Chapter 2

The CRD with a Numerical Response: Continued

This chapter continues the theme of Chapter 1. I begin with another example of a student project.

2.1 Kymn the Rower

Kymn was a member of the women’s varsity crew at the University of Wisconsin-Madison. When she could not practice on a lake, she would work out on a rowing simulation device called an ergometer. One does not simply sit down at an ergometer and begin to row. It is necessary to choose the setting for the machine. There are four possible settings, obtained by combining two dichotomies:

- One can opt for the small gear setting or the large gear setting.
- One can choose to have the vent open or closed.

Kymn decided that she was not interested in two of these settings: the large gear with the vent closed would be too easy and the small gear with the vent open would too difficult for a useful workout. As a result, Kymn wanted to compare the following two settings:

- **Treatment 1**: The small gear with the vent closed, and
- **Treatment 2**: The large gear with the vent open.

For her response, Kymn chose the time, measured to the nearest second, she required to row the equivalent of 2000 meters.

In the above, I have implicitly defined Kymn’s trial as sitting on the erg and rowing the equivalent of 2000 meters. Kymn decided to perform a total of 10 trials in her study.

Kymn’s data are in Table 2.1 with dot plots in Figure 2.1. Look at these data for a few minutes. What do you see? Below are some features that I will note.

1. Every response on treatment 2 is smaller than every response on treatment 1.
Table 2.1: Kymn’s times, in seconds, to row 2000 meters on an ergometer. Treatment 1 is the small gear with the vent closed; and treatment 2 is the large gear with the vent open.

<table>
<thead>
<tr>
<th>Trial:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment:</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Response:</td>
<td>485</td>
<td>493</td>
<td>489</td>
<td>492</td>
<td>483</td>
<td>488</td>
<td>490</td>
<td>479</td>
<td>486</td>
<td>493</td>
</tr>
</tbody>
</table>

Figure 2.1: The dot plots for Kymn’s rowing study.

Treatment 1: Small Gear, Vent Closed:

Treatment 2: Large Gear, Vent Open:

2. The variation in treatment 2 is larger than the variation in treatment 1. Having noted this fact, in both treatments there is very little within-treatment variation. It is impressive, yet perhaps unsurprising for a well-conditioned athlete, that in response times of slightly more than 8 minutes, there is so little variation in trial-to-trial performance.

If one looks at the dot plots, and remembers the center of gravity interpretation of the mean, one can see that the mean on treatment 1 is a bit larger than 491 seconds and that the mean on treatment 2 is a bit smaller than 485 seconds; these visual conclusions are supported by computation. In particular, for future reference note that the means, medians and standard deviations of these data are:

\[ \bar{x} = 491.4, \bar{y} = 492, s_1 = 1.817, \tilde{y} = 484.2, \tilde{y} = 485 \text{ and } s_2 = 3.420. \]

2.2 Sara’s Golf Study; Histograms

Sara performed a balanced CRD with 80 trials. Her response was the distance—in yards—that she hit a golf ball at a driving range. (She hit the ball into a net which displayed how far the ball would have traveled in real life. I have no idea how accurate these devices are.) Sara had two treatments: hitting the ball with a 3-Wood (treatment 1) and hitting the ball with a 3-Iron (treatment 2). If you don’t know much about golf, don’t worry; all that matters is that Sara wanted to compare two clubs with particular interest in learning which would lead to a larger response.
Table 2.2: The distance Sara hit a golf ball, in yards, sorted by treatment.

<table>
<thead>
<tr>
<th></th>
<th>3-Wood</th>
<th></th>
<th>3-Iron</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>32</td>
<td>56</td>
<td>58</td>
<td>77</td>
</tr>
<tr>
<td>101</td>
<td>101</td>
<td>107</td>
<td>107</td>
<td>108</td>
</tr>
<tr>
<td>113</td>
<td>114</td>
<td>116</td>
<td>118</td>
<td>122</td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>131</td>
<td>131</td>
<td>137</td>
</tr>
<tr>
<td>27</td>
<td>52</td>
<td>57</td>
<td>58</td>
<td>59</td>
</tr>
<tr>
<td>84</td>
<td>88</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>100</td>
<td>101</td>
<td>107</td>
<td>107</td>
<td>108</td>
</tr>
<tr>
<td>118</td>
<td>127</td>
<td>132</td>
<td>132</td>
<td>136</td>
</tr>
</tbody>
</table>

Figure 2.2: The dot plots for Sara’s golf study.

3-Wood:

Sara’s data, sorted by treatment, are presented in Table 2.2. Even a cursory examination of this table reveals that, within each treatment, there is a huge amount of variation in Sara’s responses. Dot plots of Sara’s data are presented in Figure 2.2.

I don’t like these dot plots very much, but let me begin by mentioning their good features. As with all dot plots, each plot is a valid presentation of its observations. If you want to see the exact values of all of the observations and how they relate spatially, the dot plot is great. In addition, a dot plot is good at revealing outliers: we can see the three very small response values with the 3-Wood and the one very small value with the 3-Iron. Now I will discuss, briefly, what I don’t like about these dot plots.

