot prove cause and effect.

Note: It should be noted that there are more complex forms of experiments than the types listed in this section. It may not be possible to randomly assign people into two groups. In that case, the scientist need to prove that each confounding variable is not involved. That is a more complex case that you may see in more advanced statistics classes.

Practice Problems Section 1D

Ruler Experiment Directions: Divide class into groups of three or four. Each group will need a ruler and their cell phones. It is best to stand up during this activity. Procedure: Student A will hold the cell phone in their dominant hand and then hold their non-dominant hand straight out in front of them with their hand curved. The fingers should not be very close or very far away from the thumb. While student A is texting, student B holds the bottom of the ruler up inside of student A’s non-dominate hand. Student B should hold the ruler from below student A’s hand. The top of student A’s hand should be about 5 inches on the ruler. Student B releases the ruler and student A tries to catch it. Student C records the number of inches on the top of the ruler before caught. Student C will take the catch length, subtract off the 5 inches, and then record the difference. If student A misses the ruler all together, then student C will just put “drop”. Each student should attempt to catch the ruler while texting three times.
Then repeat the process, but this time the students will attempt to catch the ruler with their non-dominant hand without a cell phone. Continue until all students have done the experiment three times without the cell phone and three times with the cell phones. Alternate the person releasing the ruler and the time before released. Collect the data for the “with phone” catches and drops in one column. In another column, collect the data for the “no phone” catches. When done, give the data to the instructor. Put the without cell phone/with cell phone data up on the board without names. The instructor or a student will collate the following results for the whole class: the mean average catch length with the cell, the mean average catch length without the cell, the total number of drops with the cell, the total number of drops without the cell.

Use your class data to answer the following questions as group. If you were absent on the day your class did the ruler experiment use the following data.

Ruler Experiment Data (Previous Class)

<table>
<thead>
<tr>
<th></th>
<th>With Phone</th>
<th>No Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Average Catch (inches)</td>
<td>10.3 inches</td>
<td>8.2 inches</td>
</tr>
<tr>
<td>Number of Drops</td>
<td>41 drops</td>
<td>7 drops</td>
</tr>
</tbody>
</table>

1. What is the explanatory (treatment) variable? What was the response variable?

2. Why did we bother to have the person catch the yardstick without the phone? Wouldn’t it of been quicker to just record the catching with the cell phone?

3. What were the two groups of people in the experiment? Were they alike? Why didn’t we randomly assign the groups?

4. What are some of the confounding variables in this experiment? What are some steps that we took to control these variables?

5. Was this experiment blind, double blind, or neither? How do you know?

6. What did the class data show? Does texting cause slow reflexes? How do you think this experiment might apply to driving while texting?

(#7-10) Directions: Determine if each of the following studies are an observational study or an experiment. Explain why. Can the study prove cause and effect or not? Why? If the study is an experiment, list some confounding variables that need to be controlled.

7. Dramamine is a common medication used in preventing and treating nausea, vomiting and dizziness caused by motion sickness. This medication has become a staple for thousands of people who travel by boat, car or plane. We need to prove that Dramamine is effective in preventing and treating the symptoms of motion sickness. Volunteers were randomly assigned into two groups. One group received Dramamine and the other received a placebo. The amount of motion was the same for all of the people. They were then asked to rank their motion sickness on a scale of 1 to 10.
8. Unemployment has become a very important topic in the United States and worldwide. We wish to understand how unemployment may be related to the tax rate. To shed light on this issue, we took a random sample of countries around the world and compared the average tax rate to the unemployment rate.

9. Tuberculosis (TB) is a disease that affects millions of people worldwide. TB is a contagious bacterial infection that affects the lungs. Doctors have long speculated that the percentage of people with Tuberculosis is higher in low income, crowded cities. A medical study was done to see if there is a relationship between low income, crowded cities and a high percentage of people with Tuberculosis. They took a random sample of cities and collected data about the size and the number of people. They then compared it to the number of cases of tuberculosis.

10. College students in the United States have long claimed that listening to music while studying causes them to retain information at a higher rate. We want to prove that this is not true. Listening to music while studying does not cause a person to retain information at a higher rate. We took a group of volunteer college students and randomly put them into three groups. The people in each group had to memorize the same information. They were ranked as high retention or low retention. One group had to listen to their favorite music, another group had to listen to a music they hated, and the third group had no music at all. The volume of music was the same for all of the people.

---

**Introduction to Categorical & Quantitative Data Analysis**

**Vocabulary**

**Data:** Information in all forms.

**Categorical data:** Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

**Quantitative data:** Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

**Population:** The collection of all people or objects to be studied.

**Census:** Collecting data from everyone in a population.

**Sample:** Collecting data from a small subgroup of the population.

**Statistic:** A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

**Parameter:** A population value, which is sometimes calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be. For example, a population mean average or a population percentage.

**Introduction**
We learned that, in order to learn about the world around us, we need to collect and analyze data. Our goal is to understand populations. Sometimes we can collect data from everyone in the population (census) and sometimes we can only collect data from a small subgroup of the population (sample). Either way, once we have the data, we need to be able to analyze it. This chapter focuses on the basics of data analysis. If you remember there are two types of data, quantitative (numerical measurements) and categorical (labels). We analyze quantitative data very differently than categorical data, so it is always vital to ask yourself a couple key questions.

- Was the data collected correctly, either an unbiased census or an unbiased large random sample?
- Is the data quantitative or categorical?
- Is there one data set or are we trying to analyze relationships between two data sets?

We will learn about rules for judging sample sizes in the next few chapters. This chapter focuses on being able to analyze the sample data or census data you have.

When analyzing data we rely on numbers calculated from the data that can help us understand the key features of the data set. If these numbers were calculated from a sample, they are called statistics. If these numbers are calculated from an unbiased census, they are called parameters. Most of the time, we only have sample data, so it is vital to understand and explain statistics.

Note on calculation: We live in the age of “big data”. No one today calculates statistics by hand, especially for a data set of ten-thousand values. Even a sample of one-hundred can be overwhelming to calculate. Statisticians and data scientists rely on computers to calculate statistics. The focus should be on understanding the meaning and correct use of the statistic, not on calculating by hand with a calculator.

### Section 1E – Categorical Data Analysis

**Vocabulary**

**Percentage (%):** An amount out of 100. For example if 72 out of every one-hundred employees opts to use a companies HMO insurance, we would say that 72% of the employees are using the HMO insurance.

**Proportion:** The decimal equivalent of a percentage. To calculate, divide the percentage by 100 and remove the percent symbol.

**Proportion and Percentage Conversions**

To analyze categorical data, we focus on exploring various types of percentages and compare them. In statistics, the decimal equivalent to a percentage is often called a “proportion”.

To convert a decimal proportion into a percentage, we multiply the proportion by 100%. This moves the decimal point two places to the right. Don’t forget to add the % symbol.

Example: Convert 0.047 into a percentage.

\[0.047 \times 100\% = 4.7\%\]

To convert a percentage into a decimal proportion, we divide by 100 and remove the percentage symbol. This moves the decimal two places to the left. Don’t forget to remove the % symbol.
Example: Convert 52.9% into a decimal proportion.

\[
52.9\% = \frac{52.9}{100} = 0.529
\]

Calculating Proportions and Percentages from Categorical Data

In order to calculate a decimal proportion from categorical data, you will need to find the amount (count, frequency) and divide by the total.

\[
\text{Decimal Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}}
\]

Counting how many people share a certain characteristic or even a total number of cars in a data set can take a long time in a big data set, however technology can help. Statistics software can count much quicker and easily than we can. In this section, we will assume we know the amount and the total.

Suppose a health clinic has seen 326 people in the last month and 41 of them had the flu. If we were analyzing their data, the first thing we would like to do is find what proportion of the patients have the flu. It is not a difficult calculation and can be done with a small calculator.

\[
\text{Decimal Proportion} = \frac{41}{326} = 0.12576687
\]

Should we round the answer? Proportions and Percentages are usually rounded to the three significant figures. Proportions are usually rounded to the thousandths place (3rd place to the right of the decimal).

Let us review rounding. We want to round the above answer to the thousandths place, which is the “5”. Always look at the number to the right of the place you are rounding to. If the number to the right is 5-9, round up (add 1 to the place value). If the number is 0-4, round down (leave the place value alone). After rounding cut off the rest of the decimals.

Therefore, in the previous answer we want to round to the thousandths place (5). The number to the right of the 5 is a 7. So should we round up or down? If you said round up, you are correct. Therefore, we will add 1 to the place value and the 5 becomes a 6. Now we cut off the rest of the decimal and our approximate answer is 0.126.

\[
\text{Decimal Proportion} = \frac{41}{326} = 0.12576687 \approx 0.126
\]

Decimal proportions are vital in the analysis of categorical data, but many people have trouble understanding the implications of a decimal proportion like 0.126. That is why we often convert the proportion into a percentage.

How to convert a decimal proportion into a percentage

To convert a decimal proportion into a percentage, multiply by 100 and put on the “%” symbol. Think of it like taking 100% of the decimal proportion. When you multiply by 100, the decimal moves two places to the right. Some people prefer to move the decimal, but I find students make fewer errors when they just multiply by 100 with their calculator.

\[
\text{Percentage} = \text{Decimal Proportion} \times 100\%
\]
Look at our previous example of the number of cases of the flu at a health clinic. We used the amount and total to calculate the decimal proportion.

\[
\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} = \frac{41}{326} = 0.12576687 \approx 0.126
\]

So what percentage of the patients had the flu? All we need to do is multiply the decimal proportion 0.126 by 100\% to get the percentage equivalent.

\[
\text{Percentage} = \text{Decimal Proportion} \times 100\% = 0.126 \times 100\% = 12.6\%
\]

So 12.6\% of the patients at the health clinic were seen for the flu. This can be alarming information to the health clinic if that is an unusually high percentage.

