Lesson 2
Graphical and Numerical Descriptions of Data

In Lesson 1 we learned about categorical and numerical data, and we developed some preliminary tools for summarizing that data (that is, descriptive statistics). In this lesson we continue that theme, developing additional methods for describing the data contained in a single variable. In addition, we continue our focus on connections to probability.

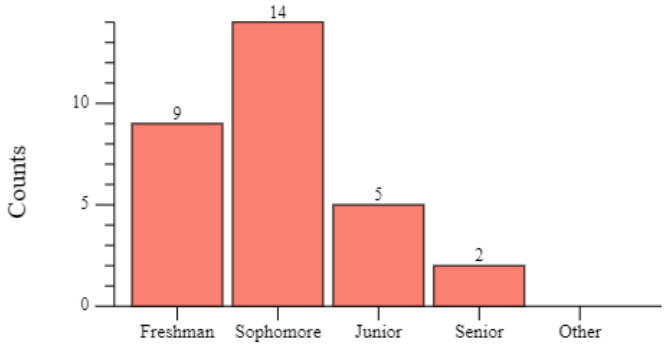
2.1 – Describing Categorical Variables

Lesson 1 already taught us how to describe categorical data numerically, using counts, proportions, and percents. We also learned how to present our summary information in the form of a frequency table, as illustrated by this example from Lesson 1:

Class Yr	Frequency	Proportion	Percent
Freshman	9	0.3000	30.00%
Sophomore	14	0.4667	46.67%
Junior	5	0.1667	16.67%
Senior	2	0.0667	6.67%
Other	0	0.0000	0.00%
Total	**30**		**100.00%**

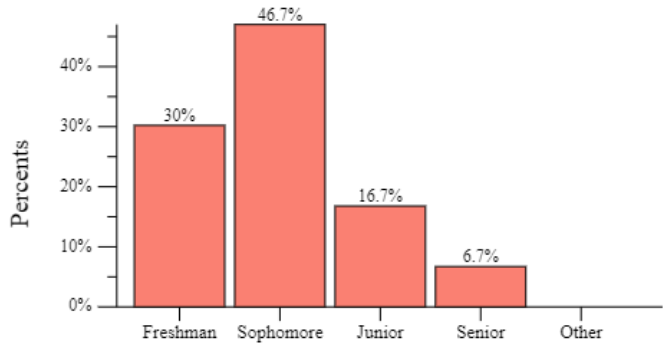
Graphical descriptions of categorical data are really quite simple. We use bar graphs or pie charts to create a picture of the data in the frequency table. These graphs help us visualize the counts or proportions/percents associated with the possible values for the variable. In Section 2.4 we will learn how to use technology to create these graphs. Here are some examples that have been created using the online statistical calculator that comes with these lessons.

The **bar graph** (or **bar chart**), in its simplest form, contains a list of the various categories along the horizontal axis, with a scale for the counts (frequencies) of those categories on the vertical axis. For each category, we have a bar whose height corresponds to the count for that category. Here, for example, is a bar chart for the frequency table given above.



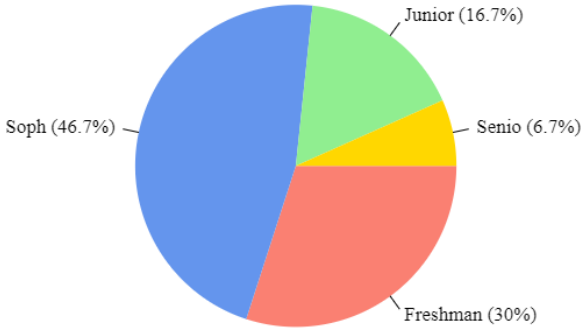
It may be useful, as was done in this example, to label the count for each category above the bar for that category. Observe that although this example was created using technology, it would also be quite easy to create by hand.

In some situations, it is more useful to know the proportion (percent) for each category rather than the raw count. A bar chart showing the percents will look identical to that for the counts, except that the vertical scale (and the labels on the bars, if they are included) will reflect percents, as illustrated in this example using the same data as in the previous bar chart example.



Again, creating this bar chart by hand would be relatively simple.

A **pie chart**, on the other hand, would be more difficult but not impossible to create by hand; as a result, one usually uses technology for this graph. It consists of a circle, with pie-slice shaped pieces whose area as a proportion of the entire circle is given by the proportions for the categories. Here is a sample pie chart for the same data as that used for the bar charts.



Comments. Observe that the "Other" category is missing from the graph, because its percentage was 0%. Also observe that the technology used to create the graph has truncated the labels for some of the categories.

2.2 – Graphs for Numerical Variables

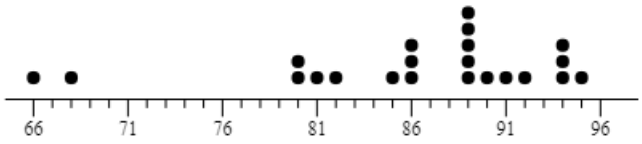
In this course, we will study four methods for graphing numerical variables.

- dot plot
- stem-and-leaf plot
- histogram
- box plot

The first three types of graphs can be described at this time. For the box plot we will first have to learn about numerical methods for summarizing our numerical data.

Dot plots and stem-and-leaf plots

The first two, the **dot plot** and the **stem-and-leaf plot**, give a direct picture of the actual data, and it is possible to recover the data from the graph. They are quite simple to execute by hand, although doing so can be tiresome if the data set is large. Here are small examples. The dot plot below depicts the grades of a class of students on the first test of the semester.



In the plot, the three dots above the 86 indicate that the data set contained the value 86 three times (three people scored 86 on the test). To create the dot plot, one would simply create a scale based on the range

of the data, then place one dot for each value in the data set. Observe that we can recreate the data set from the dot plot: 66, 68, 80, 80, 81, 82, 85, 86, 86, 86, 89, 89, 89, 89, 89, 90, 91, 92, 94, 94, 94, 95. There are 22 dots, one for each of the 22 student scores on the test.

The stem-and-leaf plot below represents this same set of data. The four numbers to the left of the vertical line are the so-called *stems* of the plot. For this set of data, the stems are the tens digits of the data. The numbers to the right of the vertical line are the *leaves*, in this case the units digits of the data. The first line of the plot, 6 | 68, represents the two scores 66 and 68. In the bottom line, the stem is 9; in the list of leaves the three 4s represent the three scores 94, 94, and 94. There are 22 scores in the set of data, resulting in 22 leaves in the plot. Again, at least for small data sets of two-digit numbers, creating the plot by hand is quite simple.

```
6 | 68
7 |
8 | 0012566699999
9 | 0124445
```

Comments.

1. Observe that for each stem the corresponding leaves are in increasing numerical order.
2. The plot contains a stem of 7 even though this stem has no leaves. There should not be any gaps in the list of stems.
3. On the other hand, it is customary not to include the stems that are smaller than the smallest stem in the data set, that is 0, 1, 2, 3, 4, and 5. Similarly, stems larger than the largest stem in the data set would be omitted.
4. There are a variety of more complicated types of stem-and-leaf plots which we choose not to cover in these lessons – for example, graphs that can be used for data that does not consist of two-digit numbers.

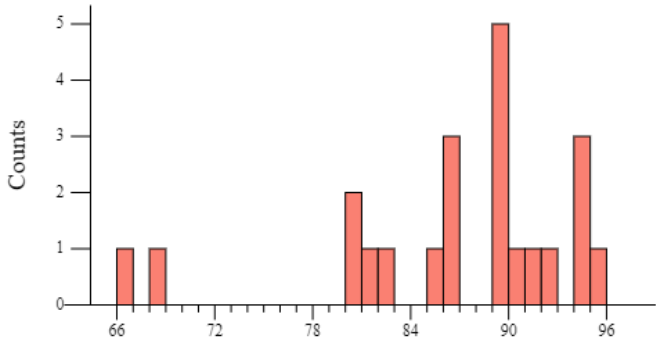
The dot plot and the stem-and-leaf plot are sometimes used because they are relatively simple to construct by hand, at least for small sets of data, and because they directly represent the individual number in the data set. However, the histogram, which we consider next, is a by far more useful tool for the graphical depiction of numerical data.

Exercise 1[1]: Create a dot plot and a stem-and-leaf plot for this data set: 15, 17, 17, 11, 12, 12, 12, 18, 19, 32, 33, 33, 33, 33, 36, 40

Histograms

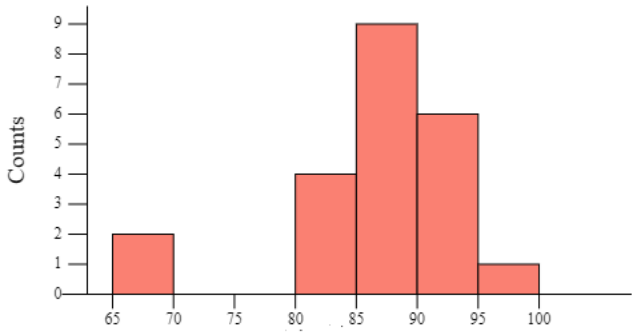
The histogram can sometimes be identical to the dot plot, except that it uses rectangles rather than dots to achieve the final graph. For example, here is a possible histogram for the same set of data in the dot plot, namely 66, 68, 80, 80, 81, 82, 85, 86, 86, 86, 89, 89, 89, 89, 89, 90, 91, 92, 94, 94, 94, 95.

[1] Solutions to the exercises may be found at the end of the lesson.



However, there is a subtle difference. The bar of height 5 that extends on the horizontal scale from 89 to 90 actually represents a count of the data between 89 and 90 – or, more precisely, all the data that is greater than or equal to 89 but less than 90. If instead of 89 five times the data contained these numbers – 89.3, 89.7, 89, 89.1, and 89.7 – the graph would have been identical to that above.

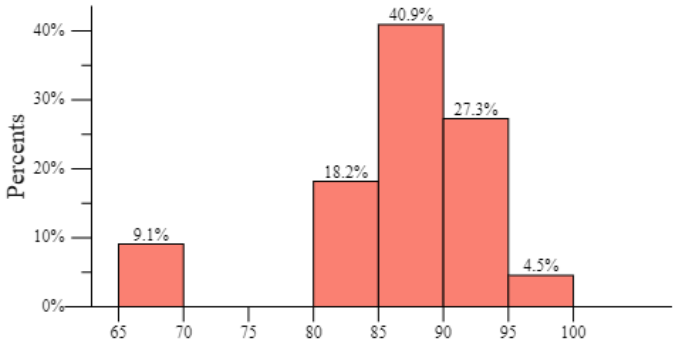
Here is another difference – the bars do not have to be of width 1 unit; in fact, they most typically are not 1 unit wide. For example, we might decide to graph this same set of data using bars that are 5 units wide, with the first one starting at 65. If we did so, we would obtain this graph:



The bar of height 9 running from 85 to 90 represents the data items that are at least 85 but less than 90, namely 85, 86, 86, 86, 89, 89, 89, 89, and 89.

Comment. We frequently use the term **bucket** to describe the pieces of the histogram. The histogram above depicts 7 buckets. The first bucket contains 2 pieces of data, those between 65 and (just below) 70. From left to right, the buckets are of size 2, 0, 0, 4, 9, 6, and 1. The buckets start at 65, and each bucket has width 5.

Just as for the bar chart, it can be helpful to label each bar (bucket) with its count. And, just as for the bar chart, it is frequently more useful to have the vertical scale represent percentages rather than counts. Here is the same histogram, but with these two changes:

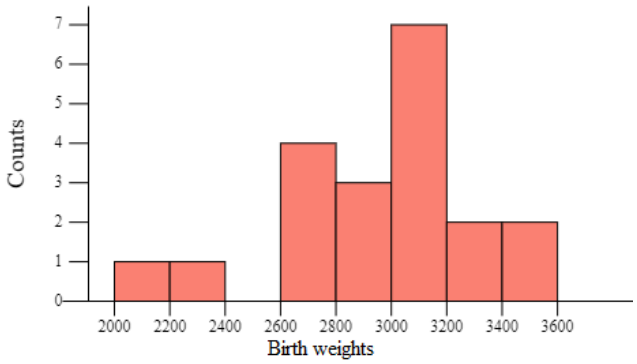


Although the histogram does not show the actual data, it is preferable to the other two for large sets of data.

Exercise 2: Create a histogram for this data set: 15, 17, 17, 11, 12, 12, 12, 18, 19, 32, 33, 33, 33, 33, 36, 40. Use buckets that start at 8 with width 5, and label each bucket with its count.

Connections to probability

Consider this histogram, which shows the birth weight (in grams) for the babies born at a certain hospital in the first 10 days of January.



Keeping in mind that, for example, the rectangle extending from 3000 to 3200 represents data starting at 3000 and ending just less than 3200, we can use the histogram to answer questions about probabilities. We give two examples that illustrate the ideas. Before we begin, we note that the histogram shows us how many babies were in the set of data: 1 in the 2000 to 2200 range, 1 in the 2200 to 2400 range, 4 in the 2600 to 2800 range, and so on, for a total of 1 + 1 + 4 + 3 + 7 + 2 + 2 = 20.

