Chapter 2: Graphical Summaries

Consider the following data
\[ x: 78, 24, 57, 39, 28, 30, 29, 18, 102, 34, 52, 54, 57, 82, 90, 94, 38, 59, 27, 68, 61, 39, 81, 43, 90, 40, 39, 33, 42, 15, 88, 94, 50, 66, 75, 79, 83, 34, 31, 36, 38, 90, 30, 49, 41, 42, 5, 2, 74, 82, 19, 11, 10, 7, 45, 80, 20, 34, 58, 55, 29, 30, 40, 22, 29, 50, 55, 51, 71, 77, 82, 94, 99, 107, 135, 150, 139, 45, 58, 53, 93, 98, 97, 20, 32, 75, 82, 49, 48, 29, 0, 23, 78, 58, 93, 47, 45, 28, 34, 59, 92, 93, 74, 58, 91, 90, 83, 84, 86, 80, 27 \]

What would your answer be if you were asked to give a typical value for this data set. Could you quickly give one? If you were asked to estimate where the middle of the data is could you quickly give a value? How about an estimate of how spread out the data is? Could you estimate how far each observation is from the average? The answer to all of these questions for most people is that they could not. Even people who regularly work with numbers have difficulty summarizing large data sets such as this one. For that reason, we try to summarize data. In the next two chapters we’ll consider two ways to do this. In this chapter, we’ll describe some methods for making graphical summaries. That is, we’ll make pictures from the data and then use those pictures to answer some of the questions above. In chapter 2, we’ll use numbers to summarize the data.

2.1 Frequency Table

Almost all of the graphical methods that we use will depend upon a frequency table.

*Definition:* A frequency table is a summary of the data that tabulates the number of observations in each of a number of groups or bins.

There are usually four columns for a frequency table. The first is the bin label. This describes what observations fall in that bin or category. For categorical data the bin label is that particular category. For numerical data the bin label is the range of values that fall in that bin. For example, one possible bin label for the data at the beginning of the chapter is 79.5 to 89.5. Then that bin would include any values that fall between those two numbers. The second column that is often included is a column for tally. This is optional but can be very helpful when you are doing a frequency table by hand. You place a single slash for each time that an observation fall in that bin. The next column is the frequency column. This is the number of observations that fall into that bin. The most efficient way to get the frequency is simply to count the number of tallies for that bin. The last column is the relative frequency. This is the frequency divided by the total number of observations, \( n \). This gives the percentage of observations that fell in that bin. This is useful for comparing the relative proportions for a large data set.
Example: Frequency table for favorite cereal of Statistics 215, Lab Section 14 students:
In this case $n = 19$.

<table>
<thead>
<tr>
<th>Favorite Cereal</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruity Pebbles</td>
<td>llll</td>
<td>4</td>
<td>4/19=0.21</td>
</tr>
<tr>
<td>Count Chocula</td>
<td>llll</td>
<td>4</td>
<td>4/19=0.21</td>
</tr>
<tr>
<td>Raisin Bran</td>
<td>lll</td>
<td>3</td>
<td>3/19=0.16</td>
</tr>
<tr>
<td>Smart Start</td>
<td>ll</td>
<td>2</td>
<td>2/19=0.11</td>
</tr>
<tr>
<td>Cheerios</td>
<td>lllll</td>
<td>5</td>
<td>5/19=0.26</td>
</tr>
<tr>
<td>Blueberry Morning</td>
<td>l</td>
<td>1</td>
<td>1/19=0.05</td>
</tr>
</tbody>
</table>

The previous example is for categorical data. The next example is a frequency table for numerical data.

Example: The following data is car length in inches for cars found in lot 56 at WVU. c: 78, 124, 98, 72, 88, 84, 80, 90, 91, 74, 83, 92, 105, 84, 84, 93, 97
Note that $n = 17$

For numerical data, the bin labels must be unambiguous. Consequently they need to go one decimal place beyond the data. It is also important that the bins be of equal width.