Notice that the percentage still has three significant figures, but is rounded to the tenths place (one place to the right of the decimal). Rounding to the tenth of a percent is a common place to round percentages in statistics.

If you want to calculate the percentage directly from the categorical data, here is another formula you may use.

\[
\text{Decimal Proportion} = \frac{\text{Amount}}{\text{Total}} \times 100\%
\]

Important Note
There are three ways to describe the proportion for categorical data: fraction, decimal, and percentage. Notice for the flu data example above, we have the three ways of describing the data: the fraction \(\frac{41}{326}\), the decimal proportion 0.126, and the percentage 12.6\%. All of them are equivalent. It is important to be comfortable with fractions, decimal proportions and percentages when describing categorical data. They are a foundation for more advanced categorical analysis later on.

Calculating a Frequency (Count) from a Percentage

How to calculate a count (frequency) from a percentage or proportion. Sometimes a percentage is given in a scientific report or in an article. For more advanced proportion analysis, the computer programs usually require the actual count (frequency). So it is important to be able to find the frequency from percentage information.

Start by converting the percentage into a proportion.

\[
\text{Proportion} = \frac{\text{Percentage}}{100} \text{ (and remove the percent symbol %).}
\]

Now multiply the proportion times the total to get the amount (frequency). This often called taking a “percentage of a total”. It is important to round your answer to the ones place since is the number of people or objects that have a certain characteristic.

\[
\text{Count (Frequency)} = \text{Decimal Proportion} \times \text{Total}.
\]

Example

According to the Center for Disease Control (CDC), about 32\% of Americans have hypertension (high blood pressure). According to suburbanstats.org, Tulsa Oklahoma has approximately 603,403 people living in it. If the CDC is correct and 32\% of Americans have hypertension, then how many people do we expect to have hypertension in Tulsa?
Step 1: Convert 32% into a decimal proportion.

\[ 32\% = \frac{32}{100} = 0.32 \]

Step 2: Multiply the decimal proportion by the total.

Amount of people with hypertension = \(0.32 \times 603403 = 193088.96 \approx 193,089\)

So approximately 193 thousand people in Tulsa have high blood pressure. This is vital information for hospitals and doctors in the Tulsa, Oklahoma area.

**Bar Charts and Pie Charts**

A quick way to count how many people or objects have a certain label is to create a Bar Chart or Pie Chart. There are many statistics software that we could use to create these graphs. They are useful to show the characteristics of categorical data.

**Creating a Bar Chart with Raw Data and StatKey**

StatKey does not create pie charts, but does have a nice bar chart feature. It not only creates the bar chart from the raw data but calculates the counts (frequencies) from each category as well as the decimal proportions.

To make a bar chart with raw data, go to [www.lock5stat.com](http://www.lock5stat.com) and click on the “StatKey” button. Now click on “one categorical variable” under the descriptive statistics and graphs button. If you have raw categorical data, click the “edit data” tab and paste your raw categorical data into StatKey. Make sure to check “raw data” at the bottom. If your data has a title, also check “data has a header row”. No click “OK”.

For example, I copied and pasted the “transportation data” from the Math 140 Fall 2015 survey data at [www.matt-teachout.org](http://www.matt-teachout.org) into StatKey and created the bar chart. Notice it not only created the graph, but also gave me the counts (frequencies) and the decimal proportions.
Creating a Bar Chart with Summary Data and StatKey

Categorical data is often summarized by the counts for each variable. When a data analyst receives categorical data to analyze, if my not be in raw form. Often it is just the counts (frequencies). In that case, when you go to the “edit data” button, you will need to type in the variables and counts as shown below. Uncheck the “raw data” box at the bottom and push “OK”. Note that you need only one space after the comma and do not type in the totals. Notice you will get the exact same graphs, counts and proportions as shown above.

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive alone</td>
<td>267</td>
<td>0.804</td>
</tr>
<tr>
<td>Dropped off by someone</td>
<td>18</td>
<td>0.054</td>
</tr>
<tr>
<td>Carpool</td>
<td>30</td>
<td>0.09</td>
</tr>
<tr>
<td>Bicycle</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td>Public Transportation</td>
<td>6</td>
<td>0.018</td>
</tr>
<tr>
<td>Walk</td>
<td>10</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>332</strong></td>
<td><strong>1.000</strong></td>
</tr>
</tbody>
</table>

Creating a Pie Chart with Raw Categorical Data and Statcato

A pie chart is a very useful graph and can give the count (or frequency) for each variable and the percentages for each variable.

To create a pie chart with Statcato, open your excel spreadsheet. Copy and paste your column of categorical data from Excel into Statcato. Before pasting, be sure to click on the gray at the top of the column in Statcato, since titles must go in the gray. Now click on the graph menu at the top and then “pie chart”. Click on “data values from a worksheet” and then under “data” put in the column. If your data is in the first column, you will click on “C1”. If it is in the second column, you will click on “C2”, and so on. Give the chart a title and click on “Show Legends” and “Show Values/Percentages for each Pie Sector”. You can sort the graph by category or by frequency (counts). If you click on “sort by category”, the pieces will be put in alphabetical order clockwise around the circle. If you click on “sort by frequency,” then the chart will be organized from the smallest section to the largest section clockwise around the circle.

Graph Menu => Pie Chart => Data Values from a Worksheet => Sort by Categories or Frequencies, Show Legend, Show Values/Percentages
Let’s use the same example, and open the transportation data from Math 140 Survey Data from Fall 2015.

**Important Reminder:** If your data set is over 300 entries, you will need to add some rows to Statcato. The math 140-survey data had close to 350 students, so we will need to add some rows to the spreadsheet in Statcato before copy and pasting from Excel. (I added 200 more rows to Statcato before I tried to copy and paste.)

Once you have added enough rows in Statcato, copy and paste the column of data that says “Transportation” in Statcato. Do not forget to put the title in the gray cell at the top. Now go to the graph menu and make a pie chart. We will show two versions of the graph. One if you sort by categories and the other if you sort by frequencies. That way you can see the difference and which one you like better. The following graph was sorted by categories. Notice it gives the same counts as StatKey, though the proportions have been converted into percentages and rounded to two significant figures. You can copy and paste the graph into a Word or Pages document, by going to the “graph” button on the left side of the graph and click on “copy graph to clipboard”.

**Pie Chart**

- **Drive alone** = 267 (80%)
- **Carpool** = 30 (9%)
- **Public transportation** = 6 (2%)
- **Dropped off by someone** = 18 (5%)
- **Walk** = 10 (3%)
- **Bicycle** = 1 (0%)

This chapter is from *Introduction to Statistics for Community College Students*, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” Creative Commons Attribution 4.0 International license – 10/1/18
Notice at the touch of a button, the computer can tell us all of the counts (frequencies) and all of the percentages. We can answer all sorts of questions about how these students get to the college.

**Creating a Pie Chart or Bar Chart with Summary Data and Statcato**

Categorical data is often given in summarized form with the variables and the counts. Statcato cannot make bar charts from raw data, but it can make a bar chart from summary counts. Statcato can also make a pie chart from summarized data. Suppose we do not have access to the raw categorical transportation data. Suppose we only knew the variable labels and the counts. First type in the variables and counts (frequencies) into two columns of Statcato. We will use the transportation data again. Note that titles like “variable” or “count” must be typed in the gray where it says “Var”.

![Table of transportation data](image)

Now go to the graph menu and then “pie chart”. Click on “Summary Data from Worksheet”. Give the columns for the categories and the columns for the frequencies.

![Pie chart settings](image)

Notice the pie chart looks the same as the one we created with raw data.

We can also create a bar chart from summary categorical data. Again, type in the summary counts and variables into two columns of Statcato. Then go to the graph menu in Statcato and click on “Bar Chart”. Statcato will want to know what column has your variable names and the column that has your counts.
Under “Select the column variable of a new series”, pick the column with your counts (frequencies). Mine was in column 2. Now click “Add Series”. Under “Select the column variable containing categories” select the column that has your variable names. Mine was in column 1. Type in a title and “show legend” and press OK. You can make the bars vertical or horizontal as well. I used vertical in this example.
Comparing Percentages

Sometimes we want to compare categorical variables and see if one variable has a significantly higher proportion or percentage than another. To compare proportion or percentages, many people often calculate the “percentage of increase”. There are three different ways of calculating the percentage of increase. Any of these formulas give the same answer.

\[
\text{Percent of Increase} = \frac{(\text{Higher Proportion} - \text{Lower Proportion})}{\text{Lower Proportion}} \times 100\%
\]

\[
\text{Percent of Increase} = \frac{(\text{Higher} \% - \text{Lower} \%)}{\text{Lower} \%} \times 100\%
\]

For example, let’s look at the transportation bar chart found with StatKey. Suppose we want to compare the percentage of Math 140 students that carpool verse the percentage that were dropped off. We can calculate the percent of increase from the counts, proportions or percentages. It is important to recognize which is the lower count (frequency) and which is the higher count. In this case, the number of students that carpool was higher than the number of students that were dropped off. The key question is was it significantly higher?
We can calculate the percent of increase from either the proportions or the percentages.

\[
\text{Percent of Increase} = \left( \frac{\text{Higher Proportion} - \text{Lower Proportion}}{\text{Lower Proportion}} \right) \times 100\% = \left( \frac{0.09 - 0.054}{0.054} \right) \times 100\% \approx 66.7\%
\]

\[
\text{Percent of Increase} = \left( \frac{\text{Higher } \% - \text{Lower } \%}{\text{Lower } \%} \right) \times 100\% = \left( \frac{9\% - 5.4\%}{5.4\%} \right) \times 100\% \approx 66.7\%
\]

Notice this tells us that the proportion of students that carpool is 66.7\% higher than the proportion that are dropped off. This difference seems statistically significant.