1. If one of the babies is selected at random, what is the probability that its birth weight is between 3200 and 3400 grams – that is from 3200 to just below 3400? From the graph, we see that 2 of the 20 weights fall in this range, so the answer is "2 out of 20." Mathematically, we calculate 2/20 = 0.1 or 10%.

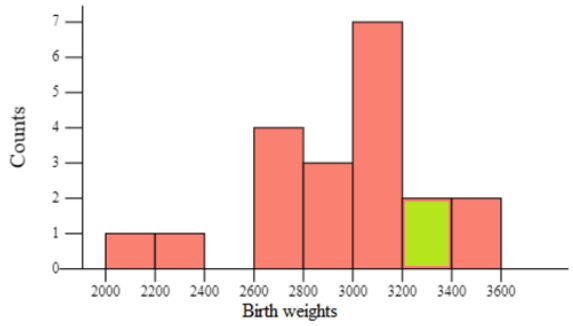
2. If one of the babies is selected at random, what is the probability that its weight is less than 2800 grams? From the graph, we see that there is a total of 1 + 1 + 4 = 6 weights in this range, so the probability is 6/20 = 0.3 or 30%.

Exercise 3: For a randomly chosen baby, calculate these probabilities:

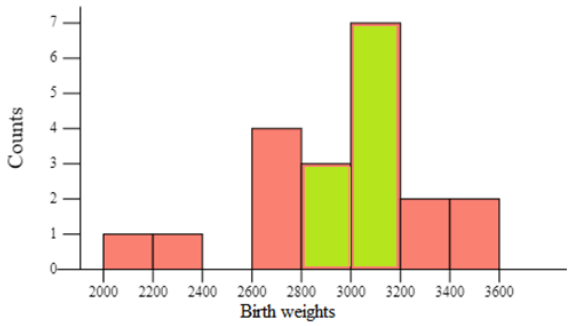
a. The weight is 3000 or higher.

b. The weight is in the 2800 to 3200 range (that is, from 2800 to just below 3200).

Note: When we have a histogram, we can think of proportions in terms of areas. For example, in the histogram above, 10% of the data (2 out of 20) is in the 3200 to 3400 range. In terms of area, 10% of the total area of the histogram is in that 3200 to 3400 rectangle. (That rectangle's area is $2 \cdot 200 = 400$, and the total area is $1 \cdot 200 + 1 \cdot 200 + 4 \cdot 200 + 3 \cdot 200 + 7 \cdot 200 + 2 \cdot 200 + 2 \cdot 200 = 4000$; and 400/4000 = 10%.) Here is a histogram in which we shade the area that corresponds to this proportion:



Similarly, this shaded histogram illustrates the probability (calculated in Exercise 3b) that the weight is in the 2800 to 3200 range (that is, from 2800 to just below 3200).

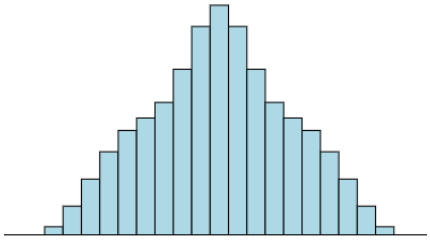


In general, probability questions can be thought of as proportion questions, which in turn can be thought of as area questions.

We will exploit this connection throughout the course.

Analyzing the graph's shape

One reason we use graphical descriptions of numerical data is that the **shape** of the graphs can display properties of the data that are not obvious just from studying the data. Many of the terms that are using to describe the shape of a histogram are illustrated by the histogram below.

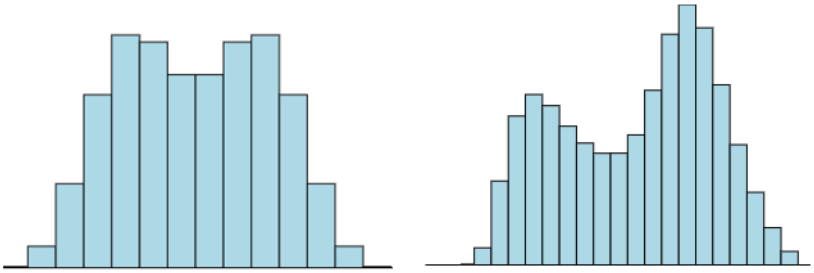


First, this graph is **symmetric**. The left half of the graph is the mirror image of the right half. This particular graph is perfectly symmetric, but we will also use the term for data where the symmetry is not quite as perfect. (We might say "roughly symmetric" or "approximately symmetric," but we will also simply say "symmetric" at times even though the symmetry is not exactly perfect.)

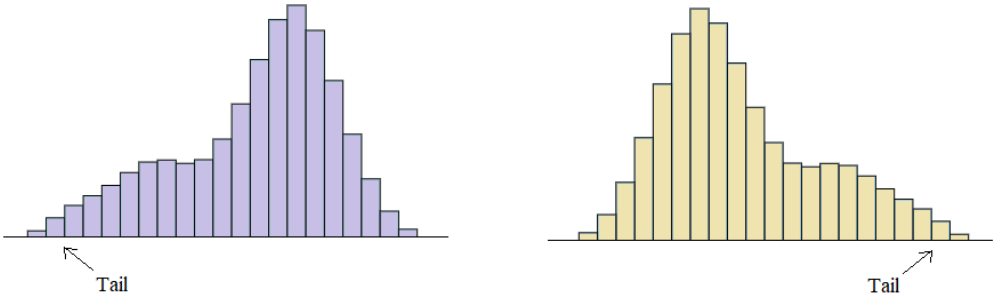
The graph has a single "peak" or high point; this is described by saying the graph is **unimodal**. We would also describe the histogram as perhaps "bell shaped," although in these lessons we will instead use the term **mound shaped** – its shape is similar to that of a mound or pile of dirt created by tossing shovelfuls of dirt onto it.

Comment. We will refer to the graph, and to the data it represents, using the term **distribution** and related words. The graph is symmetric, so we might say the data is distributed symmetrically, or that we have a symmetric distribution. Similarly, we might make the general observation that a mound shaped distribution is one that is unimodal and (at least approximately) symmetric.

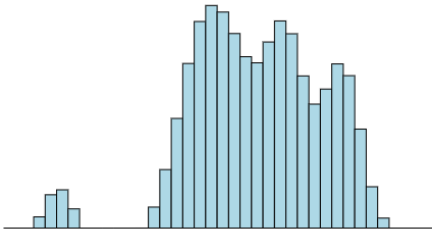
Each of the histograms shown below has two peaks, a property we will describe using the term **bimodal**. The first graph is symmetric and the two peaks are exactly the same height, but we do not need these properties to hold in order to view the distribution as a bimodal distribution. The second graph doesn't satisfy either property, but it does have two discernible, distinct peaks.



Next, consider the distributions shown below. Each has what we will describe as a **tail** trailing out away from the bulk of the data. The tail for the first histogram is to the left of the main part of the data, and we will refer to this distribution as **skewed left**; similarly, the second histogram represents a distribution that is **skewed right**.

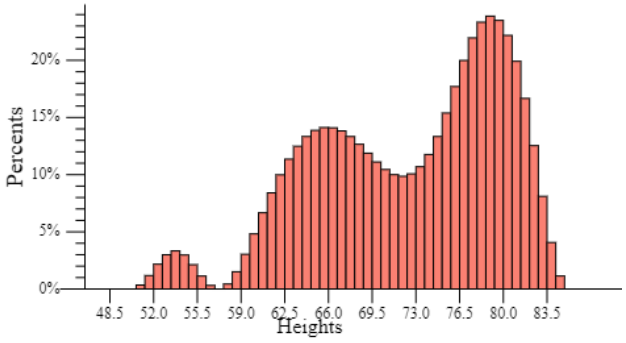


Finally, here is another histogram that is skewed left, but in this case the "tail" on the left is actually separated from the bulk of the data by several buckets containing no data. We describe this by saying that the distribution contains **outliers**, data that is significantly separated from the rest of the data.



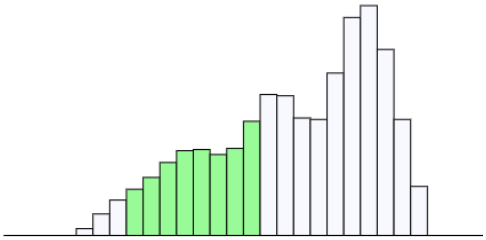
A word about sketching distributions

A standard practice when discussing distributions is to present the graph of the distribution using a smooth curve. For "real-world" distributions with lots of data, the histogram may look almost like a smooth curve. Here, for example, is a graph for which your brain may very well automatically "see" a curve representing the shape of the graph.

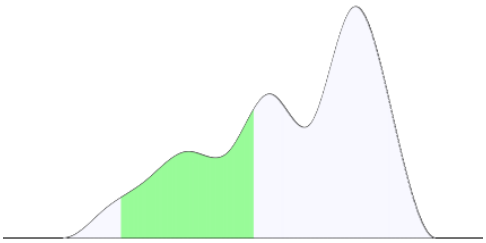


Notice, by the way, that this distribution is definitely skewed left, with a group of data at the far left which seems (almost at least) to consist of outliers.

Even when the histogram is not nearly so smooth in and of itself, it can be useful to approximate the histogram with a smooth curve. The histogram on the left represents, for example, the shading of a histogram which might illustrate the process of calculating a particular probability as we did in earlier exercises. On the right, we use a smooth curve to display the same picture but in a simplified manner.



Many buckets



Smooth curve approximation

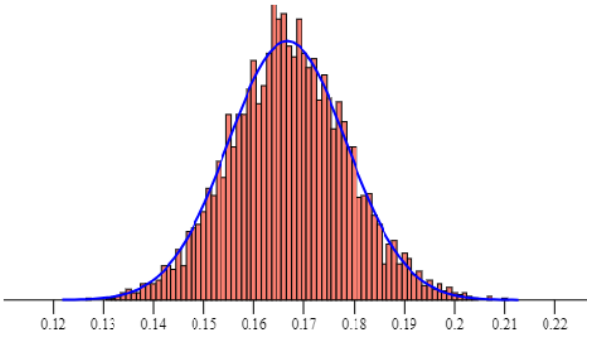
As another example, here again are the distributions we used earlier to illustrate the concepts of symmetric, skewed left, and skewed right distributions – but this time we have simplified the picture by showing smooth curve approximations to the distributions.

Symmetric

Skewed left

Skewed right

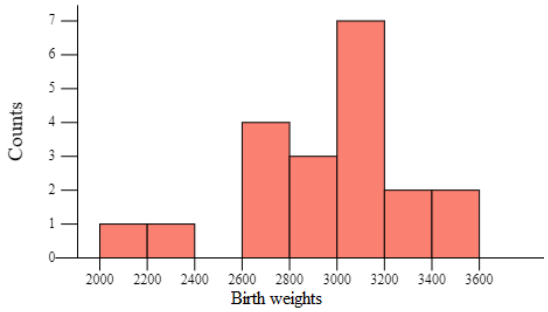
Finally, here is a diagram taken from some of the material in Lesson 6, which again illustrates the idea of using a smooth curve to approximate a distribution.



We will frequently follow this practice without comment in the remainder of the course.

2.3 – Numerical Measures of Center and Spread

In addition to questions of shape, the graph of a set of data gives a preliminary indication of the center of the data, and how spread out the data is. For example, here is the histogram we discussed earlier in this lesson, which describes the birth weight (in grams) for the babies born at a certain hospital in the first 10 days of January.



The data seems to be centered near 2900 grams, and it is spread from near 2000 grams to near 3600 grams. In this section we develop some numerical ways to describe the center of the data, and to describe how spread out the data is.

For data that is somewhat symmetric and with few or no outliers, the most useful measures of center and spread are the *mean* and *standard deviation*, respectively. On the other hand, when there are significant outliers or significant skewness, we tend to prefer the *median* and *interquartile range (IQR)* as our measures of center and spread. As you will learn, we have these preferences because the mean and standard deviation can be dramatically altered by the presence of skewness or outliers. The median and IQR are less susceptible to being influenced by skewness or outliers.

Mean and median

The *mean* is what most people usually refer to when they talk about the *average* of a set of data. It is calculated as the sum of the data items, divided by the count of the data items. Symbolically, we write

$$\bar{x} = \frac{\Sigma x}{n}$$

where x stands for "data item," $\bar{x}$ indicates the "average of the data items x," Σ indicates "sum," and n is the number of data items.

Example. Calculate the mean of the following set of data:

154 165 167 173 174 177 180

Solution: Using an ordinary calculator (or by hand), we can find the sum of the numbers, which is 1190. There are seven numbers in the data set, so the mean is $\bar{x} = \frac{1190}{7} = 170$.

Sometimes when people talk about the average of a set of data, they have in mind the *median* instead of the mean. The median is easily calculated once the data has been written down in order from lowest to highest. If there are an odd number of pieces of data, it is the one that is in the middle of the list; otherwise it is the average of the two pieces of data in the middle of the list.