The 3-Wood data range from a minimum of 22 yards to a maximum of 147 yards. This distance, 125 yards, towers over the number of observations, 40. As a result, there must be, and are, a large number of gaps in our picture and usually (there are weird exceptions) with so little data spread out so far, the peaks are very short and, hence, likely have no scientific meaning. There is another way to view the above comments: the dot plot is very bumpy; i.e., it is not very smooth. As I will
Table 2.3: Frequency tables of the distances Sara hit a golf ball, by treatment.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Width (w)</th>
<th>Freq. (f)</th>
<th>Rel. Freq. ((r_f = f/n_1))</th>
<th>Density ((d = r_f/w))</th>
<th>Freq. (f)</th>
<th>Rel. Freq. ((r_f = f/n_2))</th>
<th>Density ((d = r_f/w))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–25</td>
<td>25</td>
<td>1</td>
<td>0.025</td>
<td>0.001</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>25–50</td>
<td>25</td>
<td>2</td>
<td>0.050</td>
<td>0.002</td>
<td>1</td>
<td>0.025</td>
<td>0.001</td>
</tr>
<tr>
<td>50–75</td>
<td>25</td>
<td>2</td>
<td>0.050</td>
<td>0.002</td>
<td>8</td>
<td>0.200</td>
<td>0.008</td>
</tr>
<tr>
<td>75–100</td>
<td>25</td>
<td>4</td>
<td>0.100</td>
<td>0.004</td>
<td>11</td>
<td>0.275</td>
<td>0.011</td>
</tr>
<tr>
<td>100–125</td>
<td>25</td>
<td>18</td>
<td>0.450</td>
<td>0.018</td>
<td>11</td>
<td>0.275</td>
<td>0.011</td>
</tr>
<tr>
<td>125–150</td>
<td>25</td>
<td>13</td>
<td>0.325</td>
<td>0.013</td>
<td>9</td>
<td>0.225</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>—</td>
<td>40</td>
<td>1.000</td>
<td>—</td>
<td>40</td>
<td>1.000</td>
<td>—</td>
</tr>
</tbody>
</table>

discuss later in the subsection on *kernel densities*, smoothness is very important to scientists.

Here is what I mean by bumpy. Imagine the number line is a road and the dots are bumps in the road. Driving a car (or if you prefer a greener example, riding a bike) along the road will result in a flat road (the gaps) interrupted by numerous bumps.

Finally, it is very difficult to see a shape in either of these dot plots. This is disturbing because the quest for a shape is one of the main reasons that scientists draw pictures of data.

Admittedly, all of the dot plots we have seen in these notes have been bumpy. Most of our dot plots have not had a recognizable shape, but that is to be expected with a small amount of data, as we had in Dawn’s and Kymn’s studies in our exposition, as well as Brian’s and Reggie’s studies in the homework to Chapter 1. Arguably, policeman Kenny’s dot plots (see Figure 1.2 on page 21 in the Chapter 1 Practice Problems), based on a large number of observations and a small range of values, did reveal shapes. Below we will introduce histograms, which are smoother (statisticians often prefer this more positive term to less bumpy) than dot plots and usually reveal shape better. Finally, I will introduce you briefly to kernel density estimates that are, in the sense we will learn, better than histograms both on smoothness and revealing shapes.

In the excellent movie *Amadeus* (1984), a dramatization of the life of composer Wolfgang Amadeus Mozart (1756–1791), a jealous competitor derides one of Mozart’s works as having too many notes. In a similar spirit, one can criticize a dot plot for having too much detail. Our next picture, the histogram sacrifices some of the detail of a dot plot. The reward, one hopes, is a better, more useful, picture.

The first thing to note is that to refer to the histogram for a set of data is a bit misleading. The definite article—the—is inappropriate because many histograms can be drawn for any set of data, for two reasons. First, as we will see below, a histogram is dependent on our choice of class intervals, and there are always many possible choices for these. Second, for a given choice of class intervals, there are three possible histograms: the frequency histogram, the relative frequency histogram and the density histogram.

The first step in creating a histogram is to create a frequency table. Table 2.3 presents fre-
quency tables for both treatments for Sara’s data. Let me carefully explain these tables. The first column presents my choices for the class intervals. Because I am very interested in comparing the responses on the two treatments, I am using the same class intervals for both tables. This isn’t necessary, but I do think it’s a good idea.

In this course—exams and these notes, including homework—I will always give you the class intervals. (If you perform a project that requires a frequency table, then you will need to choose the class intervals.) Our class intervals will always follow the rules listed below. (Thus, if you need class intervals for your project, please follow these rules.) As you no doubt have already surmised, these rules do not rate high on the excitement-o-meter, but they are necessary. And there is a really annoying feature: There are two other versions of these rules—far inferior to the rules below—each of which appears in many textbooks of introductory Statistics. (Don’t let the adjective many dismay you; there are hundreds, if not thousands, of introductory Statistics texts and nearly all should be avoided. But that’s another topic.)

A valid collection of class intervals will satisfy the following five rules.

1. Each class interval has two endpoints: a lower bound(ary) and an upper bound(ary). As in Table 2.3, when one reads down the first column, the lower (and upper) bounds increase.

2. The smallest class interval boundary must be less than or equal to all observations in the data set.

3. The largest class interval boundary must be greater than or equal to all observations in the data set.

4. The upper bound of a class interval must equal exactly the lower bound of the next class interval.

5. Because adjacent class intervals have an endpoint in common, we need the following endpoint convention:

   When determining the frequencies of the class intervals (see below), each interval includes its left endpoint, but not its right endpoint.

   There is one exception to this endpoint convention: The last class interval includes both of its endpoints.

Let me say a few more words about our fourth rule. In Table 2.3, the first two class intervals are 0–25 and 25–50. There are two ways that the fourth rule could be violated; here are examples of each:

- If these intervals were changed to, say, 0–25 and 30–50, then there would be a gap between the intervals.

- If these intervals were changed to, say, 0–25 and 20–50, then these intervals would overlap.
We allow neither gaps nor overlap; either of these features would ruin our histograms.

We have spent a lot of effort on the class intervals! Let’s return to Table 2.3 and examine its remaining columns.

The second column presents the width \( (w) \) of each class interval. The width of a class interval is the distance between its endpoints. For example, the first class interval, 0–25, has \( w = 25 - 0 = 25 \), as printed. Note that in this table, all class intervals have the same width. There are reasons that a researcher may prefer to have variable-width class intervals (see Practice Problems below), but if one chooses to have variable-width class intervals, then one must use the density histogram because both of the other histograms are misleading (again, see Practice Problems below). Misleading is perhaps a bit strong. Statisticians agree that they are misleading, but, for all I know, you might be a person who is never misled by a picture.