Note: In chapter 3 and chapter 4 we will learn how to use confidence intervals, test statistics, and P-values to determine significant differences. These are generally more accurate than the percent of increase calculation.

**Statistical Significance verses Practical Significance**

Sometimes when there is a statistically significant difference, it does not necessarily mean it is of practical use. In the last example we saw that the number of students that carpool was a 66.7\% higher than the number of students that are dropped off. Does this mean that college should make a special parking lot for all of the Math 140 students that carpool? Probably not. We are only talking about a difference of 12 total students a semester. College of the Canyons has thousands of students. So even though the percent of increase is significant, the data is not really of practical use in the sense that I would be careful of making huge decisions from the 66.7\%.

**Binomial Proportions with Statcato (Optional Topic)**

Sometimes we want to know a percentage or proportion associated with a categorical event happening multiple times. One example of this is called a binomial proportion. A binomial proportion can be calculated from categorical data with only two outcomes (winning or losing, smoking or not, drinking alcohol or not). These are often referred to as “success” and “failure”. The individuals must be independent of each other and the event (success) percentage (p) must be the same all the time. To calculate a binomial percentage, you will need a computer program and three bits of information, the number of events (number of successes), the event proportion (p), and the sample size (n).

Example

Categorical data often has a requirement of at least 10 successes and at least 10 failures. Suppose we collect a random sample of 72 people and ask them whether they smoke cigarettes or not. Is 72 a large enough data set? Are we likely to get 10 or more people that smoke and 10 or more people that do not smoke? We can use Statcato to calculate this binomial percentage. According to the center for disease control, about 15.5\% of adults in the U.S. smoke cigarettes.

Probability (percentage) of 10 or more people smoking = ?

Number of Trials = Sample Size (n) = 72
Number of Events (X) = 10
Event Probability (p) = 0.155

Calculating binomial percentages can be challenging. Here is the formula that computer programs use.

\[
P(X) = C(n, X)p^X(1 - p)^{n-X}
\]

The problem with this formula is we have to calculate it for \(X = 10, X = 11, X = 12, \ldots, X = 72\) and then add all the proportions together. That is very difficult. It is best to let a computer program do the heavy lifting.
Open Statcato and click on “Calculate” menu. Then click on “probability distributions” and “binomial”. Statcato is limited in the sense that it only calculates binomial percentage for either equal to (probability density) or less than or equal to (cumulative probability). So if we are calculating a greater than question, we must think about the opposite (less than or equal to). In this problem we want to find 10 or more. The opposite of this would be 9 or less. So we will calculate the percentage for 9 or less, then subtract the answer from 100%. This is sometimes called a “complement” proportion. In Statcato, put in the following. Under “Number of trials”, put in the sample size 72. Under “constant” put in the number of events 9. Under “Event probability”, put in 0.155. Now push the “Cumulative Probability” button and push “compute”.

Notice the probability of getting 9 or less is 0.304 or 30.4%. This is the complement percentage to what we are looking for. So the probability of getting 10 or more people that smoke should be 100% – 30.4% = 69.6%. This may not be a high enough percentage to assure us that we will get at least 10 people that smoke. I would recommend collecting more data (increase the sample size).

Example

Suppose a person is playing a game or roulette that has a 1/38 or 2.63% chance of winning. The gambler plans to play the game 20 times. What is the probability that he or she wins just once?
Open Statcato and click on “Calculate” menu. Then click on “probability distributions” and “binomial”. Remember to calculate equal, you need to click on the “probability density button”.

Number of Trials = 20

Event Probability = 0.0263

Number of Events = 1 *(Put this in the “constant” box.)*

---

**Binomial Distribution: n=20, p=0.0263**

Input: 1.0

Type: Probability density

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.317003</td>
</tr>
</tbody>
</table>

Notice the answer can be found under “P(X)”. So the gambler has a 0.317 (31.7%) chance of winning the game once.
Problem Set Section 1E

1. Convert each of the following percentages into a proportion. Do not round the answers.
   a) 96.1%
   b) 2.75%
   c) 0.664%
   d) 0.082%
   e) 39.7%
   f) 8.6%
   g) 0.189%
   h) 0.0025%

2. Convert each of the following proportions into a percentage. Do not round the answers.
   a) 0.057
   b) 0.812
   c) 0.0033
   d) 0.0214
   e) 0.0613
   f) 0.451
   g) 0.00045
   h) 0.0779

3. According to an article by CBS news, approximately 15% of Americans still do not have health insurance. If approximately 78,300 people live in Chino Hills CA, how many people in Chino Hills would we expect to not have health insurance? Round your answer to the ones place.

4. According to an article online, about 30% of Americans own at least one gun. About 305,700 people live in Stockton CA. If the article was accurate, then approximately how many people in Stockton do we expect to own at least one gun? Round your answer to the ones place.

5. An article by the American Diabetes Association estimates that as of 2012, about 9.3% of Americans have diabetes. College of the Canyons has approximately 18,400 students. If the percentage were correct, how many COC students would we expect to have diabetes? Round your answer to the ones place.

6. According to a news report by www.nielsen.com, about 15.9% of Americans struggle with hunger. Lancaster CA has approximately 161,000 people living in it. If the percentage from the Nielsen report is accurate, then how many people in Lancaster CA may be struggling with hunger? Round your answer to the ones place.

7. According to an article by the Autism Society, about 1.47% of people in the U.S. have autism. The article also stated that the percentage is increasing every year and that Autism is one of the fastest growing disorders in the U.S. Van Nuys, CA has approximately 136,400 people living in it. If the percentage by the Autism Society is correct, how many do we expect to have autism?

8. An article at www.seattletimes.com was addressing the issue of whether women in the U.S. prefer traditional jeans or athletic wear like yoga pants, sweat pants or leggings. Assume that a random sample of 213 total women were asked if they prefer traditional jeans or athletic wear. Assume 139 said they prefer athletic wear and 74 said they prefer traditional jeans. Calculate the decimal proportion and the percentages. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.
9. The article at www.seattletimes.com also said that jean companies are creating more and more stretchy jeans to compete with the growing trend of women preferring athletic wear. Assume that a random sample of 197 total women were asked if they prefer stretchy jeans or athletic wear. Assume 103 said they prefer athletic wear and 94 said they prefer stretchy jeans. Calculate the decimal proportion and the percentages. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

10. A hospital is trying to decide how to allocate resources to various departments. In particular, they are comparing the medical/surgical ward to the telemetry (heart monitor) ward since these wards have similar costs per patient. Assume we looked at a random sample of 350 patients admitted to the hospital. 57 were admitted to the medical/surgical ward and 49 were admitted to telemetry. Calculate the decimal proportion and the percentages. Then calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

11. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the campus data. Use StatKey to make a bar chart with proportions and frequencies included and answer the following questions. What proportion of the students went to Valencia? What proportion of the students went to the Canyon Country campus? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

12. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the gender data. Use Statcato to make a pie chart with percentages and frequencies included and answer the following questions. What percent of the students were female? What percent of the students were male? Calculate the percentage of increase. Does the percent of increase look statistically significant? Do you think it is practically significant? Explain.

13. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the hair color data. Use Statcato to make a pie chart with percentages and frequencies included. Use StatKey to make a bar chart with proportions and frequencies included. Which graph do you like better for explaining the data? Explain why. Which hair color had the highest proportion? Which hair color had the lowest proportion?

14. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the political part data. Use Statcato to make a pie chart with percentages and frequencies included. Use StatKey to make a bar chart with proportions and frequencies included. Which graph do you like better for explaining the data? Explain why. Which political party had the most students? Which political party had the least number of students?

15. Open the “Math 140 Survey Data Fall 2015” at www.matt-teachout.org. Look at the “month of birthday” data. This data has numbers in it. Explain why this is categorical data and not quantitative. Use Statcato to make a pie chart with percentages and frequencies included. Use StatKey to make a bar chart with proportions and frequencies included. Which graph do you like better for explaining the data? Explain why. Which month had the highest percentage? Which month had the lowest percentage?

Optional Binomial Probability Questions

Directions: Use the Binomial Probability function in Statcato to answer the following questions. Assume the questions meet the requirements for calculating a binomial probability.

16. Suppose we take a random sample of 150 people and ask them if they smoke cigarettes or not. We need to have at least 10 people in the data set that smoke. Assume that the population percentage for smoking in the U.S. is 15.5%. What is the probability that we will get 10 or more people that smoke in the data set? Is this percentage high enough for us to be confident that 150 people is a large enough data set? Explain.
17. Suppose we take a random sample of 150 people and ask them if they smoke cigarettes or not. We need to have at least 10 people in the data set that do not smoke. Assume that the population percentage for non-smokers in the U.S. is 84.5%. What is the probability that we will get 10 or more people that do not smoke in the data set? Is this percentage high enough for us to be confident that 150 people is a large enough data set? Explain.