<table>
<thead>
<tr>
<th>Length</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>69.5 to 79.4 lll</td>
<td>3</td>
<td>3/17=0.18</td>
<td></td>
</tr>
<tr>
<td>79.5 to 89.4 llllll</td>
<td>6</td>
<td>6/17=0.35</td>
<td></td>
</tr>
<tr>
<td>89.5 to 99.4 lllll</td>
<td>6</td>
<td>6/17=0.35</td>
<td></td>
</tr>
<tr>
<td>99.5 to 109.4 l</td>
<td>1</td>
<td>1/17=0.06</td>
<td></td>
</tr>
<tr>
<td>109.4 to 119.4 l</td>
<td>1</td>
<td>0/17=0.00</td>
<td></td>
</tr>
<tr>
<td>119.5 to 129.4 l</td>
<td>1</td>
<td>1/17=0.06</td>
<td></td>
</tr>
</tbody>
</table>

TIP: It is always a good idea to add all the relative frequencies up. They should add to about 1.00. With rounding they can often add to slightly lower or higher, say 0.98 or 1.01. If the sum is far off say 1.15, they you need to rework your frequency table.

There are several statements we can make from the frequency table. For the first example dealing with cereal we can say which cereals were the least and which were the most popular. Cheerios was the most popular, while Blueberry Morning was the least popular. However, Blueberry Morning got more votes than say Grape-Nuts which was not chosen by anyone in that section of Statistics 215. For the numerical data we can observe that the car lengths from about 80 inches to 100 inches were the most common car lengths.

2.2 Graphical Summaries for Categorical Data

Pie Chart
Both of the methods we will discuss for summarizing categorical data stem directly from the frequency table. The first is the pie chart. For the example here we’ll use the above data from cereal favorites of Statistics 215 students. For a pie chart, the relative frequency determines how much of the pie is given to each category. Remember that for categorical data the bins of the frequency table are the categories. Below is an example of a pie chart for that data.
Example
As a reminder here is the Frequency Table for the Favorite Cereal Data.

<table>
<thead>
<tr>
<th>Favorite Cereal</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruity Peebles</td>
<td>llll</td>
<td>4</td>
<td>4/19=0.21</td>
</tr>
<tr>
<td>Count Chocula</td>
<td>llll</td>
<td>4</td>
<td>4/19=0.21</td>
</tr>
<tr>
<td>Raisin Bran</td>
<td>lll</td>
<td>3</td>
<td>3/19=0.16</td>
</tr>
<tr>
<td>Smart Start</td>
<td>ll</td>
<td>2</td>
<td>2/19=0.11</td>
</tr>
<tr>
<td>Cheerios</td>
<td>lllll</td>
<td>5</td>
<td>5/19=0.26</td>
</tr>
<tr>
<td>Blueberry Morning</td>
<td>l</td>
<td>1</td>
<td>1/19=0.05</td>
</tr>
</tbody>
</table>

Pie chart of Cereal Favorites for Statistics 215, Section 14 students.

A pie chart should always be in the proper proportions. That is the proportional area of the circle or pie corresponding to each category should be the same as it’s relative frequency. So in this example, Roughly 21% of the pie’s area should go to the slice labeled Count Chocula, since Count Chocula had 21% of the observations in our data set. Additionally the areas should all be slices of a pie. This means that the pie should be divided from the center so that the edges of a slice meet at the center of the circle.
Bar chart
The bar chart like the pie chart is based on the frequency table. For the bar chart each category has its own bar. For a particular category, the height of the bar represents either the frequency or the relative frequency.

Example
Running magazine was considering starting a Senior Running magazine to target runners over 50. They asked people over 50 their favorite sporting activity. The results are in the frequency table below. In this table the tally has been left out.

<table>
<thead>
<tr>
<th>Sport</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golf</td>
<td>138</td>
<td>138/458 = 0.30</td>
</tr>
<tr>
<td>Handball</td>
<td>17</td>
<td>17/458 = 0.04</td>
</tr>
<tr>
<td>Running</td>
<td>65</td>
<td>65/458 = 0.14</td>
</tr>
<tr>
<td>Tennis</td>
<td>87</td>
<td>87/458 = 0.19</td>
</tr>
<tr>
<td>Walking</td>
<td>109</td>
<td>109/458 = 0.24</td>
</tr>
<tr>
<td>Other</td>
<td>42</td>
<td>42/458 = 0.09</td>
</tr>
</tbody>
</table>

Bar Chart of Favorite Activities of People Over 50

To differentiate the bar chart from another graphical summary that we will see later (the histogram), it is best to keep some space between the bars.
2.3 Graphical Summaries for Numerical Data

We move our attention now to numerical summaries. With this type of data, we no longer have categories. Instead, we must break the scale of measurements – feet, meters, pounds – into bins that will substitute for the categories we had in the previous section. In this chapter we’ll consider three types of numerical summaries: stem-and-leaf plot, histogram and cumulative frequency polygon. Each has distinct advantages and disadvantages.