Example. Find the median of the following set of data. To simplify your task, the data has been sorted into increasing order, and the position number for each piece of data is displayed along with the data itself.

Position	1	2	3	4	5	6	7	8	9
Data	23	31	31	46	53	55	69	71	90

Solution. There are an odd number of data items, so the median is the middle piece of data, namely the 5^{th} data item (4 pieces of data precede the 5^{th}, and 4 pieces of data follow the 5^{th}). So the median is 53.

Example. Find the median of the following set of data.

Position	1	2	3	4	5	6	7	8
Data	23	31	31	46	53	55	69	71

Solution. Now there are an even number of data items, so the median is halfway between the two middle pieces of data, namely the 4^{th} and 5^{th} data items (3 pieces of data precede the 4^{th}, and 3 pieces of data follow the 5^{th}). So the median is halfway between 46 and 53. One way to do this calculation is to find the average of the two numbers, $\frac{46+53}{2} = 49.5$. So the median is 49.5.

The applet at the following link provides practice calculating the median of a set of data.

Calculating medians

The effect of skewness and outliers

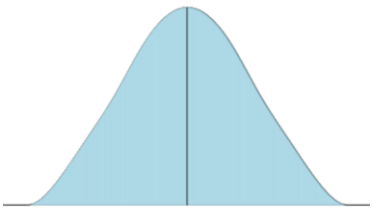
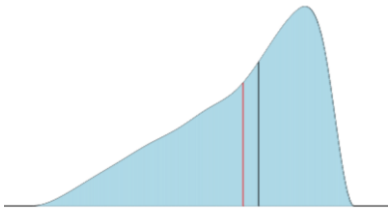
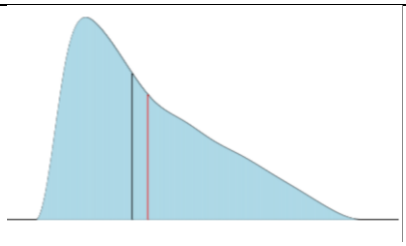
As we mentioned earlier, for data that is somewhat symmetric and with few or no outliers we will generally use the mean as our measure of center. On the other hand, when there are significant outliers or significant skewness, we tend to prefer the median as the measure of center. As a simple example of the reason for this preference, consider these two sets of data.

Set #1: 154 165 167 173 174 177 180
Set #2: 14 165 167 173 174 177 180

They are identical, except for the first item in each set. In the second set, the 14 is a significant outlier. Earlier we calculated the mean of the first set as 170. The mean of the second set is 150. The existence of the one outlier has pulled the mean down below all the data except the outlier. Similarly, in a calculation of mean salary in a neighborhood, the existence of a single resident who as a major company CEO has a salary in the millions will dramatically pull the result upward. In general, the mean is "pulled" toward the outlier.

On the other hand, the median for both data sets is the same – the middle value is 173 for both sets. In general, the median is not unduly influenced by the presence of outliers.

The same pattern can be observed for data that is skewed. When data is perfectly symmetric, the mean and the median are the same. If the data is skewed, the mean is "pulled" toward the tail of the distribution. In these diagrams, the black line is located at the median of the data, the red line at the mean. We describe the fact that the median is not unduly affected by the skewness by saying that the median is **resistant**; on the other hand, the mean is **not resistant**.

		
Symmetric: mean = median	Skewed left: mean < median	Skewed right: mean > median

Standard deviation

When the mean is used to measure the center of a set of data, the *standard deviation* is typically used to measure the spread. The standard deviation measures how far the set of data items is from the mean of the data. A naïve approach might be to just calculate the distance $x - \bar{x}$ for each data item x, and add these up. There are two problems with this. First, the negative distances "cancel" the positive distances, giving a total of 0. Second, it doesn't take into account the size of the data set. To deal with the first problem, we square each distance before adding them up; to deal with the second, we divide this sum by $n - 1$. The result is the so-called *variance*. The *standard deviation* is the square root of the variance. In symbols, using s to indicate standard deviation[2]:

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

This is a fairly complicated calculation, especially for large sets of data. The important thing to remember is that it has this property: *its value is larger for sets of data that are more spread out*. Thus it is an appropriate measure of spread. Usually you will use some form of technology to do the actual calculations. In Section 2.4 of this lesson we indicate the necessary steps for an online calculator, developed by the author, that accompanies these lessons. Your instructor may also supply information about additional types of technology, and may indicate the particular technology tool or tools you will need to learn how to use.

Range

Perhaps in a math class in elementary school you learned about a second measure of spread, the *range*. To calculate the range, you simply subtract the smallest piece of data from the largest. While easy to calculate, this measure is unfortunately extremely sensitive to outliers; it is definitely *not* resistant. If 1000 people attending a wedding are all 25 years old, and one person is 90 years old, the range of ages would be 65 years. While this is true, knowing that the range is 65 does not give a good piece of summary information about the spread of the data. In fact, except for that one person, the ages are not spread out at all.

IQR, and the 5-number summary

As we have stated, for symmetric data , and especially for what we have called *mound-shaped data* (data whose histogram resembles a mound of dirt), the mean and standard deviation are the preferred measures of center and spread. However, when the data is skewed or has significant outliers, the median

[2] You may wonder why we divide by $n - 1$ rather than by n. The full discussion of this question is well beyond the scope of these lessons. Perhaps the most important point is that we will be using the standard deviations of samples from a large population to estimate the standard deviation for the entire population, and statisticians have shown that dividing by n would yield estimates that are *biased* – they are systematically lower than they "should be."

is preferred as the measure of center because it is resistant. In this case, the preferred measure of spread is the **interquartile range**, or **IQR**..

The interquartile range measures the distance between the data item that is ¼ of the way from the smallest to the largest, and the data item that is ¾ of the way from the smallest to the largest. Put another way, it measures how spread out the middle half of the data is. Here again are the two sets of data we used earlier for calculating means and medians:

Set #1:	154	165	167	173	174	177	180
Set #2:	14	165	167	173	174	177	180

To calculate the IQR we first calculate Q1, also called the **first quartile.** This is the data item that is ¼ of the way from the smallest to the largest. We also calculate the **third quartile** Q3, the data item that is ¾ of the way from the smallest to the largest. In the following examples we illustrate the method used when doing the calculation by hand; some versions of technological calculators use a more complicated algorithm which may yield slightly different results.

Example. Calculate the IQR for Set #1 given above.

Solution. We calculate the median, use that to calculate the first and third quartiles (Q1 and Q3), then finally calculate the IQR.

1. Find the median, in this case 173. The remainder of the data is split into two halves:
   Those below the median: 154 165 167
   Those above the median: 174 177 180
2. Q1 is the median of those below the median, in this case 155.
3. Q3 is the median of those above the median, in this case 177.
4. IQR = Q3 – Q1 = 22.

Like the median, the IQR is resistant. Why? The following exercise gives a sense of the answer.

Exercise 4: Calculate the IQR for Set #2 given above. Comment on the effect of the outlier (the data item 14).

Example. Calculate the IQR for this set of data:

154 165 167 173 174 177 180 183

Solution. Because there are 8 pieces of data, the median is halfway between the fourth and fifth piece of data (173.5 in this case), splitting the data into these two pieces:

Those below the median: 154 165 167 173
Those above the median: 174 177 180 183

Q1 is the median of the first four, halfway between 165 and 167 (166). Q3 is the median of the last four, halfway between 177 and 180 (178.5). IQR = 178.5 – 166 = 12.5.

When the median and IQR are used to measure center and spread, it is common to identify five pieces of information about the data, called the ***5-number summary:***

the smallest data item
Q1
the median
Q3
the largest data item

The applet at the following link provides practice calculating the first and third quartiles, then using them to calculate the IQR.

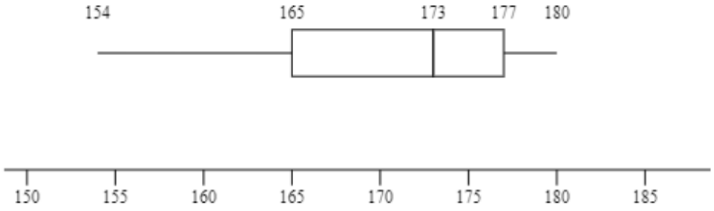
Calculating Q1, Q3, and IQR

Another type of graph: the box plot

So far we have studied several ways to graph numerical data. We now examine one final graphing method. Technically, this method does not graph the data but rather some summary information about the data.

A box plot is a graphical depiction of the 5-number summary for the data. It can be especially useful in analyzing the same variable for two different groups of people. For example, box plots might be used as a tool to examine whether college graduates in the Northeast obtain higher starting salaries than those in the Midwest.

The 5-number summary for Set#1 is: 154, 165, 173, 177, 180. Here is the corresponding box plot for that data; as for the histogram, it can be useful to supply labels for the 5-number summary.



The *box* extends from Q1 to Q3 with a vertical line at the median. There are *whiskers* from Q1 down to the minimum value and from Q3 up to the maximum value.

2.4 – Using Technology

This section provides information on using an online calculator (provided by the author of these lessons) to enter numerical data, then graph that data and calculate the mean, standard deviation, and so on[3]. Once you have read this material, you should practice using it. The apps at the following links provide tools for practicing. The first app is one you have already used to calculate the IQR by hand; this time you should use the calculator to do the calculations. The second app asks you to calculate mean and

[3] The author of the lessons has actually provided two calculators you can use in conjunction with these lessons. The first one, described here, is designed for "ad-hoc" practice, where you enter a small amount of data and carry out certain calculations with that data. The second provides the ability to analyze larger data sets in the form of data files, where each record of the file contains all the information for one person or other entity. See Section 2.7 of this lesson for a brief introduction to that calculator.

standard deviation for a set of data. For both apps, once the data is entered you may wish to also practice some of the other activities, such as creating a histogram for the data.

Calculating Q1, Q3, and IQR
Calculating mean and standard deviation

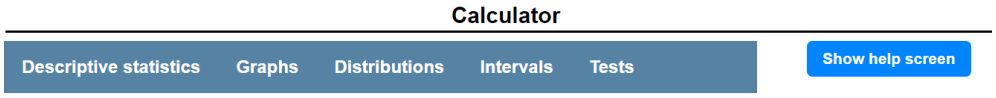
For our explanation of how to use the online calculator, we will work with the following set of numerical data.

1 1 3 3 4 5 5 5 5 5 10

We will first calculate the data set's mean, standard deviation, median, and IQR. Later on, we will create some graphs for this data. To begin, open the calculator using this link:

Statistical calculator

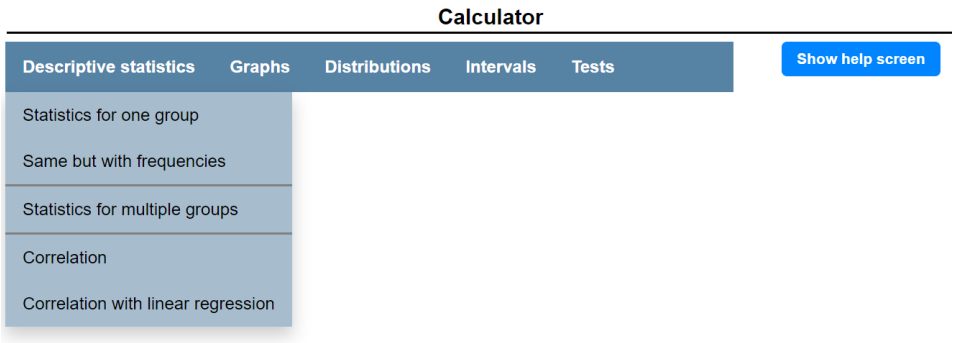
This will bring up the following menu structure:



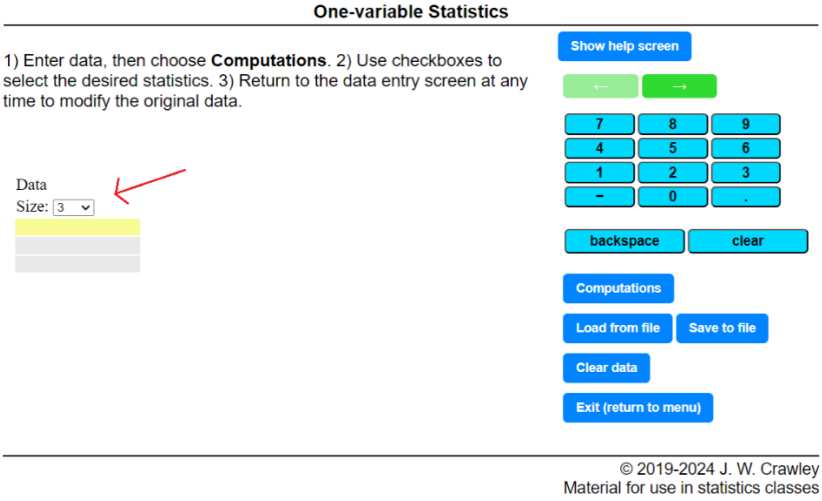
For our present purposes, we will be using options on the *Descriptive statistics* menu, and later some options on the *Graphs* menu.