18. To win at a dice game, the player must roll two dice and get a 7 or 11 sum. There is an 8/36 or 22.2% chance of winning. Suppose Lacy rolls the dice a total of 18 times.

   a) What is the probability that she wins 3 or less times?
   b) What is the probability that she wins exactly 2 or more times?
   c) What is the probability that she doesn’t win at all? (This means she wins zero times.)

19. A car company thinks that their minivan transmissions have a 12% defective rate. A total of 84 Minivans were brought in to a service center this month.

   a) What is the probability that exactly 10 of them need to have their transmission replaced?
   b) What is the probability that 8 or more of the minivans will need their transmission replaced?
   c) What is the probability that 5 or less of the minivans will need their transmission replaced?

Section 1F – Normal Quantitative Data Analysis

Vocabulary

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Normal Data: Data that is bell shaped, symmetric and unimodal. Also referred to as data that has a normal distribution.

Sample Size: Also called the total frequency.

Average: Also called the center of the data. A single number that represents a typical person or object in the data set.

Variability: Also called the spread. A measure of how spread out a data set is. A large spread tells us that the data is less consistent and the more difficult to predict. A small spread tells us that the data is more consistent and easier to predict.

Mean Average ($\bar{x}$): The balancing point for distances in a data set. The average for a data set that is normal.

Standard Deviation: The average or typical distance that points in a data set are from the mean. The measure of typical spread (typical variability) for a data set that is normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

Outliers: Unusual values in the data set.
Introduction

When analyzing numerical quantitative data, always start with finding the shape of the data set. Categorical data can be graphed, but does not have a shape. Categorical bar charts can be organized in a variety of ways depending on the order of the categories. Quantitative data is numerical measurement data and does have a shape.

Why should we find the shape?

The goal in analyzing quantitative data is to find the average, spread and unusual values. In statistics, there are many types of averages, many types of spreads. Shape helps us determine which averages and spreads are most accurate for the data.

Quantitative Statistics and Graphs with StatKey

The most common quantitative statistics we like to look at are the mean, median, standard deviation, 1st quartile, 3rd quartile, interquartile range, max, min, and range. The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the data set. Another type of graph is a histogram. This graph counts the number of data values in certain sections and makes a bar telling us how many numbers are in that section. The number of bars are also called “bins” or “buckets”. Another graph we like to look at is the boxplot. A boxplot is a graph of the 1st quartile, median, and 3rd quartile as well as potential outliers.

All of these graphs and statistics can be made with StatKey. Let’s look at an example. Go to www.matt-teachout.org and click on the “statistics” tab and then the “data sets” tab. Look for the “Health Data” excel file. Open the data set and copy the women’s heights data. Notice the data is quantitative. It measures the height in inches of the women and it seems reasonable to look for an average height of these women.

Go to www.lock5stat.com and click on the “StatKey” button. Under the “Descriptive Statistics and Graphs” menu, click on “One Quantitative Variable”. Click on the “Edit Data” button. Copy and paste the women’s height data into StatKey. Uncheck the box that says “first column is an identifier”. An identifier is a word next to every number. This data set does not have that. Check the box that says “data has a header row”. This means the data set has a title. Now push “OK”. Notice StatKey gives you the sample statistics, a dotplot, a histogram and a boxplot.
On the right of this histogram, you will see a slider that can adjust the number of “buckets” or bins. The smaller the data set the less bins you should have. This data set only has 40 numbers, so we want only a few bars. If we slide it to 3 buckets we get the following.

**Histogram Controls**

Set Limits

Number of buckets: 3
This data has a very special shape. It is called bell shaped or normal. Normally distributed data has the highest bar in the middle and about equal number of bars decreasing from the middle. It looks like a bell. We see that this data set is relatively normal (bell shaped) or “normally distributed”. StatKey has also given us summary statistics, but which statistics are most accurate for normal data?

### Summary Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>40</td>
</tr>
<tr>
<td>Mean</td>
<td>63.195</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.741</td>
</tr>
<tr>
<td>Minimum</td>
<td>57</td>
</tr>
<tr>
<td>Q1</td>
<td>61.350</td>
</tr>
<tr>
<td>Median</td>
<td>63.350</td>
</tr>
<tr>
<td>Q3</td>
<td>64.900</td>
</tr>
<tr>
<td>Maximum</td>
<td>68</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

*Important Note about Shape: The mean and standard deviation should only be used if the data set is normal. The mean and standard deviations are not accurate if the data does not have a normal shape.*

Mean ($\bar{x}$): The mean is a type of average used for data that is normally distributed. The mean balances the distances between all the numbers in the data set and the mean. Think of it this way. If you tool all the numbers in the data set below the mean, measured their distances from the mean, then added up those distances. That total distance for numbers below the mean would be equal to the total distance for numbers above the mean. The mean is calculated by adding up all the numbers in a data set ($\sum x$) and then dividing by how many numbers are in the data set (sample size “n”).

$$\bar{x} = \frac{\sum x}{n}$$
Standard Deviation (S): We said that the mean balances the distances in a data set. The standard deviation calculates the average distance numbers are from the mean. It is the most accurate measure of typical spread for data sets that are normally distributed. To calculate the standard deviation, computer programs take every single number in the data set and subtract the mean. Since those differences can be negative sometimes, they computer squares all the differences and then adds up the squares. This is a famous calculation called “sum of squares”. Since we want the average distance, we divide by \( n - 1 \) (degrees of freedom) and take the square root at the end to undo all the squares. Never calculate this by hand. It is a long calculation that should be left to a computer program.

\[
S = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}
\]

Why do we study spread? Spread is a measure of how much variability is in the data set. Think of it this way. Suppose we were looking at exam scores in a history class that are normally distributed. If a data set is very spread out, then the standard deviation would be quite large. This would mean that the scores had a lot of variability. We had A’s, B’s, C’s, D’s, and F’s. The exam scores are not consistent, and the history teacher will have a hard time predicting how her class will do. If the data set has a small spread, then the standard deviation would be quite small. The exam scores are very consistent. Maybe everyone in the class got an A or a high B. It is easier to predict how the class will do.

**Statistics for Normal Data**

Quantitative Variable and Units

Sample Size (n)

Maximum Value

Minimum Value

Average: Mean (\( \bar{x} \))

Spread: Standard Deviation (S)

Typical Values: One standard deviation from the mean. Here is a formula that is sometimes used.

\[
\bar{x} - s \leq \text{typical values} \leq \bar{x} + s
\]

Outliers (unusual values): More than two standard deviations from the mean. Here are formulas that are sometimes used.

\[
\text{Unusually Low Values (Low outliers)} \leq \bar{x} - 2s
\]

\[
\text{Unusually High Values (High outliers)} \geq \bar{x} + 2s
\]

**Women’s Height Example**

Quantitative Variable and Units: Women’s heights in inches

Sample Size (n): There were 40 women in the data set.
Maximum Value: The tallest woman in the data set was 68 inches.
Minimum Value: The shortest woman in the data set was 57 inches.
Average: Mean ($\bar{x}$) The average height of the women in the data was 63.195 inches.
Spread: Standard Deviation ($s$): The typical spread for this data was 2.741 inches. Typical women in the data were 2.741 inches from the mean.
Typical Values: Add and subtract the mean and standard deviation. Typical women in the data set have a height between 60.454 inches and 65.936 inches. We will see later that these values are the cutoffs for the middle 68% for normal data.

$$\bar{x} - s \leq \text{typical values} \leq \bar{x} + s$$

$$63.195 - 2.741 \leq \text{typical values} \leq 63.195 + 2.741$$

$$60.454 \leq \text{typical values} \leq 65.936$$

Outliers (unusual values): Add and subtract the mean and two standard deviations. Unusually tall women are 68.677 inches or higher. There are no unusually tall women in this data set. Unusually short women are 57.713 inches or lower. This means that the minimum value of 57 inches was unusually low. We will see later that these values are the cutoffs for the top and bottom 2.5% for normal data.

Unusually Low Values (Low outliers) $\leq \bar{x} - 2s = 63.195 - (2 \times 2.741) = 57.713$ inches

Unusually High Values (High outliers) $\geq \bar{x} + 2s = 63.195 + (2 \times 2.741) = 68.677$ inches

Quantitative Statistics and Graphs with Statcato

You can also make dotplots, histograms and sample statistics with Statcato. Copy and paste women’s heights into a column of Statcato. The data set is only 40 values, so you will not need to add rows to Statcato. To make a dot plot, go to the graph menu and click on dot plot. Then click on the column of data you want to use. Then push ok.

Making a dot plot in Statcato: Graph => Dot plot => Pick a column => OK

Here is the dot plot for the 40 women’s heights.
To make a histogram in Statcato, go to the graph menu, and then click on histogram. Chose a column of data and how many bars (bins) you want. Then chose ok.

Making a histogram in Statcato: Graph => Histogram => Pick a column => Chose number of bins => OK

**Note about bins:** If you chose too many bars then the histogram starts to look very crazy and you will have a hard time seeing the shape. Remember the goal is to break the dots up into groups. For example, in this health data there are only 40 women. I would not want 40 bins since that would give me about one bar per dot. If it were a small data set like the health data, I would do about three bins. Remember, the more bins you have, the more difficult it is to see the shape. This graph has five bins.
Notice again that the highest bar is close to the middle and the bars get smaller as we move away from the middle. This is often called “Bell Shaped” or “Normal Data”. Some like to describe this shape as unimodal (1 hill) and symmetric (left and right side look about the same). I prefer to call it bell shaped or normal.