Stem-and-Leaf Plot

Consider the following data. Y: 739, 279, 375, 180, 274, 350, 389, 240.5, 298, 332, 245. For this data the stem-and-leaf plot is

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4 5 7 8</td>
</tr>
<tr>
<td>3</td>
<td>0 3 5 9</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

where 3|3 represents 330.

Before we discuss how we constructed this plot, look back at the plot. First notice that the plot is split into Stems and Leaves. The stems represent hundreds and the leaves represent the tens place. We are told this by the legend at the bottom, which says 3|3 represents 330. There is one leaf for every observation in the data set though those observations have been rounded.

To construct a stem-and-leaf plot we need to follow the following steps.
1. Put the data in order, either ascending or descending.
2. Choose the number of stems that you want to use. It is best to choose a particular digit place for the stems, say tenths or hundreds. You want to choose the number of stems so that you have between 5 and 20 stems.
3. Write the stems vertically from the smallest down to the largest (Do not skip any stems).
4. Round all the observations to one decimal after the place you chose in Step 2.
5. Put the leaves on the stems in order, increasing from left to right. Be sure to have the same space between leaves on each line.
6. Label the plot. Stem|Leaf represents ….
Example
For the above data set, these steps are
1. 180, 240.5, 245, 274, 279, 298, 332, 350, 389, 739
2. Choose the hundreds place to be the stems
3.
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
4. Round all the numbers to the tens place
   180, 240, 250, 270, 280, 300, 330, 350, 390, 740
5. Place the leaves on the appropriate stem
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4 5 7 8</td>
</tr>
<tr>
<td>3</td>
<td>0 3 5 9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>
6. Add the label.
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4 5 7 8</td>
</tr>
<tr>
<td>3</td>
<td>0 3 5 9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>
| 3 | 3 represents 330.

The stem and leaf plot is a nice guide for summarizing numerical data. It is easy to pick out extreme observations, i.e. those that are too small or too large. The data is all in the plot which will not be true of the next graphical summary we consider. We’ll be able to see patterns in the data quite quickly. The drawback to stem-and-leaf plots is that they do take quite a long time to construct, if the data set is quite large (if n is large).
Histogram

The second method of graphically summarizing data is the histogram. The histogram is the most commonly used graphical summary for numerical data. It is effectively a bar chart for numerical data. The difference between the bar chart and the histogram is that the X-axis for the histogram represents some scale of measurement. Consequently, since numbers can be anywhere along the number line, then the bars will touch. This also carries the connotation that there is continuity between the bins.

Example
Let P be the length of phone conversations from Hampsfield Air Force Base to the United States in minutes.
p: 0.5, 0.5, 0.6, 0.7, 1.2, 1.3, 7.4, 8.2, 8.9, 12.3, 14.3, 19.5, 20.3, 22.4, 25.5, 28.4, 32.0, 32.1, 33.6, 47.4, 49.5, 53.4, 54.8, 57.9, 58.6, 58.7

<table>
<thead>
<tr>
<th>Length in minutes</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 to 9.95</td>
<td>I-I-I-I-I-I-I-I-I-I-I-I</td>
<td>9</td>
<td>9/26=0.35</td>
</tr>
<tr>
<td>10.00 to 19.95</td>
<td>I-I-I</td>
<td>3</td>
<td>3/26=0.12</td>
</tr>
<tr>
<td>20.00 to 29.95</td>
<td>I-I-I-I</td>
<td>4</td>
<td>4/26=0.15</td>
</tr>
<tr>
<td>30.00 to 39.95</td>
<td>I-I-I</td>
<td>3</td>
<td>3/26=0.12</td>
</tr>
<tr>
<td>40.00 to 49.95</td>
<td>I-I</td>
<td>2</td>
<td>2/26=0.08</td>
</tr>
<tr>
<td>50.00 to 59.95</td>
<td>I-I-I-I-I-I-I-I-I-I-I-I-I-I</td>
<td>5</td>
<td>5/26=0.19</td>
</tr>
</tbody>
</table>

From this frequency table we can make the following histogram.