Statistics for a single numerical variable

Click on the *Descriptive statistics* to obtain this submenu:



We will be calculating some statistics for a single group of data, so we choose the option Statistics for one group, resulting in this screen; we have added a red arrow to highlight our next step.



Notice that there is room to enter three pieces of data. Since we have 11 numbers in our data set, we first need to use this drop down list to change the size to 11. Then, using the provided key pad (and the green arrows for switching between data items), we enter the data as shown here.

Data
Size: 11

1
1
3
3
4
5
5
5
5
5
10

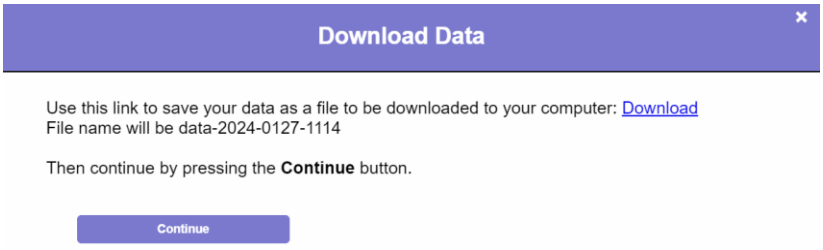
Comment. If you happen to be using a computing device, such as a PC, which includes a keyboard, you can use that keyboard rather than the online keypad to enter the data.

When we click the *Computations* button, we see a screen that displays the size, mean, and standard deviation for the data, with check boxes to allow us to select which statistics should be displayed. We remove the size, and add the median and IQR, with these results:

☐ Size	☐ Min	☐ Q3	☐ Σx
☑ Mean	☐ Q1	☐ Max	☐ Σx^2
☑ Std. dev.	☑ Median	☑ IQR	☐ Range

Mean	4.2727
Std. dev.	2.4532
Median	5
IQR	2

In the next subsection we will create a histogram for the same data. Since we want to use the same data, it will be convenient to save a copy of the data in a text file. To do this, click on *Modify data* to return to the screen where you entered the data, then click on *Save to file*. This generates the following dialog box:



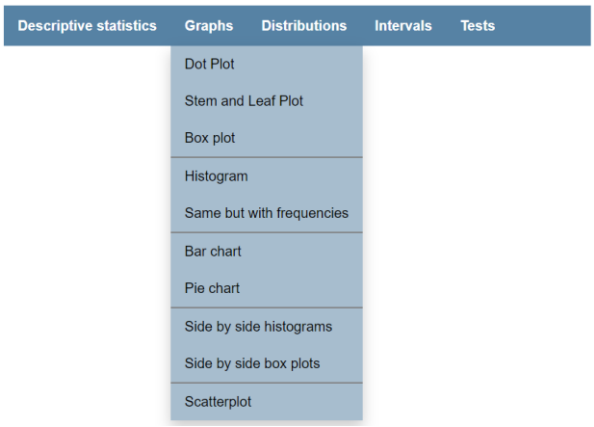
When you click on the *Download* link, a text file will be downloaded to your computer or tablet – the exact mechanism for this depends on the type of device you have. In any case, when the author saved the data, the file name was "data-2024-0127-1114" and its contents were as shown here:

> data-2024-0127-1114 created by "One-variable Statistics" application
> Data,1,1,3,3,4,5,5,5,5,5,10

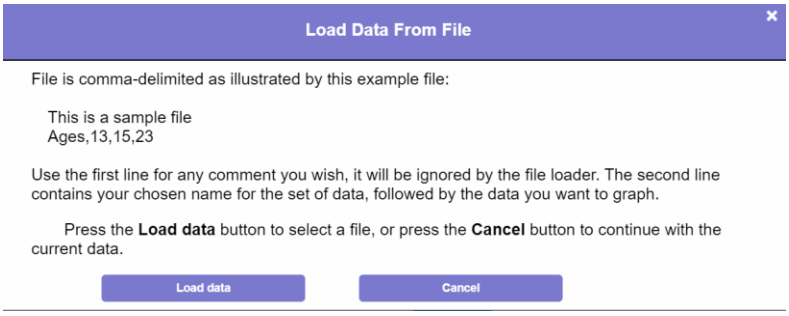
The first line of the file is simply a comment explaining how the file was created; the second line contains a description (just the generic "Data" in this case), followed by a list of the data in comma-delimited format. No matter the type of your device, your file will be generally the same as that shown above. The name will include the year, date, and time the file was generated.

Graphs for a single numerical variable

Click on *Exit (return to menu)*, then on the *Graphs* option in the menu, to obtain this submenu for graphing data:



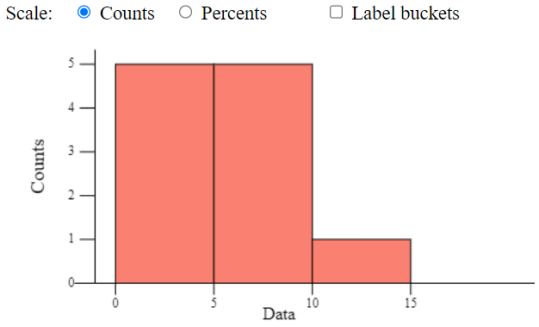
This menu has several graphing options for a single set of data; we begin with the *Histogram* option. When you click on that option, you will see a data entry screen essentially identical to what you saw earlier. We now have two options for entering the data – we can do exactly what we did in the previous example, or we can take advantage of the fact that we have saved that data to a text file. We choose the latter approach. So, click on *Load from file* to obtain this dialog box:



Comments:

1. In general for this calculator, whenever you click on *Load from file* the resulting dialog box will describe the necessary format for the file to be loaded.
2. We have created the file we want to use by entering data into the calculator, then selecting *Save to file*. We will load that same file.
3. In general, the calculator does not care how the file was created. You can use any suitable editor – for example, Notepad in Windows – to create the file, provided it is a text file.

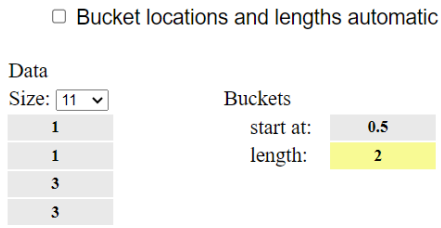
When you click the *Load data* button, the interface you see will depend on your particular device. In any case, locate the file that was downloaded earlier and select that file. The data will be loaded into your data entry screen; clicking *Computations* will generate this histogram:



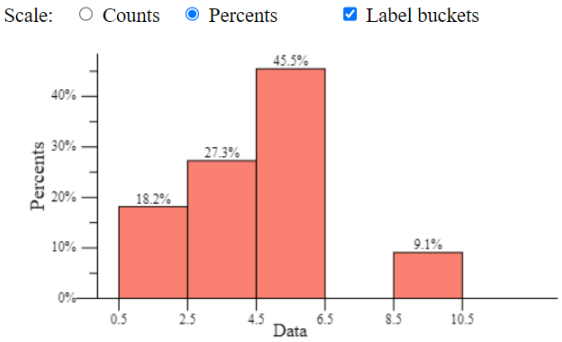
The calculator offers the option to label each "bucket" with its count, and also the option to use percents rather than counts. The calculator has also decided on the buckets:

From 0 to just less than 5
From 5 to just less than 10
From10 to just less than 15

That is, the buckets start at 0 and have width 5. If you prefer, you can return to the previous screen (click *Modify data*) and set your own start and width. For example, the author unchecked the *Bucket locations and lengths automatic* checkbox and chose to have the buckets start at 0.5 with width 2, as shown here:



which resulted in this histogram (with buckets labeled and the scale switched to percent):



Exercise 5: The *Graphs* submenu contains other graph types suitable for a single set of numerical data.

a. Use the calculator to create a dot plot, and a box plot with labels, for the same data set.

b. Create a stem-and-leaf plot for the following data set:
35, 56, 53, 45, 39, 27, 39, 35, 77, 88, 80, 79, 80, 53, 11, 19

c. Create a histogram for the data set from part (b), with the buckets starting at 9.5 with width 10.

Exercise 6: Immediately below the "Statistics for one group" option in the calculator is an option *Same but with frequencies*; the same is true for the *Histogram* option. Both provide an alternative for entering data where there are many duplicate values. Use those options to calculate the mean and standard deviation, and to create a histogram, for the following set of data for 24-hour temperature change for a TV station's 60 weather recording sites on a recent day:

Change	Frequency (count)
+3	20
+2	13
0	15
–3	10
–4	2

Notice that you could enter 60 values: +3 20 times, and so on – but the menu options referred to allow for easier data entry. After creating the data, save the file, and observe its structure.

Pie chart and bar chart for a categorical variable

This particular calculator has no way to directly enter non-numerical data, so in order to handle categorical data you will need to first create a text file which contains the frequency table for the variable. For example, here is a text file that contains the data for the class year variable described in Lesson 1 and in Section 2.1 of this lesson. As for our other text files, the first line is just a comment. Each subsequent line contains a value for the categorical variable and a count (frequency for that variable.)

```
This is the class year data for a recent intro to stats section
Freshman,9
Sophomore,14
Junior,5
Senior,2
Other,0
```

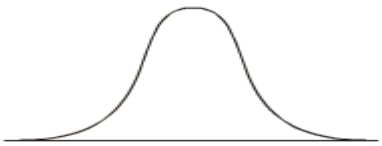
If you create a text file identical to this and load it into the pie chart and bar chart options in the calculator, you will obtain the graphs that were previously displayed in Section 2.1 of this lesson.

Exercise 7: In another introductory statistics class the instructor had a first-day survey that asked about political party affiliation, with choices Dem, Rep, Ind, and Other. In that class 77 chose Dem, 66 chose Rep, 36 chose Ind, and 17 chose Other. Create a suitable text file, then use the calculator to create a pie chart and a bar chart for this data.

2.5 – The Empirical Rule

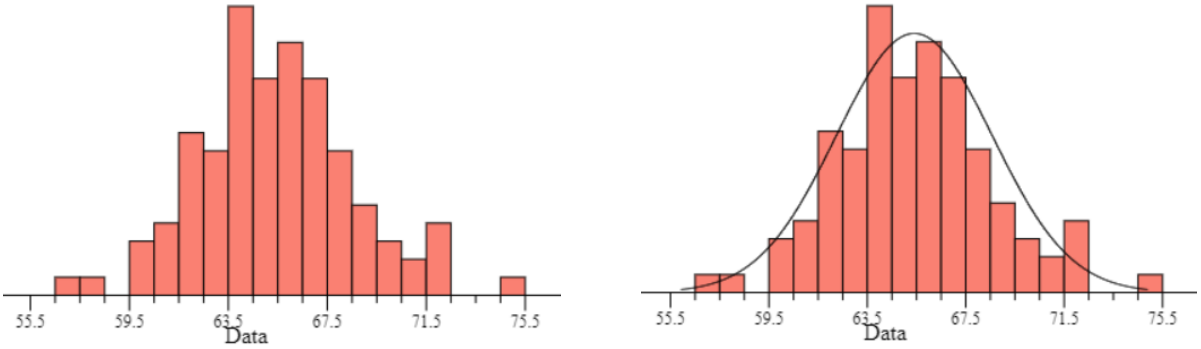
Properties of mound-shaped distributions

Think for a moment about digging a hole, and tossing the dirt aside in a single location. What happens? The dirt begins to pile up into a shape that looks something like this:



The pile of dirt is unimodal and symmetric – although perhaps not as perfectly unimodal and symmetric as this graph!

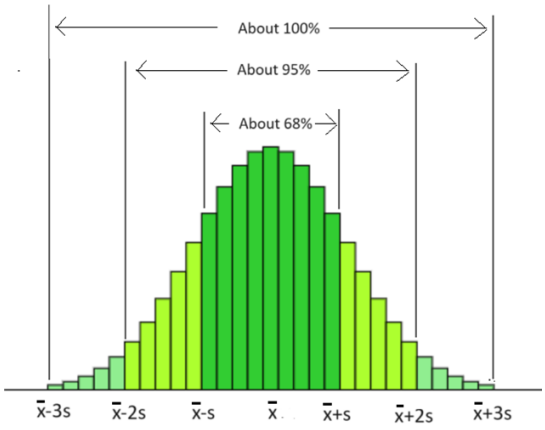
Many histograms for real-world data match this shape, at least approximately. When they do, we refer to the distributions as *mound-shaped* (or sometimes *bell-shaped*) distributions. For example, the graphs below show the heights for the female students in all sections of an introductory statistics course in a recent semester. The first graph shows the histogram, and the second shows how it approximately matches up with a smooth curve similar to that above.



Part of the importance of the standard deviation s is that it, together with the mean $\bar{x}$, describe what happens in practice for mound-shaped distributions. Statisticians have observed the following phenomenon, and they have named it the *Empirical Rule*. (It really isn't a rule, just a pattern that has been observed – but we will follow conventional terminology and use the term *rule*.) Here is a verbal description – exercise 8 contains a graphical rendition.