The frequency counts \( (f) \) are pretty evident. For example, in the 3-Wood data set, four observations—77, 81, 93 and 99—fall in the interval 75–100; hence, the frequency of this interval is 4. An interval’s relative frequency \( (r_f) \) is obtained by dividing its frequency by the number of observations in the data set. Thus, for example, the relative frequency for the 75–100 class interval in the 3-Wood data is \( r_f = 4 / 40 = 0.100 \). Finally, an interval’s density \( (d) \) is obtained by dividing its relative frequency by its width.

Let me give you an example in which the endpoint convention in the fifth rule above comes into play. In Sara’s 3-Iron data set, the observation 100 is counted as a member of the interval 100–125, not the interval 75–100.

I will now use the frequency tables for Sara’s data to draw frequency histograms. These histograms are presented in Figure 2.3. First, I will discuss how to draw a frequency histogram by hand. Second, I will discuss the information revealed by Sara’s frequency histograms. Finally, I will discuss Sara’s relative frequency and density histograms.

**Drawing a Frequency Histogram.** We proceed as follows.

1. Draw a portion of the number line and locate the various class interval boundaries on it.

2. Above each class interval draw a rectangle whose height is equal to the frequency of the class interval.

**What do we learn by inspecting a frequency histogram?** Whenever we have histograms of data from two groups we can look to see if the groups differ substantially in centers and/or spreads. For Sara’s data, we see that she definitely hit the ball farther with the 3-Wood than with the 3-Iron. This conclusion is supported by computing means and medians for her data. I also include the standard deviations below.

\[
\bar{x} = 106.875, \bar{x} = 112.0, s_1 = 29.87, \bar{y} = 98.175, \bar{y} = 99.5 \text{ and } s_2 = 28.33.
\]

Statisticians and scientists are particularly interested in assessing the shape of a histogram, although this is a very inexact activity. Below are my comments on Sara’s two histograms:
1. 3-Wood Histogram: There is one tallest rectangle, above the class interval 100–125 yards. Thus, this is the most popular class interval for Sara’s 3-Wood responses. The rectangle(s) to the right [left] of this peak rectangle is called the right tail [left tail] of the histogram. The left tail is much longer than the right tail (100 yards versus 25), but the right tail is heavier (13 observations versus 9). Because of the longer left tail, we label this histogram skewed to the left.

2. 3-Iron Histogram: This histogram exhibits almost perfect left-to-right symmetry. Note the following facts:

   - The 3-Wood histogram is skewed to the left and its mean is smaller than its median: $106.875 < 112.0$.

   - The 3-Iron histogram is approximately symmetric and its mean and median are approximately equal: $98.175 \approx 99.5$.

These are two examples of the following famous Result that is not quite a Theorem.

Result 2.1 The following are usually true:

   - If the dot plot or histogram of a set of data is approximately symmetric, then its mean and median are approximately equal.

   - If the dot plot or histogram of a set of data is clearly skewed to the right, then its mean is larger than its median.
• If the dot plot or histogram of a set of data is clearly skewed to the left, then its mean is smaller than its median.

Let’s apply this result to some of our data sets.

**Brian’s Running Data.** In the Chapter 1 homework, we learned about Brian’s study of running. Dot plots of his two data sets are in Figure 1.3 on page 24. Brian’s combat boots data look neither symmetric nor skewed to me. The mean, 333.0, is very similar to the median, 331.5. Brian’s jungle boots data look strongly skewed to the left, but the mean, 319.5, is only somewhat smaller than the median, 321.0.

Brian’s data sets help me to illustrate a common misconception about Result 2.1. For example, in my experience, sometimes a person calculates the mean and median of a data set and finds that they are equal or similar in value. The person then asserts, without drawing a picture—dot plot or histogram or any other picture—of the data, that the distribution of the data is symmetric. This can be wrong as illustrated by both of Brian’s data sets. (We can also find data sets for which the mean is larger [smaller] than the median but the distribution of the data would not be described as skewed to the right [left].)

The message is: Do not confuse an if . . . then result with an if and only if result. In math, all definitions and some results (theorems) are if and only if. Many results, in math or other disciplines, are if . . . then results. For example, at the time of my typing these words, the following is a true statement.

If a person is or has been the President of the United States, then the person is male.

The reverse is not true. I—and millions of other men—have never been the President of the United States.

**Reggie and Darts.** In the Chapter 1 homework, we learned about Reggie’s study of darts. Dot plots of his two data sets are in Figure 1.4 on page 25. Reggie’s data from 10 feet look approximately symmetric to me. The mean, 201.53, is very similar to the median, 200.0. Reggie’s data from 12 feet look strongly skewed to the left. In agreement with Result 2.1, the mean, 188.0, is smaller than the median, 196.0.

I conclude that Result 2.1 is accurate for Reggie’s data.

**Kenny and Fast Cars.** In the Chapter 1 practice problems, we learned about Kenny’s study of car speeds. Dot plots of his two data sets are presented in Figure 1.2 on page 21. Kenny’s 6:00 data are strongly skewed to the right with a large outlier, but the mean, 29.68, is only a bit larger than the median, 29.0. Kenny’s 11:00 data are not approximately symmetric, yet the mean, 34.42, is very similar to the median, 34.0.

I conclude that Result 2.1 is not very accurate for Kenny’s data.
Relative Frequency and Density Histograms. In the above, I stated that for a frequency histogram the height of a rectangle is equal to the frequency of its class interval. As you might guess or already know, a relative frequency histogram differs from a frequency histogram in only one way: The height of any of its rectangles is equal to the relative frequency of the corresponding class interval. For Sara’s data, this involves taking her frequency rectangles and dividing each height by 40, in order to convert to relative frequencies. Even if you have not seen the movie, *Honey, I Shrunk the Kids*, you likely realize that this shrinkage of each rectangle (in going from frequency to relative frequency) has no impact on the shape of the histogram. Thus, in terms of shape, it does not matter which of these two histograms we use.