We can also calculate all of the sample statistics with Statcato. Go to the “Statistics” menu, then click “Basic Statistics” and “Descriptive Statistics”. I had pasted the data into column 1, so type in “C1” under “input variable”. Check the boxes for statistics that you want and push “OK”.

![Descriptive Statistics](image)
Z-scores

In normal data, we often want to find out how many standard deviations a number (X-value) is from the mean. This is called a “Z-score”. Here is a common formula. In later chapters, we will see that we can also use the Z-score as a test statistic to measure significance.

\[ Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}} \]

Example: In the last example we saw that the women’s height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Suppose a woman is 72 inches tall. What would be the Z-score for her height? Is she unusually tall?

It is important when calculating a Z-score that you subtract the X value and the mean first. Then divide by the standard deviation. Most people in statistics round Z-scores to the hundredths place (two numbers to the right of the decimal).

\[ Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}} = \frac{(72 - 63.195)}{2.741} = +3.21233 \approx +3.21 \]

If the X-value is below the mean, the Z-score will be negative. If the X-value is above the mean, the Z-score will be positive. This Z-score was positive. So the woman that is 72 inches tall is 3.21 standard deviations above the mean. Is this unusual?

Remember the formula above for finding the cutoff for unusual values for normal data. Notice it is two standard deviations above and below the mean. Two standard deviations above the mean would be a Z-score of +2. Two standard deviations below the mean would be a Z-score of −2. So a common way to judge if a number is unusual (outlier) for normal data is to look at the Z-score.

Unusual High Values for Normal Data: \( Z \geq +2 \)

Unusual Low Values for Normal Data: \( Z \leq -2 \)

Hence since the woman’s Z-score was greater than or equal to +2, she is unusually tall compared to the women in the data set.

Example: The women’s height data was normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. One woman in the data set was 57 inches tall and we said was unusually short. If you recall, her height was below the unusual low cutoff of 57.713 inches. What would be the Z-score for her height?

\[ Z = \frac{(X \text{ value} - \text{Mean})}{\text{Standard Deviation}} = \frac{(57 - 63.195)}{2.741} \approx -2.26 \]

Since the X-value is below the mean, the Z-score will be negative. So the woman that is 57 inches tall is 2.26 standard deviations below the mean. Remember if the Z-score is less than −2, it is unusually low. This confirms what we already knew.

Typical Z-scores: Remember that typical values are within one standard deviation from the mean. This would mean that typical Z-scores are between −1 and +1.

\[-1 \leq \text{Typical Z-scores} \leq +1\]

A woman with a height of 61 inches would have a Z-score of −0.80. Notice that this Z-score is between −1 and +1 on the number line. So 61 inches is a typical height for women in this data set.
Note: Not all values are typical or unusual. A person that is 1.5 standard deviations from the mean would be neither typical (Z-score not between −1 and +1) nor unusual (Z-score not greater than +2 or less than −2).

Empirical Rule

There is common percentages that go with normal (bell-shaped) data. Usually about 68% of normal data will be within one standard deviation of the mean (typical). About 95% of normal data will be within two standard deviations of the mean. About 99.7% of normal data will be within three standard deviations of the mean. These percentages are often referred to as the “Empirical Rule” or the “68-95-99.7 Rule”.

Notice that we can use the 68%, 95% and 99.7% to figure out the sections. Since 68% makes up the middle two symmetric sections, we know each section is about 34%. Similarly, the middle 4 sections make up about 95%. Subtract out the middle two sections (68%) gives 27%. Divide that in half and you get two sections each making up 13.5% of the normal data. The middle 6 sections make up about 99.7%. Subtract out the middle four sections (95%) gives 4.7%. Divide that in half and you get two sections each making up 2.35% of the normal data. The end sections are calculated in a similar manner (100%-99.7% = 0.3%). Divide that into two symmetric tails and we get that each tail should be about 0.15%.

Remember the number of standard deviations from the mean is the Z-score. You can write the Z-scores for the bottom values in the Empirical rule. This is often called the “Standard Normal Curve”. Notice the center of the curve is the mean (Z-score of zero) and the standard deviation of this curve is exactly one. When a computer program refers to a normal curve with a mean of zero and a standard deviation of one, they are talking about Z-scores and the Standard Normal Curve.
Many data sets are normal. We will see in the next chapter that many sampling distributions have a normal shape as well. It is therefore important to be able to calculate percentages associated with normal data and normal curves. Confidence Intervals and P-value are both extremely important topics that we will cover in chapter 3 and chapter 4 that involve the empirical rule and calculating percentages associated with normal curves.

**Calculating Percentages for Normal Curves with StatKey**

Computer software programs can calculate percentages associated with normal quantitative data. Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Theoretical Distributions” menu click on “Normal”. Notice the parameters are set at a mean of zero and a standard deviation of one. Remember this means it is set up to find Z-scores or to find percentages associated with Z-scores. The curve is sometimes called a “density curve”. The idea is that the total area under the curve is 100%, so to find a percentage you find the area under the curve.

### Normal Distribution

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**StatKey Theoretical Distribution**

Notice that the curve has three buttons on the top left (Left Tail, Two-Tail, and Right Tail).

Example: Suppose we want to find what percent of normal data has a Z-score of −2.371 or above. Since we are looking for above, click the right tail button. The upper box is the percentage and the lower box is the Z-score. In this case we know the Z-score and are looking for the percentage. So in the bottom box type in −2.371.
Notice the top box is the answer, 99.1% of normal data values will have a Z-score of −2.371 or higher.

Example: Push the “reset plot” button. Suppose we want to find the two Z-scores that 90% (0.9) of normal data values are in between. Since we are looking for “in between”, click the two-tail button. The upper boxes are the percentages in each tail and in the middle the lower boxes are the two Z-scores. In this case we know the percentage in between and are looking for the Z-scores. So in the upper middle box type in the decimal proportion equivalent of 90% (0.9).
Notice the Z-score answers we are looking for are at the bottom. So the middle 90% of normal data values have a Z-score between $-1.645$ and $+1.645$. These are famous Z-scores for 90% confidence intervals that we will study in chapter 3.

**Percentages for any normal data**

We often want to calculate percentages for normal quantitative data without calculating Z-scores first. StatKey can do that as well. Push the “reset plot” button. Right now the mean is set at zero and the standard deviation is at one.

**Normal Distribution**

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Example: Suppose we want to calculate percentages associated with the women’s height data we studied earlier. We found that the women’s heights were normally distributed with a mean of 63.195 inches and a standard deviation of 2.741 inches. Click on the button that says “edit parameters” and put those numbers into StatKey.
Suppose we want to know what percentage of women in the data have a height of 69 inches or less. Since we are looking for “less than”, click on left tail. Remember the top box is the percentage (proportion). The bottom box is now the height. Since we know the height is 69, type in 69 into the bottom box. The proportion in the top box is our answer. So about 98.3% of the women in the sample data have a height below 69 inches.

Note: Be careful about generalizing results of sample data to the population. This does not mean that 98.3% of all women have a height of 69 inches or below. As we learned in chapter one, samples may have bias and not represent the population.

Example: Suppose we wish to find the heights that the middle 35% of the women are in between. Just push the “two-tail” button and put 0.35 in the upper middle box. The answer will be in the two lower boxes.
So about 35% of the women in the data have a height between 61.951 and 64.439 inches.

Note: These percentages are based on perfectly normal curves, yet real data is rarely perfectly normal. There are actually 15 women in the data had a height between 61.951 and 64.439. This was actually 37.5%. This is off from the theoretical percentage because the data was not perfectly normal. It is important to realize that theoretical distributions rarely match up exactly with real data.

Calculating Percentages for Normal Curves with Statcato

Z-scores, X-values and percentages for normal curves can also be calculated with Statcato. Go to the “Calculate” menu, click on “Probability Distributions” and then “Normal”.
If you leave the mean at zero and the standard deviation at 1, then Statcato is set up to calculate Z-scores or percentages from Z-scores. To calculate a Z-score from a percentage less than the Z-score, put in the proportion (decimal equivalent of the percentage) into the box that says “constant”. Then click “inverse cumulative” and “compute”. For example, what is the Z-score that 85% of values in a normal data set are less than? The answer is under “X”. The Z-score is 1.0365.

**Normal Distribution: mean = 0.0 stdev = 1.0**
Input: 0.85
Type: Inverse cumulative probability

\[ P(<=X) \times \]

0.85 1.0365
Suppose we want to find the percentage less than a Z-score of 2.36. Put 2.36 in the constant box and press “Cumulative Probability”. The answer is under “P(<=X)”. So the answer 0.990863 or about 99.1%.

**Normal Distribution: mean = 0.0 stdev = 1.0**

Input: 2.36

Type: Cumulative probability

\[
\begin{array}{ccc}
X & P(<=X) \\
2.36 & 0.990863
\end{array}
\]
We can also calculate X-values and percentage for those X-values for normally distributed data. We need to input the mean and standard deviation into Statcato. For example, earlier we saw some random sample data for women’s heights was normally distributed with a mean of 63.195 and a standard deviation of 2.741. Suppose we want to find the percentage of women in the data that have a height below 64 inches. We see that the answer is 0.615502 or about 61.6%. Note that Statcato can only calculate for less than. If we want to know what percent of women in the data have a height above 64 inches, we first calculate less than and then subtract the answer from 100%. In this case, 100% − 61.6% = 38.4%.