- Approximately 68% of the data lies between $\bar{x} - s$ and $\bar{x} + s$ (that is, within one standard deviation of the mean).
- Approximately 95% of the data lies between $\bar{x} - 2s$ and $\bar{x} + 2s$ (that is, within two standard deviations of the mean).
- Approximately 100% of the data lies between $\bar{x} - 3s$ and $\bar{x} + 3s$ (that is, within three standard deviations of the mean).

Exercise 8 : Use this graph to help you answer the following questions.

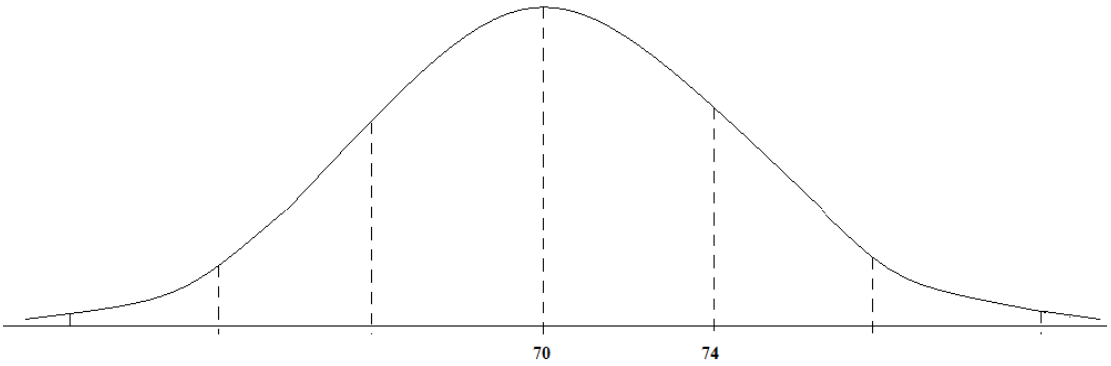


a. If about 68% of the observations are within one standard deviation of the mean, then about _____% are **not** within one standard deviation of the mean. Of these, about _____% are larger than $\bar{x} + s$, and about _____% are smaller than $\bar{x} - s$.

b. Because about 95% of the observations lie between $\bar{x} - 2s$ and $\bar{x} + 2s$, the largest _____% of the observations are larger than $\bar{x} + 2s$.

c. About _____% of the observations are *smaller* than $\bar{x} + 2s$. **Hint**: These are the one that are *not* larger than $\bar{x} + 2s$. Use your answer to the previous question.

Exercise 9: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use this information to finish labeling the following graph with numerical values for the mean ±1, 2, and 3 standard deviations. Then use the empirical rule and the symmetry to fill in the blanks.



a. Approximately 68% are between _____ inches and _____ inches tall. Therefore, about 32% are either taller than _____ inches or shorter than _____ inches, with about 16% in each of these groups.

b. Approximately 95% are between _____ inches and _____ inches tall. Therefore, about 5% are either shorter than _____ inches or taller than _____ inches, with about 2.5% in each of these groups.

c. Almost everyone is between _____ inches and _____ inches tall.

d. Approximately what percent are in each indicated height range:

i. _____ % between 66 and 74 inches
ii. _____ % between 66 and 70 inches
iii. _____ % between 70 and 74 inches
iv. _____ % less than 70 inches
v. _____ % less than 74 inches
vi. _____ % greater than 74 inches
vii. _____ % greater than 78 inches
viii. _____ % less than 78 inches
ix. _____ % between 74 and 78
x. _____ % less than 66 inches
xi. _____ % less than 62 inches
xii. _____ % between 62 and 74 inches

e. Anyone taller than 78 inches is in the tallest _____%.

f. Anyone shorter than _____ inches is in the shortest 16%.

Connections to probability

In this brief section we just want to remind you that for any situation involving proportions it is equally possible to think in terms of probabilities. For example, in the previous exercise the empirical rule tells us that about 68% of adult male heights lie between 66 inches and 74 inches, with 16% less than 66 inches and another 16% greater than 74 inches. These statements can be reworded as probabilities. A handy notation, used by the book, is P(…..) to stand for "the probability of ….." So, "P(taller than 74)" is shorthand for "the probability that a randomly selected adult male is taller than 74 inches." Using this notation, we can restate the previous percentage observations as follows:

P(66 to 74) = 68%, or 0.68
P(greater than 74) = 16%, or 0.16
P(less than 66) = 16%, or 0.16

Recall that we more frequently use decimal notation (0.68) rather than percentage notation when we write probabilities.

Exercise 10: Refer to Exercise 9, and its solution, to calculate these probabilities.

a. P(66 to 74)	b. P(66 to 70)	c. P(70 to 74)
d. P(less than 70)	e. P(less than 74)	f. P(greater than 74)
g. P(greater than 78)	h. P(less than 78)	i. P(74 to 78)
j. P(less than 66)	k. P(less than 62)	l. P(62 to 74)

The *z*-score

It turns out that mound-shaped distributions occur quite frequently in naturally occurring data. Moreover, a specific mound-shaped distribution called the *normal distribution* is very important in the theory of statistical inference, to be studied in the second half of this course.

For mound-shaped distribution, knowing how many standard deviations separate a piece of data from the mean gives a good indication of approximately where the data lies. For example, data which is more than two standard deviations away from the mean is either in the smallest 2½% or the largest 2½% of the data, since approximately 95% of the data is closer than two standard deviations. Because of this, it is useful to calculate the separation between a piece of data and the mean, in terms of standard deviations. The result of this calculation is called the *z*-score for the data. The *z*-score shows not only how far away from the mean the data lies, but also in which direction. The score is positive if the data is larger than the mean, negative if the data is smaller than the mean). So, for example:

A *z*-score of 3 indicates the data is 3 standard deviations above the mean.
A *z*-score of −2 indicates the data is 2 standard deviations below the mean.
A *z*-score of 0.15 indicates the data is 0.15 standard deviations above the mean.

Notice that in terms of *z*-scores, the empirical rule indicates that:

- Approximately 68% of the data will have a *z*-score between −1 and +1.
- Approximately 95% of the data will have a *z*-score between −2 and +2.
- Approximately 100% of the data will have a *z*-score between −3 and +3.

The calculation for the *z*-score is simple: 1) calculate the distance from the data item to the mean, and 2) find how many standard deviations this difference represents, by dividing by the standard deviation. Using *x* to stand for the data item, we have:

$$z = \frac{x - mean}{standard\ deviation}$$

Given the data item, we can calculate the *z*-score. Conversely, if we know the *z*-score we can calculate the data item by plugging in what we know (*z*, the mean, and the standard deviation) and solving for the unknown variable *x*.

Example. Heights are measured for a group of women. The distribution is mound shaped with mean 64 inches and standard deviation 3.4.

a. Find the *z* score for Sue, whose height is 60 inches.
Solution. For this problem, *x* is 60, the mean is 64, and the standard deviation is 3.4. We calculate the *z* score as:

$$z = \frac{x - mean}{standard\ deviation} = \frac{60-64}{3.4} = -\frac{4}{3.4} = -1.1765$$

We have rounded to 4 places, which we will frequently do for *z* scores. Another option might be to round to 2 places, yielding –1.18.

b. Betty's *z* score is 1.47. What is her height? Round to the nearest inch.
Solution. For this problem, *z* is 1.47, the mean is 64, and the standard deviation is 3.4. We calculate the *x* score by substituting these values and solving for *x*.

$$1.47 = \frac{x - 64}{3.4}$$

$$x - 64 = 1.47(3.4)$$

$$x = 64 + 1.47(3.4) = 69 \text{ inches}$$

Exercise 11: Adult female heights for a different group of women have a mean of 65 inches and a standard deviation of 3.5 inches.

a. Cynthia is 69.5 inches tall. Calculate her *z*-score.
b. Betty is 59 inches tall. Calculate her *z*-score.
c. Mary's *z*-score is 1.96. How tall is she?
d. Connie's *z*-score is –2.56. How tall is she?

The applets at the following links provides additional practice working with *z* scores.

Calculating z scores

Finding z scores for given data values

"Unusual" observations

When you see an adult male walk into the room, you can instinctively judge his height, perhaps identifying him as "about average height" or as "unusually tall" or perhaps as "unusually short." Using the Empirical Rule, we can quantify this idea, not only for heights but for any variable whose distribution

is approximately mound-shaped. Of course, "unusual" is a vague notion, but in the framework of statistics one possible way to think of it is this: *Let us agree, for the current discussion, that the middle 95% of the data is **not** unusual*. In other words, anything within two standard deviations of the mean is not unusual; anything *not* within two standard deviations *is* unusual. The next exercise gives you some practice thinking about these ideas.

Note: The notion of *unusual* might also be expressed as *unexpected*. If we select an adult male at random, the probability is about 95% that that person's height will be within two standard deviations of the mean. We could say that we *expect* the height to be within two standard deviations of the mean. If the person is taller (or shorter) than expected, we might express that by saying, "Having the height be that large (or small) was unexpected."

Exercise 12: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use the empirical rule to fill in the blanks.

a. Approximately 95% are between _____ inches and _____ inches tall. Sketch a picture.

b. Anyone taller than 78 inches is in the tallest _____%.
c. Anyone shorter than _____ inches is in the shortest 2.5%.
d. If by "unusual" we mean unusually far away from the average of 70 inches, then both short people and tall people qualify as unusual. The 5% of most unusual people are either shorter than _____ inches or taller than _____ inches.
e. If Sam is 79 inches tall, he is unusually tall because he falls in the top _____%. If Joe is 61 inches tall, he is unusually short because he falls in the bottom _____ %. Both these people are unusual; they are in the rarest _____% of all adult male heights.
f. Sam is 79 inches tall and Bill is 81 inches tall. Both fall in the top 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Sam's or Bill's?

g. Joe is 61 inches tall and Ted is 58.5 inches tall. Both fall in the bottom 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Joe's or Ted's?

2.6 – Identifying Outliers

When we graph data, outliers are pieces of data that are "far away from" the rest of the data. Depending on whether the data is or is not relatively symmetric, there are two common ways to quantify the idea of "far away from."

When the data is reasonably symmetric, we use the mean $\bar{x}$ and the standard deviation s to measure center and spread. In this case, the guidelines for identifying possible outliers are:

- any piece of data larger than $\bar{x} + 3s$
- any piece of data smaller than $\bar{x} - 3s$

For mound-shaped data, "approximately 100%" of the data lies within three standard deviations of the mean – so in this case, data that is an outlier might be thought of informally as "really, really unusual."

On the other hand, for skewed data, we measure center and spread using the median and IQR. Using these measures, the guidelines for possible outliers are:

- any piece of data larger than $Q3 + 1.5 \cdot IQR$
- any piece of data smaller than $Q1 - 1.5 \cdot IQR$

2.7 – Data File Analysis, Part 1

As we mentioned earlier, the author of the lessons has provided two calculators for your use. In this section we begin our description of the second calculator, which is designed to work with data files similar to those that might be created by spreadsheet software. Typically spreadsheet software is able to generate a comma-delimited text file, perhaps with a .csv file extension. We will be working with a text file created in this manner.

Several semester ago we surveyed students in several introductory statistics classes on the first day of class. Here are some of the questions that we asked them:

1. What is your gender? (M) Male (F) Female
2. What is your class year? (FR) Freshman (SO) Sophomore (JR) Junior (SR) Senior
3. How many states have you visited?
4. Do you currently smoke? (Y) Yes (N) No
5. How tall are you (in inches)?
6. How many days per week do you read a newspaper?

Sixty-three students responded to the survey and their answers are contained in the text file at this link:

First day survey

To follow along with the descriptions, you should open the link, then save the file as a text file on your own device. Here are the first few lines of that text file:

```
From first day survey
C,C,N,C,N,N
Gender,Class_Year,States_Visited,Smoke,Height(in),Newspaper
F,JR,12,N,70,2
M,FR,8,N,73,0
```

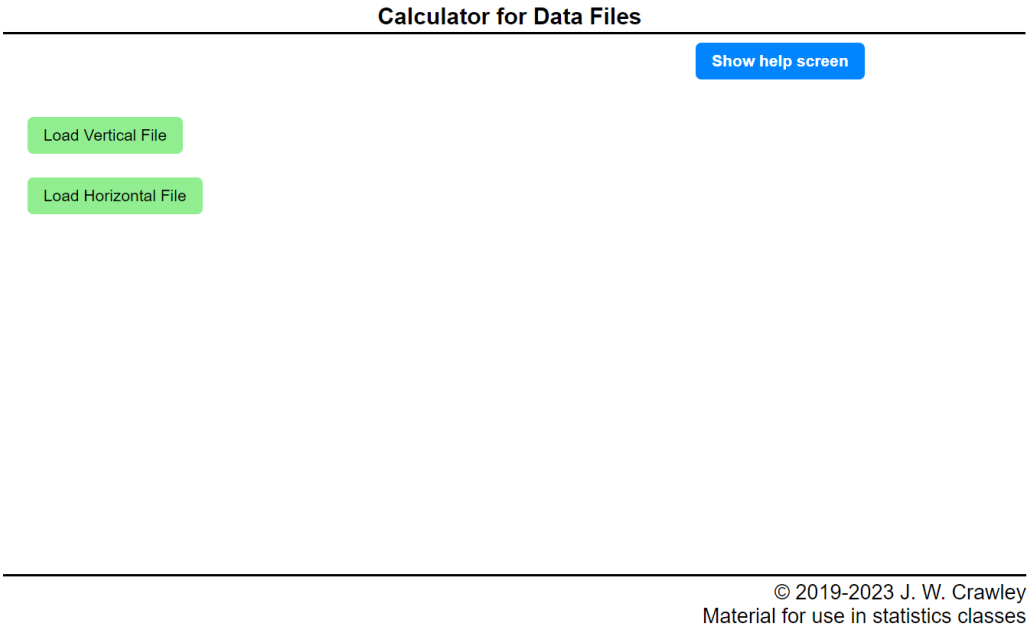
The file was created by first creating a standard spreadsheet file with 6 columns, headed with the variable descriptions Gender, Class_Year, States_Visited, Smoke, Height(in), and Newspaper. In each row below the headers we recorded one student's answers to the questions on the survey. Then we inserted two rows above the headers. The first simply contained a comment about the file ("From first day survey"). The second indicated, for each variable, whether it was a categorical (C) or a numerical (N) variable[4]. Finally, we saved the file as a comma-delimited file, then loaded that file into a text file editor, removed the extra commas in the first line of the file, and saved it as a text file. This final step was not strictly necessary – the calculator is also capable of loading a .csv file created by spreadsheet software.