Similarly, a density histogram differs from the previous two histograms only in terms of the heights of its rectangles: The height of any of its rectangles is equal to the density of the corresponding class interval. For Sara’s data, this involves taking her relative frequency rectangles and dividing each height by the constant width, 25, in order to convert to densities. Again, as in the movie, *Honey, I Shrunk the Kids*, this shrinkage (because $w > 1$) has no impact on the shape of the histogram. Thus, in terms of shape, for histograms with constant-width class intervals, it does not matter which of these three histograms we use. Please note the following items:

1. As we will see in the Practice Problems, if the class intervals do not have constant widths, you should not use the frequency or relative frequency histogram. They will have the same misleading shape. In this situation, the density histogram will have a different shape and should be used.

2. If we have constant-width class intervals and if $w < 1$, then the densities are larger than the relative frequencies, but all three histograms still have the same shape. (See the totally unnecessary sequel *Honey, I Blew Up the Kid.*)

3. In a frequency or relative frequency histogram we look at the height of a rectangle. The height reveals either how many or what proportion of observations are in the class interval, depending on the histogram. For a density histogram, one should look at the area of a rectangle, not its height. In particular, for a density histogram, the area of a rectangle is the relative frequency of the class interval:

$$\text{Area} = \text{Base} \times \text{Height} = w(r_f/w) = r_f.$$

4. In view of the previous item, we see that the total area of a density histogram equals 1.

In view of the above, which of the three histograms is best? Or does it matter?

1. If one chooses to have variable-width class intervals, the density histogram must be used.

2. If for theoretical or other reasons—see later developments in these notes—one wants the total area of the picture to equal one, the density histogram must be used.

3. If neither of the above apply, then all three histograms give the same picture. In this situation, I avoid the extra work involved in constructing the density histogram. So, how do I choose between frequency and relative frequency histograms?
(a) If the number of observations is small, I prefer frequency: with, say, 10 observations I prefer to say, “Four of the observations are . . .” rather than “Forty percent of the observations are . . .”

(b) If the number of observations is large, I prefer relative frequency: with, say, 16,492 observations I prefer to say, “Twenty-five percent of the observations are . . .” rather than “Four thousand one hundred twenty-three of the observations are . . .”

(c) If I am comparing two data sets, as we always do in a comparative study, and the sample sizes \( n_1 \) and \( n_2 \) are substantially different, then I prefer relative frequency histograms over frequency histograms.

### 2.2.1 Kernel Densities

Let us return to my earlier discussion of Sara’s dot plots being bumpy. Look again at her histograms in Figure 2.3 and remember my road analogy on page 30. Here each road has long expanses that are flat, making it much smoother than its corresponding dot plot. Unfortunately, these flat, smooth roads result in the careless travel periodically hitting a wall or going over a cliff! (Well, periodically, provided such encounters prove to be neither incapacitating nor sufficiently discouraging to end one’s journey.) Clearly, these roads require warning signs and elevators (lifts in the U.K.)! A solution to these dangers is provided by what I will simply call kernel density histograms, or kernel densities, for short. Later in these notes you will learn that a better name is kernel density histogram estimates. But right now we don’t know what estimates are and we certainly don’t know what kernels are.

Kernels are fairly easy to explain—though not until we develop several more ideas. They are somewhat difficult to implement and require some careful computations. Software packages exist that will perform the computations for you, but we won’t be covering them in this course. I use the statistical software package Minitab to create all of the kernel densities in this course. For our purposes, a kernel density provides a smoother picture of our data than either a dot plot or a histogram. In my experience, kernel densities frequently appear in online articles and in published reports—books, journals and magazines. As a result, even though I won’t teach you how to construct a kernel estimate, there is value in my introducing you to the idea of one.

Figure 2.4 provides kernel densities for both of Sara’s data sets. I want to make several comments on these pictures.

1. For a given set of data, there is not the kernel density, there are many. Think of kernel densities as being a range of possibilities between two extremes: the dot plot which is very bumpy and a histogram that has one class interval for all data which, of course, will be one rectangle and, hence, smooth. The kernel densities I have plotted are in some sense the best kernel densities. If the idea of best interests you, read the next item; if not, you may ignore the next item.

2. If you want more information on kernel densities, see the Wikipedia entry for kernel density estimation. Following the terminology in Wikipedia, I chose the Normal kernel (also called the Gaussian kernel) with bandwidth \( h = 15 \) for both pictures. This choice of bandwidth is
Figure 2.4: Kernel densities for Sara’s 3-Wood and 3-Iron data.

3. The 3-Wood kernel density is skewed to the left, in agreement with my histogram. The 3-Iron kernel density is approximately symmetric, again in agreement with my histogram.

4. Given the small amount of time we will spend on this topic, I don’t want you to be concerned about too many issues. Essentially, kernel estimates are good because they give us a smooth picture of the data. They can also be used to calculate areas; hence, my inclusion of a vertical scale on these pictures. The area under a kernel density equals one, which explains why they include the word density in their name.

5. Kernel densities are a reasonable descriptive tool provided the response is a measurement. They should not be used if the response is a count. Thus, I would definitely avoid creating a kernel density for Dawn’s study of Bob the cat. Some statisticians relax this directive if the count variable exhibits a large amount of variation. For example, some statisticians might use a kernel density to describe Reggie’s dart scores. Sadly, we can’t spend additional time on this subject.