**Normal Distribution: mean = 63.195 stdev = 2.741**

Input: 64.0
Type: Cumulative probability

<table>
<thead>
<tr>
<th>X</th>
<th>P(&lt;=X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64.0</td>
<td>0.615502</td>
</tr>
</tbody>
</table>
You can also use the “Inverse Cumulative Probability” function to calculate the height that 15% of women are taller than. Remember, Statcato only works with less than, so if 15% of women are greater than this height, than 85% of women are less than this same height. So we will enter 85% (0.85) into the constant box. We see the answer under “X”. So 85% of women have a height less than 66.0358 inches. This also means that 15% of women have a height above 66.0358 inches.

**Normal Distribution: mean = 63.195 std.dev = 2.741**

Input: 0.85
Type: Inverse cumulative probability

\[ P(\leq X) \ 
\]

\[
\begin{array}{c}
0.85 \\
66.0358
\end{array}
\]
Calculating between is challenging with Statcato. It does not have a between button so we must work off of percentages less than an X-value. If we want to find the two values that the middle 40% are in between, we have to think about the percentages less than each X-value. If 40% is in the middle, that means that the remaining 60% is divided into the two tails. So each tail is 30%. So the X-value on the left will have 30% (0.3) less than. The X-value on the right will have 70% (0.7) less than. Put 0.3 into the “Constant” box and press inverse cumulative. Then put 0.7 into the “Constant” box and press inverse cumulative. For women’s heights we would get that the middle 40% of women’s heights are between 61.7576 inches and 64.6324 inches.

**Normal Distribution: mean = 63.195 stdev = 2.741**

Input: 0.3
Type: Inverse cumulative probability

\[ P(<X) \times \\
0.3 \quad 61.7576 \]

**Normal Distribution: mean = 63.195 stdev = 2.741**

Input: 0.7
Type: Inverse cumulative probability

\[ P(<X) \times \\
0.7 \quad 64.6324 \]
Note on Rounding Statistics for Quantitative Data

It is often best to not round if you are unsure. Data analysts usually prefer better accuracy and can round to their own specifications. Rounding too much interferes with accuracy. If you must round, here are some general guidelines.

Percentages and proportions are usually rounded to three significant figures. Proportions are rounded to the thousandths place and percentages are rounded to the tenths place.

Quantitative statistics like the mean or standard deviation are usually rounded to one more decimal place to the right than the original data has. Notice the women’s heights data is rounded to the tenths place (one number to the right of the decimal). So statistics calculated from this data would usually be rounded to the hundredths place (two numbers to the right of the decimal).

Mean (women’s height) = 63.195 \approx 63.20 \text{ inches}

Standard Deviation (women’s height) = 2.741 \approx 2.74 \text{ inches}

<table>
<thead>
<tr>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women Ht (in)</td>
</tr>
<tr>
<td>64.3</td>
</tr>
<tr>
<td>66.4</td>
</tr>
<tr>
<td>62.3</td>
</tr>
<tr>
<td>62.3</td>
</tr>
<tr>
<td>59.6</td>
</tr>
<tr>
<td>63.6</td>
</tr>
</tbody>
</table>

Practice Problems Section 1F

1. Answer the following questions:
   a) What is meant by saying that data is normally distributed or “normal”?
   b) Define the mean average and explain how it is calculated.
   c) Define the standard deviation and explain how it is calculated.

2. Answer the following questions:
   a) If a data set is normally distributed, what measure of average should we use?
   b) If a data set is normally distributed, what measure of spread should we use?
   c) If a data set is normally distributed, how many standard deviations from the mean is considered typical?
   d) If a data set is normally distributed, what is the formula for finding typical values?
   e) If a data set is normally distributed, approximately what percentage is typical?
f) If a data set is normally distributed, how many standard deviations from the mean is considered unusual?

g) If a data set is normally distributed, approximately what percentage of the data is unusually high?

h) If a data set is normally distributed, approximately what percentage of the data is unusually low?

Directions: Analyze the following data sets. Open “Bear” data and the “Health” data from my website www.matt-teachout.org. (Look under “Statistics” tab and then click the “data sets” tab.) Use StatKey or Statcatoto to create a dotplot, histogram and find summary statistics. Verify that each data set is normal and that the mean and standard deviation are accurate. Remember, for normal data we should use the mean as our average and the standard deviation as our measure of typical spread. Calculate the typical range by adding and subtracting the mean and standard deviation. Find the unusual cutoff values by adding and subtracting the mean and two standard deviations. List any unusual values in the data set. Do not round.

3. Bear neck circumference (inches)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the data set have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no high outliers, put “none”.

4. Bear Chest Size (inches)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the data set have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no high outliers, put “none”.

This chapter is from Introduction to Statistics for Community College Students, 1st Edition, by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-By” Creative Commons Attribution 4.0 International license – 10/1/18
5. Women's Diastolic Blood Pressure
   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the data set have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no low outliers, put “none”.

6. Women's Wrist Circumference (Inches)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the data set have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no low outliers, put “none”.

7. Men's Height (Inches)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the data set have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no low outliers, put “none”.
8. Men’s Weight (Pounds)

   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) Is the data set normally distributed? (Yes or No)
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? *(Give the number and the name of the statistic used.)*
   g) How much typical spread does the data set have?
      *(Give the number and the name of the statistic used.)*
   h) Find two numbers that typical values fall in between.
   i) What is the unusual high (high outlier) cutoff for this data?
   j) What is the unusual low (low outlier) cutoff for this data?
   k) List all high outliers in this data set. If there are no high outliers, put “none”.
   l) List all low outliers in this data set. If there are no low outliers, put “none”.

9. Write the definition of a Z-score and explain how we can use Z-scores to tell if a number is unusual?

10. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Use this information to answer the following Z-score questions.

   a) Bud has an IQ of 143. Calculate the Z-score for Bud’s IQ. Is his IQ unusually high compared to other people in the data set? How do you know?
   b) Jan has an IQ of 89. Calculate the Z-score for Jan’s IQ. Is her IQ unusually low compared to other people in the data set? How do you know?

11. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Use this information to answer the following Z-score questions.

   a) Maria spent $105.12 on merchandise in the store. Calculate the Z-score for Maria. Is $105.12 unusually high compared to other people in the data set? How do you know?
   b) Julie spent $13.61 on merchandise in the store. Calculate the Z-score for Julie. Is $13.61 unusually low compared to other people in the data set? How do you know?

12. Draw the standard normal curve and explain the percentages that make up the empirical rule.
13. The salaries of employees at a company are normally distributed with a mean of 31.4 thousand dollars and a standard deviation of 2.1 thousand dollars. Use the Empirical Rule graph below to answer the following questions.

![Empirical Rule Graph](image)

a) What percentage of the employees have a salary between 27.2 thousand dollars and 35.6 thousand dollars?
b) What percentage of the employees have a salary between 29.3 thousand dollars and 33.5 thousand dollars?
c) What percentage of the employees have a salary between 25.1 thousand dollars and 37.7 thousand dollars?
d) What percentage of the employees have a salary greater than 33.5 thousand dollars?
e) What percentage of the employees have a salary less than 27.2 thousand dollars?
f) Typical values for a normal curve are one standard deviation from the mean. Find two salaries that typical employee salaries fall in between?
g) The unusual high cutoff is two standard deviations above the mean. What salary represents the unusual high cutoff, that is the salary that 2.5% of the employees are greater than?
h) The unusual low cutoff is two standard deviations below the mean. What salary represents the unusual low cutoff, that is the salary that 2.5% of the employees are less than?

14. A random sample of IQ tests is normally distributed with a mean of 99.8 and a standard deviation of 15.3. Use this information to answer the following questions.

a) Use StatKey or Statcato to calculate what percent of people in the IQ sample data that have an IQ greater than 77.
b) Use StatKey or Statcato to calculate what percent of people in the IQ sample data that have an IQ less than 108.
c) Use StatKey or Statcato to calculate what percent of people in the IQ sample data that have an IQ between 95 and 120.
d) Use StatKey or Statcato to find the IQ score that 60% of people are less than.
e) Use StatKey or Statcato to find the IQ score that 85% of people are greater than.
f) Use StatKey or Statcato to find two IQ scores that the middle 40% of people are in between.
15. A clothing store wants to study the amount of money spent in their store by customers. Census data indicated that the data is normally distributed with a mean of $46.89 and a standard deviation of $12.44. Use this information to answer the following Z-score and percentage questions.

   a) Use StatKey or Statcato to calculate what percent of people that spent more than $25.
   b) Use StatKey or Statcato to calculate what percent of people that spent less than $50.
   c) Use StatKey or Statcato to calculate what percent of people spent between $35 and $60.
   d) Use StatKey or Statcato to find the amount of money spent that 37% of people are less than.
   e) Use StatKey or Statcato to find the amount of money spent that 15% of people are more than.
   f) Use StatKey or Statcato to find two amounts that the middle 60% of people are in between.

Section 1G – Quantitative Data Analysis for Non-Normal Data and Summary Statistics

Vocabulary

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Normal Data: Data that is bell shaped, symmetric and unimodal.

Skewed Right Data: Also called positively skewed. Data where the center is on the far left and has a long tail to the right.