[4] Many statistical packages try to figure out, from the data itself, whether a particular column is categorical or numerical. Unfortunately, that analysis is frequently incorrect. In designing this calculator, we have instead chosen to have the file creator specify the intended nature of each column of data.

With this background, we are ready to use the calculator to analyze the data generated by the responses to the survey. First, open the data file calculator using this link:

Data file calculator

This will bring up this screen:

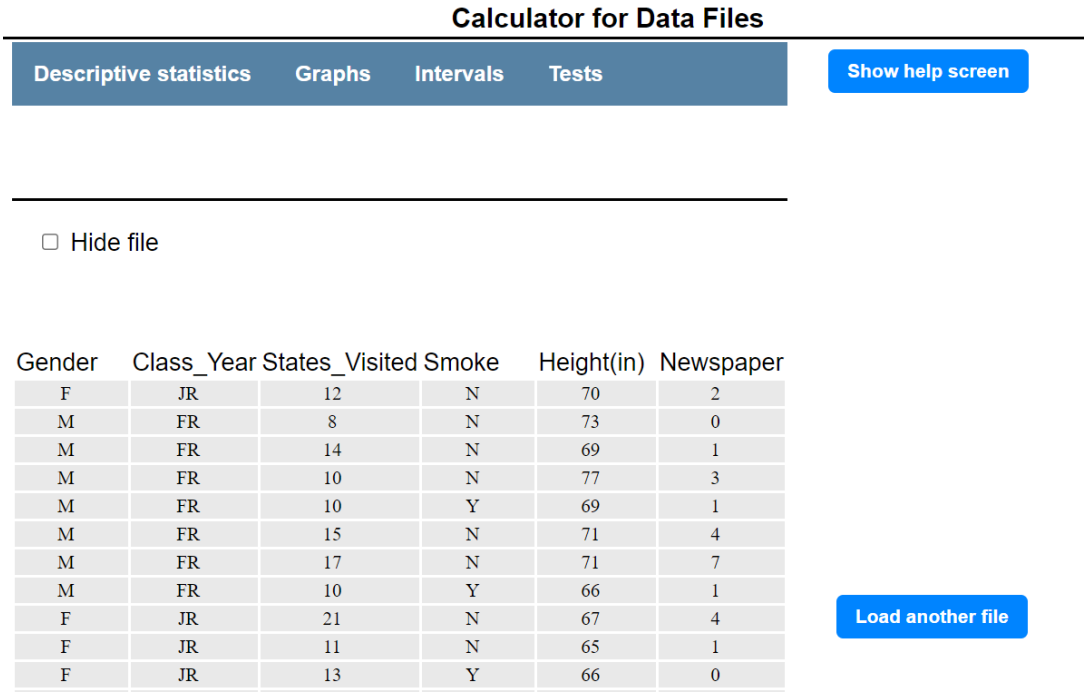


Click on the *Load vertical file* button to load the file[5]. The details of the file loading interface vary depending on the particular device you are using, but in any case once the file has loaded you will see the following:

[5] The file we are using is in what we will refer to as "vertical" format – each variable is stored in one column of the original spreadsheet. The calculator can also work with "horizontal" format files, with each variable in a row rather than a column. In this case the first few lines of the text file might look like this:

From first day survey,,,,,
C,Gender,F,M, (and 61 more pieces of data for the Gender variable)
C,Class_Year,JR,FR, (and so on)
N,States_Visited,12,8, (and so on)
C,Smoke,N,N, (and so on)
N,Height(in),70,73, (and so on)
N,Newspaper,2,0, (and so on)

If you create the file as a text file using a text editor, this may be easier than the vertical format; but for a file generated by entering data in a spreadsheet, the vertical format is the more natural format.



Calculator for Data Files

Descriptive statistics | Graphs | Intervals | Tests

Show help screen

☐ Hide file

Gender	Class_Year	States_Visited	Smoke	Height(in)	Newspaper
F	JR	12	N	70	2
M	FR	8	N	73	0
M	FR	14	N	69	1
M	FR	10	N	77	3
M	FR	10	Y	69	1
M	FR	15	N	71	4
M	FR	17	N	71	7
M	FR	10	Y	66	1
F	JR	21	N	67	4
F	JR	11	N	65	1
F	JR	13	Y	66	0

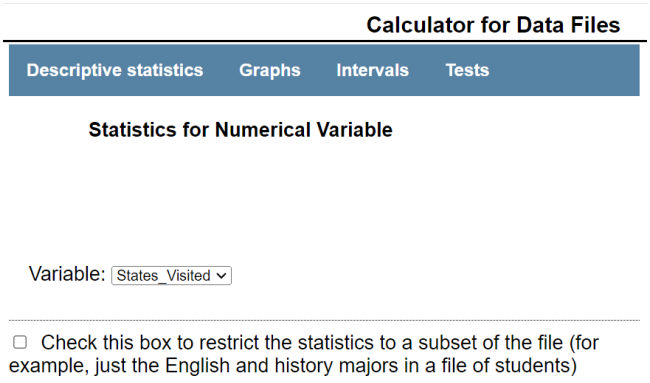
Load another file

The menu structure is quite similar to that in the other calculator covered in Section 2.4 – the big difference is that any calculation you do will refer to the data in the data file you have loaded. We will give a few examples of calculations and graphs you have studied in this lesson.

Comment: To help us focus on the interface issues, we have checked the *Hide file* checkbox prior to starting. If you wish to see the data in the file at any point, simply un-check that box.

Example. Find the mean, standard deviation, and 5-number summary for the number of states visited by students participating in the survey.

Solution. Select *Descriptive statistics* then *One-variable statistics* to obtain the following



Calculator for Data Files

Descriptive statistics | Graphs | Intervals | Tests

Statistics for Numerical Variable

Variable: States_Visited

☐ Check this box to restrict the statistics to a subset of the file (for example, just the English and history majors in a file of students)

Use the drop-down list to select the variable to be analyzed – in this case, it is already selected since it is the first numerical variable in the file. Click *Computations*, then select the desired statistics. Here is the result:

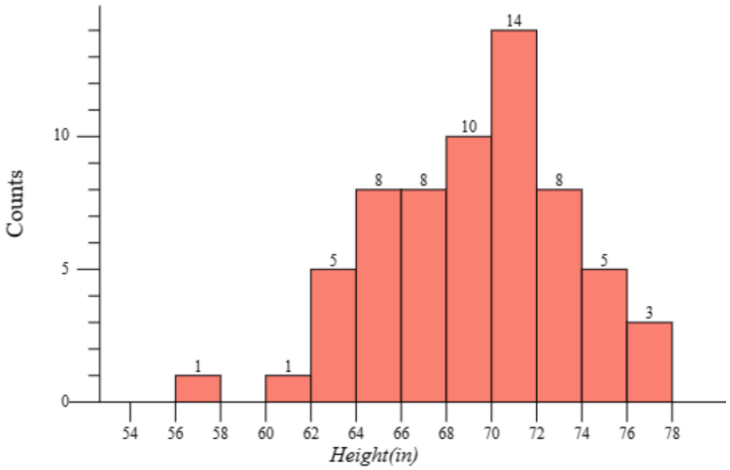
☐ Size	☑ Min	☑ Q3	☐ Σx
☑ Mean	☑ Q1	☑ Max	☐ Σx^2
☑ Std. dev.	☑ Median	☐ IQR	☐ Range

Statistics for the *States_Visited* variable

Mean	15.8571
Std. dev.	7.5303
Min	6
Q1	10
Median	14
Q3	20
Max	40

Example. Create a labeled histogram for the heights of the students. Accept the default values for the bucket locations and widths.

Solution. Use menu option *Graphs*, submenu choice *Histogram*. In the drop down list choose *Height (in)* as the variable, then click *Computations.* Here is the result:



If we wish, we can return to the menu, uncheck the *Bucket locations and widths automatic* checkbox, and choose our own values for the start and width of the buckets. For example, choosing 54 and 1 as those values yields this graph:

Exercise 13: Use the calculator to do the following:

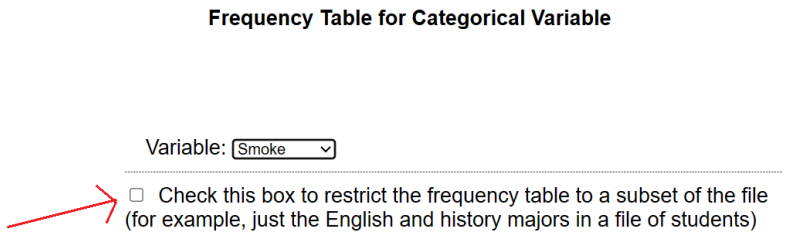
a. Find the min, max, and range for the height variable.
b. Create a bar chart for the *Class_Year* variable.
c. Create a histogram of the states visited, including the count for each bucket. Comment on the shape of the distribution.

Analyzing a Subset of the Data File

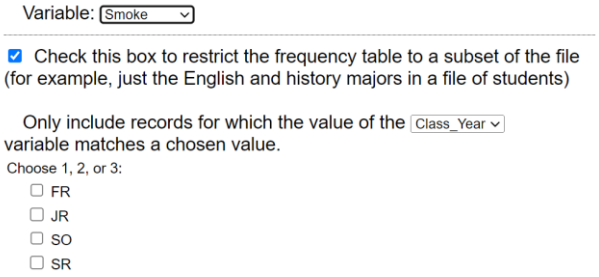
All the calculator options we have been using provide the capability to perform our calculations or graphing using a subset of the entire data file.

Example. Analyze the smoking rate for the females in the survey, by creating a frequency table and a pie chart.

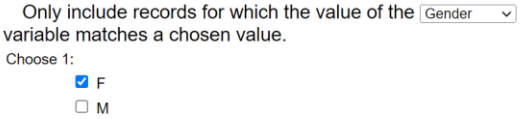
Solution. Choose *Descriptive statistics, Frequency table.* For the variable choose Smoke (we want to analyze the proportion of yes/no answers for this variable). Observe the checkbox just below the drop-down list. Each menu option we have used contains a similar checkbox.



Check that checkbox to obtain the following:



If we wanted to study smoking rate restricted to particular classes (FR, etc.) we could select those. Instead, we change the *Class_Year* choice to *Gender* and check the F box as shown here:



Clicking *Computations* yields these results:

Counts and percentages for the *Smoke* variable
Restricted to records where *Gender* is equal to:
F

Smoke	Count	Percent
N	22	91.67%
Y	2	8.33%
Totals	24	100%

We can do exactly the same using menu option *Graph, Pie chart* with variable *Smoke* and restricted to records where *Gender* is F as shown here:

Pie Chart

Variable: Smoke

☑ Check this box to restrict the pie chart to a subset of the file (for example, just the English and history majors in a file of students)

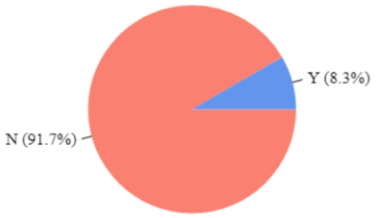
Only include records for which the value of the Gender variable matches a chosen value.

Choose 1:

☑ F

☐ M

Clicking *Computations* yields this graph:



Smoke

Restricted to records where *Gender* is equal to:

F

Example. Find the mean, standard deviation, and five number summary for the *States_Visited* variable, but only for the so-called "underclassmen" in the file – that is, the freshmen and sophomores.

Solution. Use *Descriptive statistics*, *One-variable statistics*, choose the *States_Visited* (in) variable, then checking the check box and selecting *Class_Year* as the variable, obtaining the following:

Statistics for Numerical Variable

Variable: States_Visited

☑ Check this box to restrict the statistics to a subset of the file (for example, just the English and history majors in a file of students)

Only include records for which the value of the Class_Year variable matches a chosen value.