2.3 Interpreting the Standard Deviation

Please excuse a slight digression. Years ago at a conference on Statistics education, I heard a wonderful talk about students’ understanding of the standard deviation. The speaker had interviewed her students approximately one month after the end of her one semester course on introductory Statistics. She found that very few students—even among the students who earned an A in her course—could adequately explain the meaning of the standard deviation. I really admired the speaker for talking about something that many (most? all?) of us teachers suspect: There is
something about the standard deviation that our students just don’t get. I conclude that teachers—including me—need to improve our explanations of the standard deviation. This section is my attempt to do better than I have in the past.

Note: When we have a comparative study, for example a CRD, we have two standard deviations—one for each set of data—and we distinguish them with subscripts. Similarly, we distinguish our two means by using different letters of the alphabet, $x$ and $y$. When I discuss means and standard deviations in general terms, my data are represented by

$$w_1, w_2, \ldots, w_m, \text{ with mean } \bar{w} \text{ and standard deviation } s.$$ 

Let’s revisit what we know. We have decided to measure spread by looking at how much each observation differs from our arithmetic summary, the mean. The discrepancy between an observation and the mean is called the deviation of the observation. A deviation can be negative, zero or positive. The sign of a deviation tells us whether its observation is smaller than—if negative—or greater than—if positive—the mean. The magnitude (absolute value) of a deviation tells us how far the observation is from the mean, regardless of direction. Our goal is to find a way to summarize all these deviations with one number which will be our measure of the spread in the data.

It is a waste of time to summarize deviations by calculating their mean because for every data set the mean deviation is 0. It seems to make sense to summarize deviations by computing the mean of the magnitudes, but, alas, this summary is of no use in Statistics. Instead, we do something very strange. Something I never saw in all my years of studying math until I took my first Statistics course.

- We square each deviation (equivalently, square each magnitude);
- We compute almost the mean (remember: we divide by degrees of freedom, not sample size) of the squared deviations, calling the resulting number the variance;
- We compensate for having squared the deviations by next taking the square root of the variance and call the result the standard deviation.

Now, consider the name: standard deviation. Why this name? Well, the deviation part makes sense; the summary we obtain is function of the set of deviations. In my experience, it’s the word standard that befuddles people. Why the modifier standard? I actually don’t know. My guess is that we say standard because, as we will see repeatedly in these notes, the standard deviation is essential for the process of standardizing. We will see that standardizing is very useful. But, for me, this is a chicken versus egg situation: my guess is that the idea of standardizing is more basic and, hence, a key number in its process is called the standard deviation. But, I might have this backwards; or the truth might be something else entirely. If the proper person sees these notes, perhaps he/she will tell me the answer and I can improve this presentation!

Let us agree to accept that, perhaps, standard deviation is a strange name for $s$, and let’s proceed to learning what it means. For example, I stated earlier that for Sara’s 3-Iron data,

$$\bar{y} = 98.175 \text{ and } s_2 = 28.33.$$
Figure 2.5: Elastic Man capturing approximately 68% of Sara’s 3-Iron data.

\[ \bar{y} - s_2 = 69.845 \quad \text{and} \quad \bar{y} + s_2 = 126.505 \]

(Recall that the 3-Iron was Sara’s treatment 2; thus, we use \( y \)'s for the data and a subscript of 2 on the \( s \).) How do we interpret the value 28.33 for \( s_2 \)? First, recall that we have an exact interpretation of the value 98.175 for \( \bar{y} \); namely, 98.175 is exactly the center of gravity of the dot plot of the data. Our interpretation of \( s_2 \) is weaker in that it is not exact, it is only an approximation. To make matters worse, sometimes it’s a bad approximation. One positive note: if we have access to a picture of the distribution of the data, then we will know whether the approximation is bad and how it is bad. (See the Practice Problems.)

Our approximation is given in the result below. I recommend that you quickly skim this result and read my motivation which follows it.

**Result 2.2 The Empirical Rule for interpreting the value of the standard deviation.** Suppose that we have a set of data. Denote its mean by \( \bar{w} \) and its standard deviation by \( s \). The three approximations below collectively are referred to as the Empirical Rule.

1. Approximately 68% of the observations lie in the closed interval \([\bar{w} - s, \bar{w} + s]\),

2. Approximately 95% of the observations lie in the closed interval \([\bar{w} - 2s, \bar{w} + 2s]\), and

3. Approximately 99.7% of the observations lie in the closed interval \([\bar{w} - 3s, \bar{w} + 3s]\).

Here is the idea behind the Empirical Rule. The superhero Elastic Man has the ability to stretch his arms as much as he desires. He is standing on the number line at the mean of the data. He poses the following question to himself:

How far do I need to stretch my arms in order to encompass 68% of Sara’s 3-Iron data?

The first approximation in the Empirical Rule answers this question; it tells Elastic Man to stretch enough so that one hand is at \((\bar{y} - s_2)\) and the other hand is at \((\bar{y} + s_2)\). This activity is pictured in Figure 2.5. This picture is a bit busy, so let me spend a few minutes explaining it. Elastic Man (a.k.a. square-headed man) is standing above the mean of the data, at 98.175. His hands extend from

\[ \bar{y} - s_2 = 98.175 - 28.33 = 69.845 \quad \text{to} \quad \bar{y} + s_2 = 126.505. \]
Table 2.4: The performance of the Empirical Rule for Sara’s golf data. The values in this table are the number (percentage) of observations, out of 40, in each interval. Remember that the mean and standard deviation (SD) are \( \bar{x} = 106.175 \) and \( \bar{y} = 98.175 \) and \( s_1 \) and \( s_2 \) for the 3-Wood [3-Iron] data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Mean ± SD</th>
<th>Interval:</th>
<th>Mean ±2 SD</th>
<th>Mean ±3 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-Wood</td>
<td>29 (72.5)</td>
<td>37 (92.5)</td>
<td>40 (100)</td>
<td></td>
</tr>
<tr>
<td>3-Iron</td>
<td>22 (55.0)</td>
<td>39 (97.5)</td>
<td>40 (100)</td>
<td></td>
</tr>
</tbody>
</table>