**Histogram of length of phone calls in minutes from Hampsfield Air Force Base to the United States**

![Histogram graph](image)
From the histogram on the previous page, we can get some summaries of the data quite quickly. The key to interpreting a histogram as with most of the other graphical methods is that frequency (or relative frequency) is equated to height of the bars or more generally to area of the bars. In the case of the histogram, the higher the bars the more observations are in that interval. Just from looking at that histogram we can estimate the shortest calls and the longest calls. We can also notice that more calls fall into the first bin than into any other. Thus more calls are less than ten minutes than for any other ten-minute interval. We can also estimate the point in the data where half of the observations fall above and half of the observations fall below. In the histogram of phone call lengths, this is approximately 18 to 25. To judge this you want to choose a point on the X-axis where the area of the bars to the right of that point is about half of the total area. Likewise then the area of the bars to the left of that point is about half of the total area.

The histogram has several advantages. First, as was illustrated in the previous paragraph, the histogram gives a lot of information quickly. Instead of trying to gauge and process the 26 numbers from the phone call length data set, we can look at a picture. This picture was easier to digest. Histograms are also quite common in their usage, so that most people know how to deal with them. In the example above the data set was relatively small, \( n=26 \). The histogram, however, can be a very good way of summarizing a very large number of observations. In addition to these advantages, the histogram can be found in a large number of software packages, e.g. Microsoft Excel.

The histogram does have two disadvantages. First, it may not be good at detecting gaps in the data or observations that are extreme (large or small) relative to the rest of the data. Look back at the phone call length data, there is a gap from the 33.6 to 47.4. However, there is no gap in the histogram to reflect this gap of over 10 minutes. Also, the shape of the histogram would change if we changed the particular number of bins that were used. Thus, if two people chose different bins, they could get somewhat different histograms for the same dataset.

**TIP:** For a histogram to accurately convey data to the human eye, it is important to have each bin of the frequency table be of equal width. That is, the bins must all have the same range, the high value minus the low value. Additionally, then the widths of the bars must all have the same widths. The reason for these restrictions is that the human eye is not good at perceiving areas if the widths are different. Since area is equivalent to frequency we then don’t interpret the histogram correctly.

**Cumulative Relative Frequency Polygon (CRF)**

There are times when it is important to know how much of the data is above or below a certain value in the data. In order to accomplish this, we will add another column to the frequency table. This column will be labeled cumulative frequency. The cumulative relative frequency is the percentage of observations that are equal to or below are particular value. For the frequency table, it is the relative frequency in that bin plus the sum of the relative frequencies of the bins with values that are less than the values in that bin.
Example
Consider the phonecall length data from above.

<table>
<thead>
<tr>
<th>Length in minutes</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 to 9.95</td>
<td>lllll lllll</td>
<td>9</td>
<td>9/26=0.35</td>
<td>9/26=0.35</td>
</tr>
<tr>
<td>10.00 to 19.95</td>
<td>III</td>
<td>3</td>
<td>3/26=0.12</td>
<td>12/26=0.46</td>
</tr>
<tr>
<td>20.00 to 29.95</td>
<td>llll</td>
<td>4</td>
<td>4/26=0.15</td>
<td>16/26=0.62</td>
</tr>
<tr>
<td>30.00 to 39.95</td>
<td>III</td>
<td>3</td>
<td>3/26=0.12</td>
<td>19/26=0.73</td>
</tr>
<tr>
<td>40.00 to 49.95</td>
<td>II</td>
<td>2</td>
<td>2/26=0.08</td>
<td>21/26=0.81</td>
</tr>
<tr>
<td>50.00 to 59.95</td>
<td>llllll</td>
<td>5</td>
<td>5/26=0.19</td>
<td>26/26=1.00</td>
</tr>
</tbody>
</table>

To get the cumulative relative frequency for the first bin, just take the relative frequency for that bin. For the rest of the bins, take the relative frequency for that bin and add it to the cumulative relative frequency for the bin above it.