Choose 1, 2, or 3:

☐ FR

☐ JR

☐ SO

☐ SR

You can select one, two, or three of the four available checkboxes. To obtain just the underclassmen we would check FR and SO, then click *Computations*. On the resulting output screen we check the desired statistics as shown in this result:

☐ Size	☑ Min	☑ Q3	☐ Σx
☑ Mean	☑ Q1	☑ Max	☐ Σx^2
☑ Std. dev.	☑ Median	☐ IQR	☐ Range

Statistics for the *States_Visited* variable
Restricted to records where *Class_Year* is one of:
FR, SO

Mean	13.8529
Std. dev.	5.8524
Min	6
Q1	10
Median	12
Q3	17
Max	33

Earlier we did the same calculation for the entire file, obtaining mean 15.8571 and median 14. Notice that both the mean and median are slightly lower for the underclassmen than they are for the entire set of students.

Example. Create a pie chart for the Class_Year variable, for the female students.

Solution. Choose *Graphs*, then *Histogram*. Choose the *Class_Year* variable and check the box to restrict the pie chart to a subset of the data. Select the *Gender* variable and check the F box, as shown here:

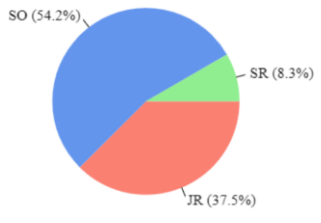
Variable: Class_Year

☑ Check this box to restrict the pie chart to a subset of the file (for example, just the English and history majors in a file of students)

Only include records for which the value of the Gender variable matches a chosen value.
Choose 1:
☑ F
☐ M

Click *Computations* to obtain this graph:

Display includes: ◉ Percents ○ Counts ○ Neither



Class_Year
Restricted to records where *Gender* is equal to:
F

Exercise 14: Use the calculator to do the following:

- **a.** Find the min, max, and range for the height variable, for the upperclassmen (juniors/seniors)
- **b.** Create a bar chart for the *Class_Year* variable, just for those who do not smoke.
- **c.** Create a histogram of the heights for the sophomores in the file.

Exercise 15: Create labeled boxplots for the height variable, for each of the following:

- **a.** The entire file
- **b.** Just the males
- **c.** Just the females

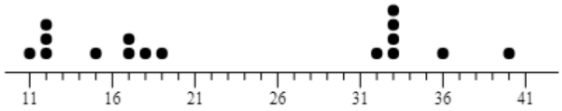
Exercise 16: Use spreadsheet software and/or a text file editor to create a comma-delimited file for the survey data shown in Lesson 1 (page 2). Then do the following for that data file:

- **a.** Create a bar chart for the *Politics* variable, showing percentages for each category.
- **b.** For the *Commute* variable, calculate the mean, standard deviation, median, and IQR, and create a histogram labeled with counts.
- **c.** Create a pie chart for the *Political party* variable, for the males in the file.

Comment. Your instructor may want you to use this calculator for data analysis projects, perhaps with data files generated by questionnaires administered in your own class. Or perhaps they will simply want you to turn in the result of carrying out exercises similar to Exercises 13 through 16. One way to obtain a copy of your output is to use the *Copy results* button, which takes you to a screen that contains the results in a form that can be copied to the clipboard. The mechanics of doing this depend on your device. For example, on a PC you might right click the output to bring up a menu that includes "Copy image" as an option. On a tablet, you might touch and hold the image on your screen to obtain a similar menu.

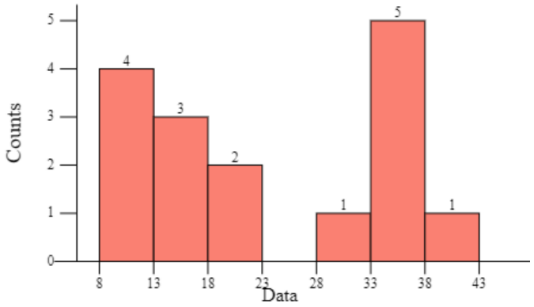
Solutions to Exercises

1: Create a dot plot and a stem-and-leaf plot for this data set: 15, 17, 17, 11, 12, 12, 12, 18, 19, 32, 33, 33, 33, 33, 36, 40



1	122257789
2	
3	233336
4	0

2: Create a histogram for this data set: 15, 17, 17, 11, 12, 12, 12, 18, 19, 32, 33, 33, 33, 33, 36, 40. Use buckets that start at 8 with width 5, and label each bucket with its count.



3: For a randomly chosen baby, calculate these probabilities:

a. The weight is 3000 or higher.

$\frac{11}{20} = 0.55 = 55\%$

b. The weight is in the 2800 to 3200 range (that is, from 2800 to just below 3200).

$\frac{10}{20} = 0.50 = 50\%$

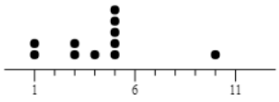
4: Calculate the IQR for Set #2 given above. Comment on the effect of the outlier (the data item 14).

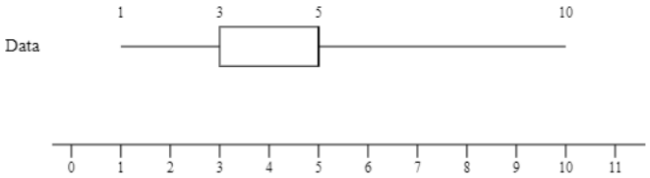
The median, Q1, Q3, and IQR are the same as those for set #1. In this example, having one outlier has no effect on the IQR.

5: The *Graphs* submenu contains other graph types suitable for a single set of numerical data.

a. Use the calculator to create a dot plot, and a box plot with labels, for the same data set.

Here are the graphs; observe that for this dataset the median and Q3 are both equal to 5.



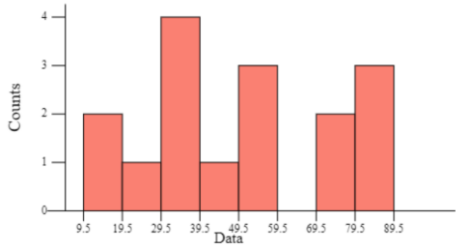


b. Create a stem-and-leaf plot for the following data set:

35, 56, 53, 45, 39, 27, 39, 35, 77, 88, 80, 79, 80, 53, 11, 19

1	19
2	7
3	5599
4	5
5	336
6	
7	79
8	008

c. Create a histogram for the data set from part (b), with the buckets starting at 9.5 with width 10.



Observe that the shape is essentially the same as that for the stem-and-leaf plot.

6: Immediately below the *Statistics for one group* option in the calculator is an option *Same but with frequencies*; the same is true for the *Histogram* option. Both provide an alternative for entering data where there are many duplicate values. Use those options to calculate the mean and standard deviation, and to create a histogram, for the following set of data for 24-hour temperature change for a TV station's 60 weather recording sites on a recent day:

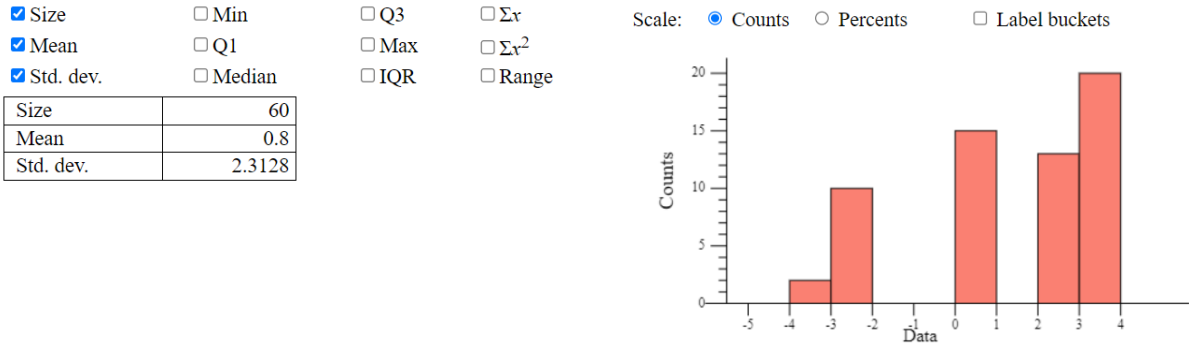
Change	Frequency (count)
+3	20
+2	13
0	15
–3	10
–4	2

Notice that you could enter 60 values: +3 20 times, and so on – but the menu options referred to allow for easier data entry. After creating the data, save the file, and observe its structure.

Here are the data entry screen, the file that was saved, the statistics, and the histogram created with automatic bucket start/width.

Data	Frequency
Size: 5	Size: 5
3	20
2	13
0	15
-3	10
-4	2

```
data-2024-0127-1327 created by "One-variable Statistics (with Frequencies)" application
Data,3,2,0,-3,-4
Frequency,20,13,15,10,2
```

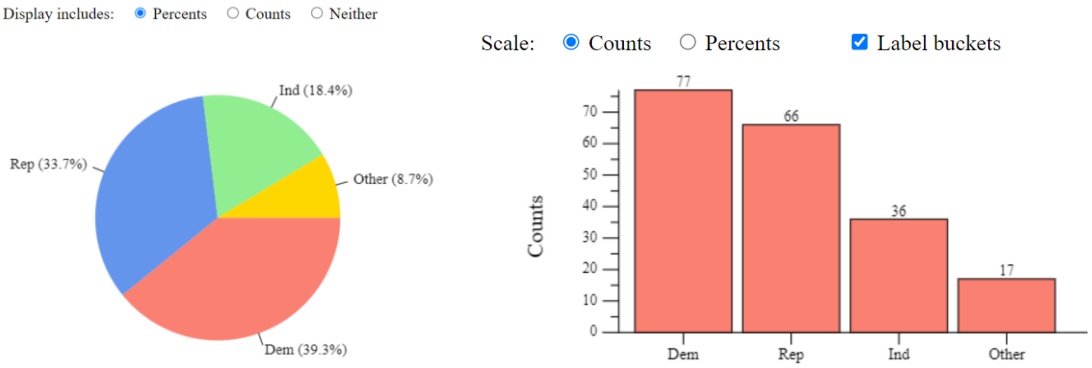


7: In another introductory statistics class the instructor had a first-day survey that asked about political party affiliation, with choices Dem, Rep, Ind, and Other. In that class 77 chose Dem, 66 chose Rep, 36 chose Ind, and 17 chose Other. Create a suitable text file, then use the calculator to create a pie chart and a bar chart for this data.

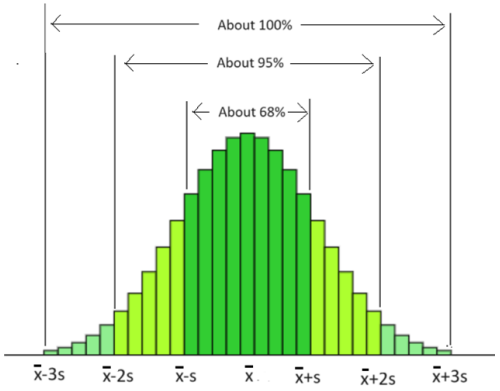
One way to set up the file is as in this example:

```
Political party
Dem,77
Rep,66
Ind,36
Other,17
```

With this file, here are the two graphs (if you use a different order for the categories, your graphs may vary):

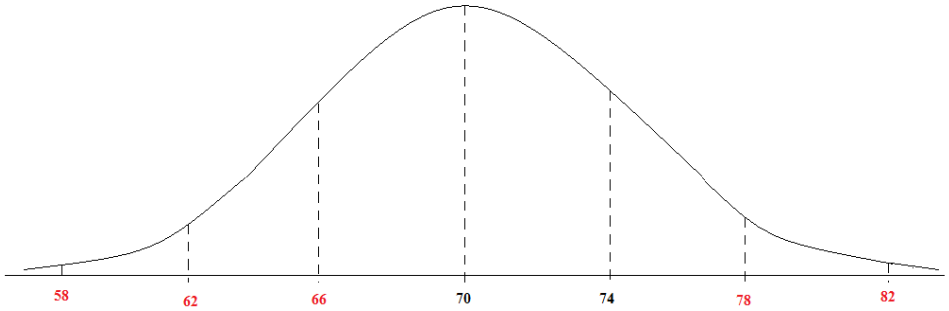


8: Use this graph to help you answer the following questions.



a. If about 68% of the observations are within one standard deviation of the mean, then about __32__ % are **not** within one standard deviation of the mean. Of these, about __16__ % are larger than $\bar{x} + s$, and about __16__ % are smaller than $\bar{x} - s$.
b. Because about 95% of the observations lie between $\bar{x} - 2s$ and $\bar{x} + 2s$, the largest __2.5__ % of the observations are larger than $\bar{x} + 2s$.
c. About __97.5__ % of the observations are *smaller* than $\bar{x} + 2s$. **Hint**: These are the one that are *not* larger than $\bar{x} + 2s$. Use your answer to the previous question. (100% minus 2.5%)

9: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use this information to finish labeling the following graph with numerical values for the mean ±1, 2, and 3 standard deviations. Then use the empirical rule and the symmetry to fill in the blanks.



a. Approximately 68% are between __66__ inches and __74__ inches tall. Therefore, about 32% are either taller than __74__ inches or shorter than __66__ inches, with about 16% in each of these groups.
b. Approximately 95% are between __62__ inches and __78__ inches tall. Therefore, about 5% are either shorter than __62__ inches or taller than __78__ inches, with about 2.5% in each of these groups.
c. Almost everyone is between __58__ inches and __82__ inches tall.
d. Approximately what percent are in each indicated height range: (for each we show you one way to calculate the answer; there are other ways)
   i. __68__ % between 66 and 74 inches
   ii. __34__ % between 66 and 70 inches (1/2 of 68%)

iii. 34 % between 70 and 74 inches (1/2 of 68%)
iv. 50 % less than 70 inches (by the symmetry)
v. 84 % less than 74 inches (50% less than 70, plus 34% 70 to 74)
vi. 16 % greater than 74 inches (100% minus 84%)
vii. 2.5 % greater than 78 inches (5% either below 62 or above 78, so 2.5% above 78)
viii. 97.5 % less than 78 inches (100% minus 2.5%)
ix. 13.5 % between 74 and 78 (97.5% less than 78, minus the 84% less than 74)
x. 16 % less than 66 inches (32% either below 66 or above 74, so 16% below 66))
xi. 2.5 % less than 62 inches (5% either below 62 or above 78, so 2.5% below 62)
xii. 81.5 % between 62 and 74 inches (84% less than 74, minus the 2.5% less than 62)

e. Anyone taller than 78 inches is in the tallest 2.5 %.

f. Anyone shorter than 66 inches is in the shortest 16%.