According to the Empirical Rule, his reach encompasses approximately 68% of the data. Let’s see whether the Empirical Rule is accurate. If you look at Sara’s 3-Iron data in Table 2.2 on page 29, you will see that nine observations are smaller than 69.845 and nine observations are larger than 126.505. Thus, in actuality, \( 40 - (9 + 9) = 22 \) observations lie within the reach of Elastic Man. Sadly, \( 22/40 = 0.55 = 55\% \). The Empirical Rule’s 68% is a poor approximation of 55%. But is it really that bad? Looking at the list of observations again, we see that Elastic Man barely misses three observations at 68 yards and one observation at 127 yards. Add these four observations to the previous total of 22 and we get 26 of 40 observations, which is 65% of the 40 observations and is close to the Empirical Rule’s approximation of 68%. Thus, the approximation fails for these data because Elastic Man needs to stretch a bit farther than \( s_2 \) yards in each direction in order to encompass approximately 68% of the data.

I did additional arithmetic and counting to investigate the Empirical Rule’s performance for both sets of Sara’s data. The results are in Table 2.4. My recommendation: Don’t bother checking these numbers; you will get your chance to create such a table in a Homework problem. The Empirical Rule states that the calculated intervals will encompass approximately 68%, 95% and 99.7% of the data. For five of the intervals the approximations are good and I have already discussed the other interval.

Actually, the Empirical Rule tends to work better for larger amounts of data; 40 observations really aren’t many in this setting. But even with thousands of observations, there are situations in which the Empirical Rule, gives one or more poor approximations. The first interval, mean ± SD, is particularly problematic, just as it was for Sara’s 3-Iron data. This topic is not central to our development in these notes; thus, I will save further examples to the Practice Problems and Homework.

### 2.4 Cathy’s Running Study

I end this chapter with a very small balanced CRD. This will serve as a simple, yet real, example for several ideas presented later in these notes.

Cathy was a very busy student, wife and mother enrolled in my class. One of her favorite escapes was to run one mile. She had two routes that she ran: one through a park and one at her local high school. She decided to use her project assignment to compare her two routes. In
Table 2.5: Cathy’s times, in seconds, to run one mile. HS means she ran at the high school and P means she ran through the park.

<table>
<thead>
<tr>
<th>Trial:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location:</td>
<td>HS</td>
<td>HS</td>
<td>P</td>
<td>P</td>
<td>HS</td>
<td>P</td>
</tr>
<tr>
<td>Time:</td>
<td>530</td>
<td>521</td>
<td>528</td>
<td>520</td>
<td>539</td>
<td>527</td>
</tr>
</tbody>
</table>

particular, Cathy performed a balanced CRD with six trials. A trial consisted of Cathy running one mile and the response was the time, measured to the nearest second, required for Cathy to complete her run. Her treatments were: running at the high school (treatment 1) and running through the park (treatment 2). She assigned trials to treatments by randomization. Her data are presented in Table 2.5. Below are the means, medians and standard deviations for Cathy’s data.

\[
\bar{x} = 530, \bar{x} = 530, s_1 = 9.00, \bar{y} = 525, \bar{y} = 527 \text{ and } s_2 = 4.36.
\]

### 2.5 Computing

Given a set of numerical data, the website:

[http://www.wessa.net/rwasp_density.wasp#output](http://www.wessa.net/rwasp_density.wasp#output)

will create a kernel density. Time limitations prevent me from discussing this site further. A student of mine discovered the following website that will create a histogram of your data:


To be clear, you will **never** be asked to use either of these sites in this course.
2.6 Summary

For many data sets, a dot plot does not provide a satisfactory picture of the distribution of the data. In such cases, a researcher might opt for a histogram.

The quest for a histogram begins with the construction of the frequency table. A frequency table consists of five columns, with headings: class interval, width, frequency, relative frequency and density. Note the following:

1. **Class Interval:** There are many valid choices for the class intervals in a frequency table. The collection of class intervals, however, must satisfy the five rules listed on page 31. In these notes I will always provide you with class intervals that obey these rules.

2. **Width:** The width of a class interval is equal to its upper bound minus its lower bound. It is worth noting whether a frequency table has constant-width or variable-width class intervals.

3. **Frequency:** For each class interval count the number of observations that lie within it, using our endpoint convention: every class interval includes its left endpoint, but not its right, with the exception that the last class interval includes both of its endpoints. The frequencies sum to the number of observations in the data set.

4. **Relative Frequency:** Divide each frequency by the number of observations in the data set; the result is the relative frequency. The relative frequencies for any table sum to one.

5. **Density:** The density of a class interval is equal to its relative frequency divided by its width.

To draw a frequency histogram, follow the two steps on page 32. In particular, the height of a rectangle equals the frequency of its class interval. By contrast, in a relative frequency histogram the height of each rectangle equals the relative frequency of its class interval. For a density histogram, the height of each rectangle equals the density of its class interval which implies that the area of each rectangle equals the relative frequency of its class interval.

For a frequency table with constant-width class intervals, all three histograms have the same shape. For a frequency table with variable-width class intervals, one should use the density histogram; the other two types of histograms are misleading.

We learned in Chapter 1 that the mean of a set of data is exactly equal to the center of gravity of the data set’s dot plot. Thus, for example, if a dot plot is exactly symmetric then the mean (and median) equal the point of symmetry. Result 2.1 on page 33 extends this relationship to dot plots that are not exactly symmetric. This is not a totally satisfactory result because the best I can say is that its conclusions are usually true. Despite this weakness, Result 2.1 is considered to be useful.

As discussed earlier in this chapter, a dot plot can be a very bumpy picture of a distribution of data. A histogram replaces the bumps with a picture that is flat between (up or down) jumps. A kernel density goes one step further: it has neither bumps nor jumps; it is a smooth picture of the distribution of the data. You will never be asked to construct a kernel density.