In the example above, the cumulative relative frequency (CRF) for the first bin (0.00 to 9.95) is 9/26 or 35%. The CRF for the second bin (10.00 to 19.95) is the CRF for the bin above it (9/26) plus the relative frequency for that bin (3/26) which is 9/26+3/26 = 12/26 = 0.46. Likewise, the CRF for the third bin is the CRF for the second bin plus the relative frequency which is 12/26 + 4/26 = 16/26. Similarly continue for the remainder of the table.

We can then turn this table into a graph to yield a cumulative relative frequency polygon. To achieve this we plot the high values for each bin (9.95, 19.95, etc.) against the CRF for each bin. We include the low value of the first bin to make the graph easier to interpret. Since no values fall below it, the CRF for that value is 0.00. So for the above table we would plot the following values. The X-values are the high end of the bins, the Y-values are the cumulative relative frequencies.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9.95</td>
<td>0.35</td>
</tr>
<tr>
<td>19.95</td>
<td>0.46</td>
</tr>
<tr>
<td>29.95</td>
<td>0.62</td>
</tr>
<tr>
<td>39.95</td>
<td>0.73</td>
</tr>
<tr>
<td>49.95</td>
<td>0.81</td>
</tr>
<tr>
<td>59.95</td>
<td>1.00</td>
</tr>
</tbody>
</table>

After plotting these points, connect the points with line segments.
Cumulative relative frequency of length of phone calls in minutes from Hampsfield Air Force Base to the United States

We can use this to identify certain aspects of the dataset. For example, if we start from the point on the Y-axis where the CRF is 0.50. If we go across to the right until we hit the line, then from that drop down to the X-axis. That point about 23, say, is the middle of the data. There should be about half or 50% of the data below 23 and half of the distribution above 23.

2.4 Peaks and Skews

For summarizing a data set it is often important to talk about the shape of the dataset. A dataset is often referred to as a histogram which is the most common form of graphically representing a data sets. The methods described below will be used for describing either histograms or stem-and-leaf plots.

Peak
One way to describe a histogram is to describe the number of humps or peaks that the data sets has. If a histogram has one peak, we refer to it as unimodal. If it has two peaks, we refer to it as bimodal and if it has more than two peaks, we refer to it as multimodal. Finally, if it seems to have no peaks, then we call that histogram uniform. It is important to remember when considering the peaks of a data set, that we are looking not for the number of high bars but rather the number of sections of the data that our distinct.
Examples

This histogram is unimodal, since there is only one peak.

![Unimodal Histogram](image1)

This histogram (below) is bimodal since it has two peaks.

![Bimodal Histogram](image2)

This histogram (below) is unimodal, though there are several high bars, there is only one distinct section of the data where those bars exist.

![Unimodal Histogram with Multiple Peaks](image3)
The histogram below is multimodal since there are several distinct humps.

The next histogram is roughly uniform since there is no distinct peaked region.
Skew

Skewness relates to the shape of a dataset. Generally, it is the tendency for the data to fall off to one side or the other. It can also be seen as relating some numerical quantities together as we’ll see in Chapter 4. This is best seen graphically.

Examples:
Below is an example of data that is skewed right. The data falls off from the peak and has a longer run toward the right, so the data is right skewed.

![Histogram of skewed right data]

Below is an example of data that is skewed left. The section of the distribution that is longer as it moves away from the peak is referred to as the tail of this distribution. In this case the tail is to the left, so the distribution is skewed left.

![Histogram of skewed left data]
The distribution below is symmetric since it is about even on either side of the middle.

It is possible to have distribution that are not unimodal be skewed. Below we have a distribution that is bimodal and skewed right.

2.5 Summary

In this chapter we began a discussion of ways to summarize a data set. First, we introduced some graphical summaries. We started with the frequency table that is the basis for a wide variety of graphical summaries. Next we looked at pie charts and bar charts. The two summaries are useful for categorical data. For numeric data we have the stem-and-leaf plots as well as the histogram for encapsulating the data. For both of these summaries, we looked at ways to describe them. Specifically we looked at peakedness and skew for these graphs.