10: Refer to Exercise 9, and its solution, to calculate these probabilities.

These answers come directly from the solution to Exercise 9 part (d).

a. P(66 to 74) = 0.68
b. P(66 to 70) = 0.34
c. P(70 to 74) = 0.34
d. P(less than 70) = 0.50
e. P(less than 74) = 0.84
f. P(greater than 74) = 0.16
g. P(greater than 78) = 0.025
h. P(less than 78) = 0.975
i. P(74 to 78) = 0.135
j. P(less than 66) = 0.16
k. P(less than 62) = 0.025
l. P(62 to 74) = 0.815

11: Adult female heights for a different group of women have a mean of 65 inches and a standard deviation of 3.5 inches.

a. Cynthia is 69.5 inches tall. Calculate her *z*-score.

$$z = \frac{69.5-65}{3.5} = 1.29$$

b. Betty is 59 inches tall. Calculate her *z*-score.

$$z = \frac{59-65}{3.5} = -1.71$$

c. Mary's *z*-score is 1.96. How tall is she?

$$1.96 = \frac{x-65}{3.5}$$

$$x - 65 = 1.96(3.5)$$

$$x = 65 + 1.96(3.5) = 71.86 \text{ or approximately 72 inches}$$

d. Connie's *z*-score is –2.56. How tall is she?

$$-2.56 = \frac{x-65}{3.5}$$

$$x - 65 = -2.56(3.5)$$

$$x = 65 - 2.56(3.5) = 56.04 \text{ or approximately 56 inches}$$

(Notice the general pattern *x* = mean + *z* times standard deviation.)

12: Adult male heights are mound-shaped, with a mean of 70 inches and a standard deviation of 4 inches. Use the empirical rule to fill in the blanks.

a. Approximately 95% are between 62 inches and 78 inches tall. Sketch a picture.
See the graph in Exercise 9.

b. Anyone taller than 78 inches is in the tallest __2.5__ %.
c. Anyone shorter than __62__ inches is in the shortest 2.5%.
d. If by "unusual" we mean unusually far away from the average of 70 inches, then both short people and tall people qualify as unusual. The 5% of most unusual people are either shorter than __62__ inches or taller than __78__ inches.
e. If Sam is 79 inches tall, he is unusually tall because he falls in the top __2.5__ %. If Joe is 61 inches tall, he is unusually short because he falls in the bottom __2.5__ %. Both these people are unusual; they are in the rarest __5__ % of all adult male heights.
f. Sam is 79 inches tall and Bill is 81 inches tall. Both fall in the top 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Sam's or Bill's? Bill's
g. Joe is 61 inches tall and Ted is 58.5 inches tall. Both fall in the bottom 2.5% of heights using the empirical rule, but whose height would you say is more unusual, Joe's or Ted's? Ted's

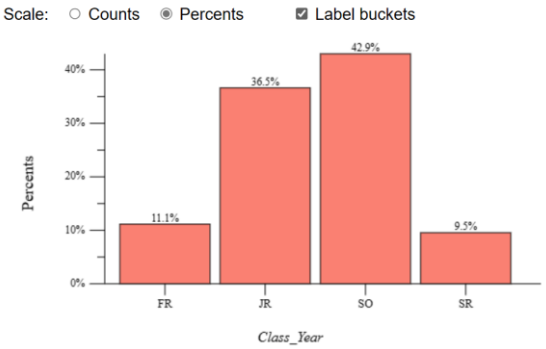
13: Use the calculator to do the following:

a. Find the min, max, and range for the height variable.

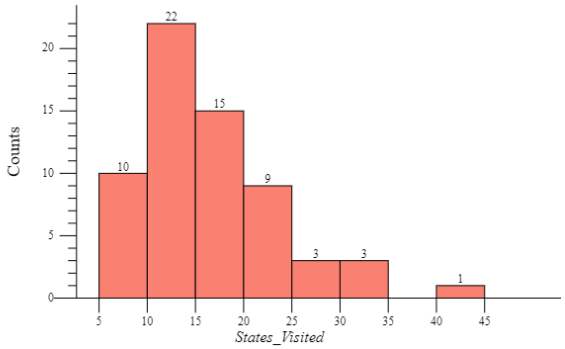
Statistics for the *Height(in)* variable

Min	56
Max	77
Range	21

b. Create a bar chart for the *Class_Year* variable.



c. Create a histogram of the states visited, including the count for each bucket. Comment on the shape of the distribution.
Here is the graph, using the defaults for the start and lengths for the buckets. The graph is skewed right, with what seems to be an outlier on the right.



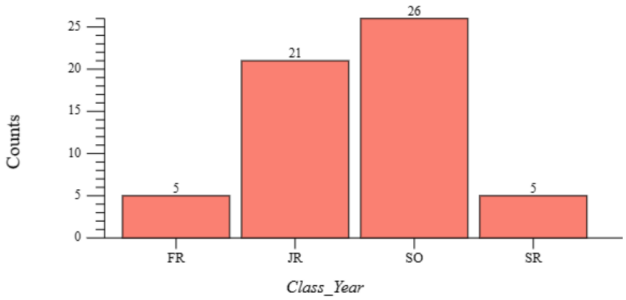
14: Use the calculator to do the following:

a. Find the min, max, and range for the height variable, for the upperclassmen (juniors/seniors)

Statistics for the *Height(in)* variable
Restricted to records where *Class_Year* is one of:
JR, SR

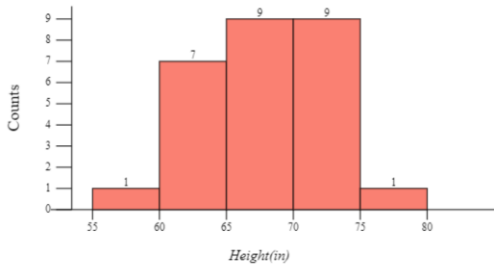
Min	61
Max	76
Range	15

b. Create a bar chart for the *Class_Year* variable, just for those who do not smoke.

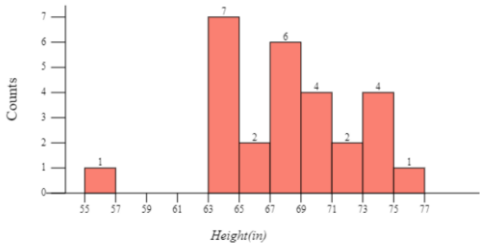


Restricted to records where *Smoke* is equal to:
N

c. Create a histogram of the heights for the sophomores in the file.
Here are two versions of the histogram. In the first graph we accepted the defaults for the bucket start and width; the second starts the buckets at 55 with a width of 2.

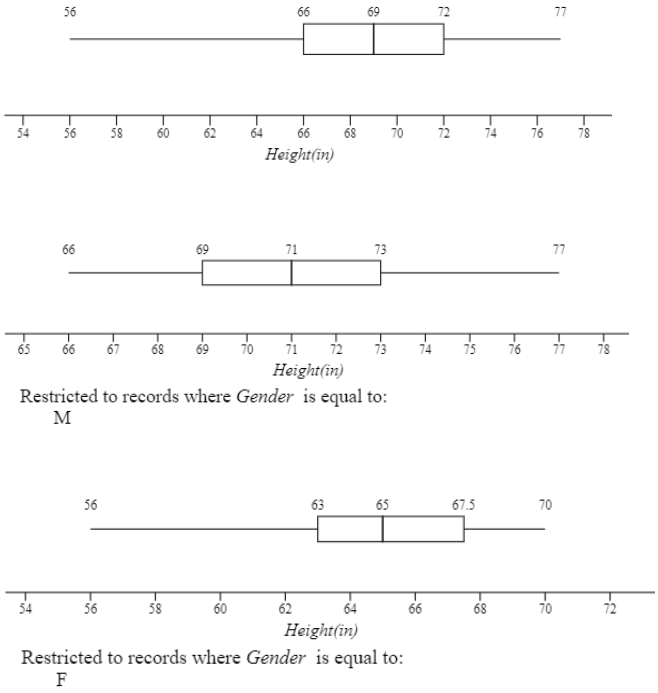


Restricted to records where *Class_Year* is equal to:
SO



Restricted to records where *Class_Year* is equal to:
SO

15: Create labeled boxplots for the height variable, for each of the following: **a.** The entire file; **b.** Just the males; **c.** Just the females



Comment: At a quick glance, it would appear that the women are taller than the men, since the box for the women is farther to the right than for the men. Upon closer inspection, we see that this is only because the scales are different. If we wanted to compare the two groups, we should graph the boxplots on the same scale – this will be discussed in some detail in the next lesson.

16: Use spreadsheet software and/or a text file editor to create a comma-delimited file for the survey data shown in Lesson 1 (page 2).

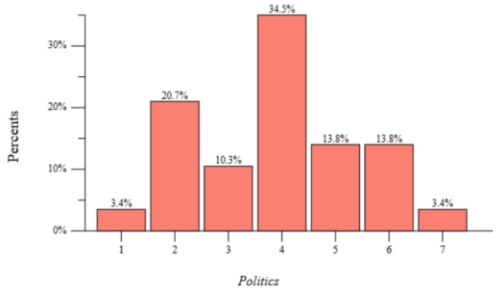
Here is a link to the file created by the author.

https://webspace.ship.edu/jwcraw/stats/Lessons/firstDayLesson1Data.txt

NOTE. There is some missing data, where the questions were not answered. When you load the file the calculator will inform you of this fact; any time an operation would involve that missing data, that person's data will not be included in the final results.

Then do the following for that data file:

a. Create a bar chart for the *Politics* variable, showing percentages for each category.

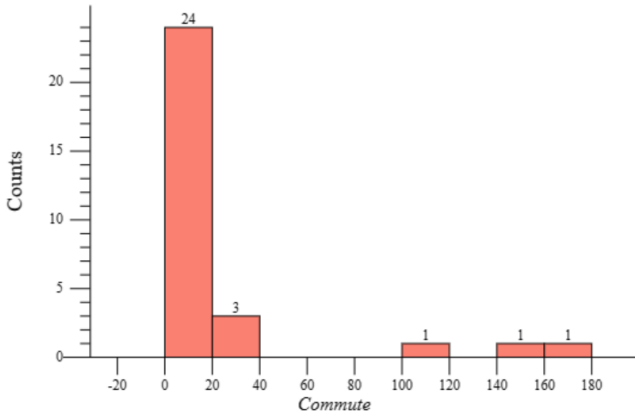


b. For the *Commute* variable, calculate the mean, standard deviation, median, and IQR, and create a histogram labeled with counts.

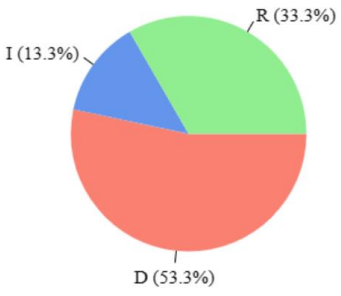
Comment. Notice the huge difference between the mean and the median (and also the standard deviation and the IQR). The histogram explains this difference – the data is very heavily skewed right, with outlier to the right as well.

Statistics for the *Commute* variable

Mean	17.1233
Std. dev.	42.8829
Median	0.75
IQR	1.2



c. Create a pie chart for the *Political party* variable, for the males in the file.



Political Party

Restricted to records where *Gender* is equal to:
Male