The Empirical Rule (Result 2.2) provides us with an interpretation of the value of the standard deviation, $s$: Approximately 68% [95%; 99.7%] of the deviations have a magnitude that is less than or equal to $s$ [$2s$; $3s$].

42
Table 2.6: Sorted speeds, in MPH, by time-of-day, of Kenny’s 100 cars.

### Speeds at 6:00 pm

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 28  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 29  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  | 30  |

### Speeds at 11:00 pm

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 33  | 33  | 33  | 33  | 34  | 34  | 34  | 34  | 34  | 34  | 34  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  | 35  |
| 37  | 37  | 37  | 37  | 38  | 38  | 38  | 39  | 39  | 39  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  | 40  |

2.7 Practice Problems

Recall from Chapter 1 that Kenny the policeman conducted a comparative study on the speeds of cars. Kenny’s data are reprinted in Table 2.6. We will use these data in some of the problems below.

1. Using class intervals 26–29, 29–32, 32–35, 35–38, 38–41, and 41–44, construct the frequency tables for both sets of Kenny’s data. Remember to use the endpoint convention; for example, the observation 32 is placed in the interval 32–35.

2. Using your tables from problem 1, draw the frequency histograms for both sets of Kenny’s data.

3. Kernel densities for Kenny’s data are in Figure 2.6 on page 47. Comment on these pictures.

4. The means and standard deviations of Kenny’s data are:

   \[ \bar{x} = 29.68, \quad s_1 = 2.81, \quad \bar{y} = 34.42 \quad \text{and} \quad s_2 = 3.25. \]

   Use these values and Kenny’s actual data in Table 2.6 to check the performance of the Empirical Rule.

5. The purpose of this problem is to give you some practice working with the three types of histograms. The parts of this problem presented below are variations on the theme *my dog ate my homework* in the sense that each part provides only partial information about a histogram.

   (a) I have a single rectangle from a histogram. Its endpoints are 12 and 15 and its height is 0.03. Given that there are \( n = 1000 \) observations in the data set, *how many* observations are in this interval (remembering our endpoint convention) if it is:

   i. A frequency histogram?

   ii. A relative frequency histogram?
iii. A density histogram?

(b) I have a single rectangle from a histogram. Its endpoints are 10.00 and 10.05 and its height is 3. Given that there are \( n = 500 \) observations in the data set, how many observations are in this interval (remembering our endpoint convention) if it is:

i. A frequency histogram?

ii. A relative frequency histogram?

iii. A density histogram?

(c) I have a single rectangle from a histogram. Its endpoints are 20 and 22, but I am not going to tell you its height. Given that there are \( n = 600 \) observations in the data set, and that 10% of these observations are in this interval, how tall is the rectangle if it is:

i. A frequency histogram?

ii. A relative frequency histogram?

iii. A density histogram?

The remaining Problems in this section do not follow my usual Question–Answer format. Rather, they are extended examples, where I illustrate some ideas, but don’t ask you to do any work. These remaining problems are important, so please read them carefully.

6. This is an extreme example of skewed data, but the data are real. I want to use strongly skewed data because, as you will see, having variable-width class intervals is very useful for skewed data. One of my favorite books for skimming through is *The Baseball Encyclopedia*. Before the internet this book was the first place I would look if I had a question about the history of American professional baseball. A huge segment of *The Baseball Encyclopedia* is devoted to the career statistics of every man who has played major league baseball, dating back to the 1880s, as I recall. The men are separated into two sections, players and pitchers. (A few men appeared in both sections, most notably Babe Ruth who had substantial, and glorious, years as both a pitcher and a player.) A few years ago I selected 200 men at random from the tens of thousands in the player section. I suspect that you have a good idea what I mean by at random; I will discuss the concept carefully in Part II of these notes. For my present purposes, your understanding of this issue is not important.

For each man selected, I recorded a simple statistic: the total number of games in which he appeared during his major league career. I won’t list the 200 observations here, but I will give you some highlights:

(a) The shortest career was one game, a value possessed by 11 players in my sample.

(b) The longest career in my sample was 3,308 games.

(c) My three favorite summary statistics are below. Because these data are not from a comparative study, I will denote my data by \( w \)'s, the mean by \( \bar{w} \), the standard deviation by \( s \) and the sample size by \( m \).

\[
\bar{w} = 354.1, \; \tilde{w} = 83.5 \; \text{and} \; s = 560.4.
\]
Note that for this nonnegative response, the ratio of the mean to the median is

\[
\frac{354.1}{83.5} = 4.24!
\]

(d) In a data set for which almost one quarter (actual count, 47) of the observations are fewer than 10 games, 21 players had careers of more than 1,000 games, of which six had careers of more than 2,000 games. In fact, the 21 players with the longest careers played a total of 35,151 games, which is almost one-half of the total of 70,810 games played by all 200 players! Any way you look at this data set, these data are strongly skewed to the right.

Figure 2.7 presents a constant-width frequency histogram of these data. The class intervals are: 0–100, 100–200, 200–300, . . . , 3300–3400 games. I do not like this picture! Here is the main feature that I don’t like about it:

Despite having 34 class intervals, over one-half of the data (103 observations) are placed into one interval. At the other extreme, there are twelve class intervals with no observations, six class intervals with one observation and a total of 24 class intervals with three or fewer observations!

This is a bit like having a road atlas (we used these before google maps and gps) with hundreds of pages devoted to Alaska and one-quarter page to New York City. Well, in my actual road atlas, two pages are devoted to New York City and only one page to Alaska. My road atlas *puts emphasis on* the place with lots of streets and people and *discounts* the state with only a handful of highways. We should do the same in Statistics. We accomplish this goal by using variable-width class intervals. The guiding principle is:

- In regions of the number line in which data are plentiful, we want detail. Thus, we make the class intervals narrow.
- In regions of the number line in which data are scarce, we group data more coarsely. Thus, we make the class intervals wide.