Skewed Left Data: Also called negatively skewed. Data where the center is on the far right and has a long tail to the left.

Sample Size: Also called the total frequency. The number of values are in a data set.

Median Average: The center of the data when the numbers are put in order. Also called the 50th percentile \( P_{50} \), since about 50% of the numbers in the data set are less than the median. It is also called the 2nd Quartile \( Q_2 \). The average for a data set that is not normal.

1st Quartile \( Q_1 \): The number that about 25% of the data values are less than. Used for typical values for data that is not normal.

3rd Quartile \( Q_3 \): The number that about 75% of the data values are less than. Used for typical values for data that is not normal.

Interquartile Range \( IQR \): The distance between the middle 50% of the numbers in a data set. Calculated by subtracting the 1st and 3rd quartiles. The measure of typical spread for a data set that is not normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

Range: A quick measure of total spread. Calculated by subtracting the minimum and maximum values in a data set.

Outliers: Unusual values in the data set.
Introduction

When a data set is normal (or bell-shaped), we use the mean as our average and the standard deviation as our measure of typical spread. Not all data sets are normal though. Let’s explore some data that is not normally distributed.

Let us look at another example from the health data. This time we will look at women’s pulse rates in beats per minute (BPM). Go to [www.matt-teachout.org](http://www.matt-teachout.org) and click on the “Statistics” tab and then the “Data Sets” tab. Open the health data in Excel. Copy the women’s pulse rate data. Now go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Under the “Descriptive Statistics and Graphs” menu, click on “One Quantitative Variable”. Under “Edit Data”, paste the women’s pulse rate data into StatKey. Uncheck the box that says “first column is identifier”, check the box that says “data has header row”, and push “OK”. Here are the graphs and summary statistics.
Notice first that this is not normal data. The highest bar (center) is on the far left. The graph has a short tail to the left of the highest bar and a long tail to the right of the highest bar. This shape is called “skewed right” or “positively skewed”. We can adjust the number of bars (buckets) by using the slider on the right of the graph.

### Summary Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>40</td>
</tr>
<tr>
<td>Mean</td>
<td>76.300</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>12.499</td>
</tr>
<tr>
<td>Minimum</td>
<td>60</td>
</tr>
<tr>
<td>Q₁</td>
<td>68.000</td>
</tr>
<tr>
<td>Median</td>
<td>74.000</td>
</tr>
<tr>
<td>Q₃</td>
<td>80.000</td>
</tr>
<tr>
<td>Maximum</td>
<td>124</td>
</tr>
</tbody>
</table>

Remember the mean and standard deviation are only accurate if the data is normal. So for this data set, we should not use the mean as the average and we should not use the standard deviation as our typical spread.

So what statistics should we use? Here is the general rule for skewed data or any data that is not normal.

**Summary statistics for non-normal data**

Average: Median  
Typical Spread: Interquartile Range (IQR)  
Typical Values: Between the 1st quartile ($Q₁$) and the 3rd quartile ($Q₃$)  
Outliers: Boxplot will indicate if there are outliers.

Quartiles are based on the numbers in order, so are much more accurate for data that is not normally distributed. The median is also called the 2nd quartile or the 50th percentile. It is the center of the data when the numbers are in order. About 50% of the numbers will be less than the median and about 50% of the numbers will be greater than the median. When a data set is not normally distributed, we use the median as our average. It is much closer to the center. Look at the histogram above. The summary statistics provided by StatKey show us that the mean was 76.3 beats per minute (bpm) and the median was 74 bpm. Notice 74 is closer to the highest bar in the data set. In other words the median is closer to the center and a more accurate average than the mean. Mean averages are based on distances so will be pulled off of the center in the direction of the skew.

The median is calculated by first putting the numbers in order from smallest to largest. If there is one number in the middle (sample size $n$ is odd), then that is the median. If there are two numbers in the middle (sample size $n$ is even), then the median will be half way between the two numbers in the middle.

The 1st quartile ($Q₁$) is also called the 25th percentile and is the number that about 25% of the data is less than. The 3rd quartile ($Q₃$) is also called the 75th percentile and is the number that about 75% of the data is less than. The 1st and 3rd quartiles are markers that mark the middle 50% of the data when it is in order. The middle 50% is
considered “typical” in a data set that is not normally distributed. For normal data we want the middle 68% (empirical rule) because there is more data in the middle.

The distance between the 1st and 3rd quartiles is called the interquartile range (IQR). This is the best measure of typical spread for data that is not normally distributed. StatKey does not list the IQR in its summary statistics, but we can calculate it with the following formula.

\[
IQR = Q_3 - Q_1
\]

Since our women’s pulse rate data was skewed right, we would use the following statistics.

Variable and Units: Women’s pulse rates in beats per minute (bpm)

Minimum: The lowest pulse rate for these women was 60 bpm.

Maximum: The highest pulse rate for these women was 124 bpm.

Average: The average pulse rate for these women is 74 bpm (median).

Typical spread: 

\[
IQR = Q_3 - Q_1 = 80 - 68 = 12 \text{ bpm}
\]

Typical women in the data set had a pulse rate within 12 bpm of each other.

Typical Values: Typical pulse rates are between 68 bpm \((Q_1)\) and 80 bpm \((Q_3)\).

Finding outliers for non-normal data

To find outliers for data sets that are not normally distributed, we will introduce another graph. The graph is called a “box and whisker plot” or “box plot” for short.

![Box plot](image.png)

A box plot is a graph of the 1st quartile, median, 3rd quartile and outliers. It is the perfect graph to look at when a data set is not normal. The left of the box is \(Q_1\) (68 bpm) and far right of the box is \(Q_3\) (80 bpm). So the box represents the typical values (middle 50%). The line inside the box is the median average of 74 bpm. The lines that go to the left and right of the box are called whiskers. The whiskers go to the lowest and highest numbers in
the data set that are not unusual (not outliers). The outliers are usually denoted by stars in StatKey and circles and triangles in Statcato. See the two stars the far right. Those are both outliers. There are two unusually high pulse rates in the data set. In StatKey, you can hold your curser over the stars and they will tell you what the numbers are. In this case the two high outliers are at 104 bpm and 124 pbm. There are no unusually low values since we do not see any stars on the left of the graph.

In case you are wondering, here are the formulas used by computer programs to determine outliers in a box plot. You do not need to calculate these yourself. The computer has already found your unusual values.

Unusual high (high outlier) cutoff: \( Q_3 + (1.5 \times IQR) \)

Unusual low (low outlier) cutoff: \( Q_1 - (1.5 \times IQR) \)

Note about box plots and normal data: Remember, a box plot is a graph of the quartiles and the median. They work really well for data that is not normal. However, they do not show the mean or standard deviation, so it is important to be careful how you interpret box plots for normal data. Normal data has different characteristics than those shown on a box plot. For example, typical values for normal data are not between \( Q_1 \) and \( Q_3 \). Also the outlier cutoffs are different for normal data so there may be differences in what is considered an outlier.

In the last section we saw that we can also calculate dot plots, histograms, box plots and summary statistics with Statcato. Copy and paste the data into a column of Statcato. Then go to the graph menu and click on “dot plot”, “histogram” or “box plot”.

**Dot Plot of Women’s Pulse Rates (Beats Per Minute)**
Notice that something is wrong with the Statcato box plot. The outliers have been left off. This is a common problem. To fix this, right click on the box-plot. Click on “zoom out” and “range axis”. You may have to do this multiple times. You want to be able to see the minimum value (60 bpm) and maximum value (124 bpm) on the scale of the graph. Here is the correct box plot.
Notice Statcato designated 104 with a circle (regular outlier) and 124 with a triangle (far out outlier). The dot in the middle of the box plot is the mean. Most box plots do not have the mean, but Statcato puts it in so that you can compare it to the median.

Let us look at some other examples.

Here is some salary data from a small company with 26 employees. The salaries are given in dollars per hour. We created a dot plot and histogram for this data.
Notice the highest bar and most dots are on the far right, while there is a long tail to the left. Therefore, this is called “skewed left” or “negatively skewed”.  

Note: Real data rarely has a perfect shape. Most data has a shape somewhere in between bell shaped and skewed, and you will need to make a decision. Look for a significant difference in the length of the tail to classify something as skewed. If my highest hill is toward the middle and I had 2 bars to the right and 3 bars to the left of the highest bar, I would still classify that bell shaped or normal. Some say that is “nearly normal”. If the highest hill is on the far right and I have 2 bars to the right of the highest hill and 7 bars to the left of the highest hill, I would classify that as skewed left. Some call this “negatively skewed” since negative numbers are to the left on the number line.

Here are a couple unusual shapes that sometimes appear.

A graph that looks like a rectangle is called “uniform”. A graph with two distinct high bars is called “bimodal”.

**Histogram with Uniform Shape**
Summary Statistics: Measures of Center, Spread and Position

Though the mean, median, standard deviation and IQR are used most often in data analysis, there are many different types of statistics that can be used to dig deeper into the data. We will not be covering these statistics in depth, but it is good to at least have an idea of what they measure.

Measures of Center

**Mean Average:** The balancing point in terms of distances. The measure of center or average used when a data set is bell shaped (normal).

**Median Average:** The center of the data in terms of order. Also called the second quartile (Q2) or the 50th percentile. Approximately 50% of the data will be less than the median and 50% will be above the median. This is the measure of center or average used when a data set is skewed (not bell shaped).

**Mode:** The number that occurs most often in a data set. Data sets may have no mode, one mode, or multiple modes. It is also sometimes used in bimodal or multimodal data.

**Midrange:** A quick measure of center that is usually not very accurate, but can be calculated quickly without a computer. (Max + Min) / 2

Measures of Spread

**Standard Deviation:** How far typical values are from the mean in a bell shaped data set. It is the most accurate measure of spread for bell shaped data. If you add and subtract the mean and standard deviation, you get two numbers that typical values in a bell shaped data set fall in between. It can also be used to find unusual values in bell shaped data. Should not be used unless the data is bell shaped.

**Variance:** The standard deviation squared. A measure of spread used in ANOVA testing. Only accurate when the data is bell shaped.

**Range:** A quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical values in the data set. It can be calculated quickly without a computer. (Max – Min)

**Interquartile range (IQR):** How far typical values are from each other in a skewed data set. Measures the length of the middle 50% of the data. It is the most accurate measure of spread for skewed data sets. Should not be used when data is bell shaped. (Q3–Q1)
Measures of Position
Minimum: The smallest number in the data set. Is sometimes classified as an unusual value (outlier).

Maximum: The largest number in the data set. Is sometimes classified as an unusual value (outlier).

First Quartile (Q1): The number that approximately 25% of the data is less than and 75% of the data is greater than. Used for finding typical values for skewed data sets.

Third Quartile (Q3): The number that approximately 75% of the data is less than and 25% of the data is greater than. Used for finding typical values for skewed data sets.

Frequency or Sample Size (n)
The frequency or sample size of a data set (n) is not a measure of center, spread or position, but is important bit of information. It tells us how many numbers are in the data set.

Practice Problems Section 1G
1. Answer the following questions:
   a) Describe a skewed right shape?
   b) Describe a skewed left shape?
   c) Define the median average and explain how it is calculated.
   d) Define the interquartile range (IQR) and explain how it is calculated.

2. Answer the following questions:
   a) If a data set is not normally distributed, what measure of average should we use?
   b) If a data set is not normally distributed, what measure of typical spread should we use?
   c) If a data set is not normally distributed, what are the two statistics that typical values are in between?
   d) If a data set is not normally distributed, approximately what percentage is typical?
   e) If a data set is not normally distributed, how can we use a box plot to find high outliers in the data set?
   f) If a data set is not normally distributed, how can we use a box plot to find low outliers in the data set?

Directions: Analyze the following data sets. Open “Bear” data and the “Health” data from my website www.matt-teachout.org. (Look under “Statistics” tab and then click the “data sets” tab.) Use StatKey or Statcato to create a dotplot, histogram, box plot and find summary statistics. Use the graphs and statistics to answer the following questions. Do not round. Remember if a data set is not normal, we should use the median as our average, IQR as our typical spread and $Q_1$ and $Q_3$ for two numbers that typical values fall in between. Also use the box plot to find any outliers (unusual value) in the data set. If you use Statcato, don’t forget to zoom out on the box plot to see the outliers.
3. Bear ages (months)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the dataset?
   c) What is the shape of the dataset?
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the dataset have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) List all high outliers in this dataset. If there are no high outliers, put “none”.
   j) List all low outliers in this dataset. If there are no low outliers, put “none”.

4. Bear Weights (pounds)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the dataset?
   c) What is the shape of the dataset?
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the dataset have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) List all high outliers in this dataset. If there are no high outliers, put “none”.
   j) List all low outliers in this dataset. If there are no low outliers, put “none”.

5. Women’s Systolic Blood Pressure in millimeters of mercury (mm of Hg)
   a) What is the data measuring and what are the units?
   b) How many numbers are in the dataset?
   c) What is the shape of the dataset?
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? (Give the number and the name of the statistic used.)
   g) How much typical spread does the dataset have?
      (Give the number and the name of the statistic used.)
   h) Find two numbers that typical values fall in between.
   i) List all high outliers in this dataset. If there are no high outliers, put “none”.
   j) List all low outliers in this dataset. If there are no low outliers, put “none”.

6. Men’s Diastolic Blood Pressure (mm of Hg)

   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) What is the shape of the data set?
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? *(Give the number and the name of the statistic used.)*
   g) How much typical spread does the data set have?
      *(Give the number and the name of the statistic used.)*
   h) Find two numbers that typical values falls in between.
   i) List all high outliers in this data set. If there are no high outliers, put “none”.
   j) List all low outliers in this data set. If there are no low outliers, put “none”.

7. Women’s Cholesterol in milligrams per deciliter (mg/dL)

   a) What is the data measuring and what are the units?
   b) How many numbers are in the data set?
   c) What is the shape of the data set?
   d) What is the minimum value?
   e) What is the maximum value?
   f) What is the average (center)? *(Give the number and the name of the statistic used.)*
   g) How much typical spread does the data set have?
      *(Give the number and the name of the statistic used.)*
   h) Find two numbers that typical values falls in between.
   i) List all high outliers in this data set. If there are no high outliers, put “none”.
   j) List all low outliers in this data set. If there are no low outliers, put “none”.

8. For each of the following statistics, classify it as a measure of center, spread or position.

   a) Q1
   b) Mean
   c) Variance
   d) Standard Deviation
   e) Minimum
   f) Q3
   g) Mode
   h) IQR
   i) Median
   j) Range
   k) Maximum
   l) Midrange
9. Write a brief description of each of the following statistics.

   a) Q1
   b) Mean
   c) Variance
   d) Standard Deviation
   e) Minimum
   f) Q3
   g) Mode
   h) IQR
   i) Median
   j) Range
   k) Maximum
   l) Midrange

Chapter 1 Review

Key Terms

Data: Information in all forms.

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Statistic: A number calculated from sample data in order to understand the characteristics of the data. For example, a sample mean average, a sample standard deviation, or a sample percentage.

Parameter: A population value, which is sometimes calculated from an unbiased census, but is often just a guess about what someone thinks the population value might be. For example, a population mean average or a population percentage.

Random: When everyone in the population has a chance to be included in the sample.

Simple Random Sample: Sample data in which individuals are selected randomly. This method tends to minimize sampling bias and is generally considered a good way to collect data.

Convenience Sample: Sample data that is collected in a way that is easy or convenient. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.
**Voluntary Response Sample:** Sample data that is collected by putting a survey out into the world and allowing anyone to fill it out. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

**Cluster Sample:** Sample data that collects data from groups of people in a population instead of one at a time. The groups should be chosen randomly to avoid sampling bias.

**Stratified Sample:** Sample data used to compare populations. The individuals should be chosen randomly to avoid sampling bias. For example, we may take a random sample of people living in Palmdale, CA and another random sample of people living in Valencia, CA and use the data to compare the average salaries.

**Systematic Sample:** Sample data that is collected with some type of system like choosing every twentieth person on a list.

**Bias:** When data does not represent the population.

**Sampling Bias:** A type of bias that results from collecting data without using a census or random sample.

**Question Bias:** A type of bias that results when someone phrases the question or gives extra information with the goal of swaying the person to answer a certain way.

**Response Bias:** A type of bias that results when people do not answer truthfully.

**Non-response Bias:** A type of bias that results when people refuse to participate or give data.

**Deliberate Bias:** A type of bias that results when people lie about their data results, manipulate data to get a certain result, or leave out groups of people from their data set.

**Experimental Design:** A scientific method for controlling confounding variables and proving cause and effect.

**Explanatory Variable:** The independent or treatment variable. In a cause and effect experiment, this is the cause variable.

**Response Variable:** The dependent variable. In an experiment this the variable that measures the effect.

**Treatment Group:** The group of people or objects that has the explanatory variable.

**Control Group:** The group of people or objects that does not have the explanatory variable.

**Confounding Variables:** Also called lurking variables. Other variables that might influence the response variable other than the explanatory variable being studied.

**Random assignment:** A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

**Placebo Effect:** The capacity of the human brain to manifest physical responses based on the person believing something is true.

**Placebo:** A fake medicine or fake treatment used to control the placebo effect.
Chapter 1 Review Problems

1. Tell if the following data is categorical or quantitative and explain why.
   a) The types of cars in the different parking lots.
   b) The average number of hours spent practicing ping-pong.
   c) Areas in North Dakota that have wild mustangs.
   d) Each person is asked if they wear glasses, contacts, neither, or both.
   e) The average speed of racecars at the Indianapolis 500.
   f) Exam scores for various students on a history exam.

2. Jim wants to know how much money the average working COC student makes. Describe how Jim could use each of the following techniques to collect data. For each technique, will there be a significant amount of sampling bias or not too much sampling bias?
   a) Systematic
   b) Voluntary Response
   c) Random Sample
   d) Convenience Sample
   e) Cluster Sample
   f) Stratified Sample
   g) Simple Random Sample
   h) Census

3. Define the following key terms and give an example of each.
   a) Population
   b) Census
   c) Sample
   d) Random
   e) Bias
   f) Parameter
   g) Statistic
4. Describe and give an example of each of the following types of bias.
   a) Sampling Bias
   b) Question Bias
   c) Response Bias
   d) Deliberate Bias
   e) Non-Response Bias

5. Rachael needs to do an experiment that will show that the nicotine patch causes a person to stop smoking. Set up the experiment for Rachael. Write a description of the experiment and include the following. What are some lurking variables that she will need to control? How can Rachael control the lurking variables? Include a description of how we will deal with the placebo effect.

6. Compare and contrast the similarities and differences between an experiment and an observational study. How can we tell if we should use an experiment or an observational study?