Following this principle, I drew a density scale histogram for these baseball data with the following class intervals:

- Four intervals with width = 25: 0–25, 25–50, 50–75 and 75–100;
- Four intervals with width = 100: 100–200, 200–300, 300–400 and 400–500;
- One interval with width = 500: 500–1000;
- One interval with width = 1000: 1000–2000; and

This new histogram is presented in Figure 2.8. I will make two comments about it.
(a) In the earlier picture, 103 observations were grouped together into the first class interval, 0–100 games. In the new histogram, this interval has been divided into four narrower intervals. With this extra detail, we can see that almost two-thirds (actual count, 67 out of 103) of these careers were shorter than 25 games (remember out endpoint convention: 0–25 does not include 25). In fact, the rectangle above 0–25 has area:

$$25(0.0134) = 0.335.$$

Recall that for a density histogram, area equals relative frequency. Thus, 33.5% of the observations are in the interval 0–25. Finally, 33.5% of 200 equals 67 players, as I mention above parenthetically.

(b) Beginning with the class interval 500–1000 and moving to the right, this new histogram is much smoother than the earlier constant-width frequency histogram. I like this because, as a baseball aficionado, I can think of no reason other than chance variation for the numerous bumps in the earlier histogram. Note that the area of the rectangle above 1000–2000 is:

$$1000(0.000075) = 0.075.$$

Thus, 7.5% of the 200 players—i.e., 15 players—had careers of 1000–2000 games.

Finally, Figure 2.9 is the frequency histogram for the same class intervals as in Figure 2.8. When you compare these two figures you will see why statisticians label the frequency histogram misleading. It is misleading because even if we are told to focus on the height of each rectangle, we see the area.

7. The goal of this problem is to give you additional insight of the Empirical Rule, Result 2.2.

I used Minitab to generate three artificial data sets, each of size 1000. I will not give you a listing of these data sets, nor will I draw their histograms. Instead, note the following:

(a) The data sets all have the same mean, 500, and standard deviation, 100.

(b) The first data set has a symmetric, bell-shaped histogram. The second data set has a symmetric, rectangular-shaped histogram. The third data set is strongly skewed to the right.

Table 2.7 presents a number of summaries of these three data sets. Please examine this table before reading my comments below. Remember: I have not given you enough information to verify the counts in this table; trust me on these please.

(a) The Empirical Rule approximations are nearly exact for the symmetric, bell-shaped histogram. The Empirical Rule does not work well for the other shapes.

(b) For the symmetric, rectangular-shaped histogram the Empirical Rule approximation count for the interval $\bar{w} \pm s$ is much larger than the actual count. For the interval $\bar{w} \pm 2s$ the Empirical Rule approximation count is much smaller than the actual count.
(c) For the strongly skewed histogram, the Empirical Rule approximation count for the interval $\bar{w} \pm s$ is much smaller than the actual count. Because of this discrepancy, statisticians sometimes abuse the language and say that for skewed data, the standard deviation is *too large*. Obviously, the standard deviation is simply an arithmetic computation; it is what it is and is neither too large nor too small. But, in my opinion, the misspeak has some value. First, in the Empirical Rule, instructing Elastic Man to reach $s$ units in both directions in order to encompass 68% of the data is, indeed, telling my favorite superhero to *reach too far*. Second, as we will see often in these notes when looking at real data, even one extreme value in a data set has a large impact on the value of the standard deviation—making it larger, often much larger. Skewed data almost always contain at least one extreme value.

Also, the Empirical Rule approximation count for the interval $\bar{w} \pm 2s$ is quite close to the actual count. Finally, the Empirical Rule approximation count for the interval $\bar{w} \pm 3s$ is substantially larger than the actual count.

(d) Table 2.7 also presents counts for the data set on length of baseball careers that we studied in the previous problem. These baseball data had 200 observations, not 1,000, so I needed to adjust my Empirical Rule approximation counts. The pattern for these baseball data matches the pattern for the artificial strongly skewed data set.
Figure 2.7: Frequency histogram of the number of games played by $n_1 = 200$ major league baseball players.

Figure 2.8: Variable-width density histogram of the number of games played by $n_1 = 200$ major league baseball players.
Figure 2.9: Misleading frequency histogram of the number of games played by $n_1 = 200$ major league baseball players.

Table 2.7: An examination of the performance of the Empirical Rule for the three artificial data sets in Practice Problem 7 and the real baseball career data in Practice Problem 6.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Shape</th>
<th>Min.</th>
<th>Max.</th>
<th>Actual Counts:</th>
<th>$\bar{w} \pm s$</th>
<th>$\bar{w} \pm 2s$</th>
<th>$\bar{w} \pm 3s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Symmetric, bell</td>
<td>177</td>
<td>823</td>
<td>682</td>
<td>954</td>
<td>998</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Symmetric, rectangular</td>
<td>327</td>
<td>673</td>
<td>578</td>
<td>1,000</td>
<td>1,000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Skewed to the right</td>
<td>402</td>
<td>989</td>
<td>868</td>
<td>941</td>
<td>979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical Rule Approximation:</td>
<td></td>
<td></td>
<td>680</td>
<td>950</td>
<td>997</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseball Careers:</td>
<td>0</td>
<td>3,308</td>
<td>175</td>
<td>192</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical Rule Approximation:</td>
<td></td>
<td></td>
<td>136</td>
<td>190</td>
<td>199</td>
<td></td>
</tr>
</tbody>
</table>
2.8 Solutions to Practice Problems

1. The frequency tables are in Table 2.8.

2. The histograms are in Figure 2.10.

3. These pictures are smooth, of which I approve! The 6:00 PM kernel density is skewed to the right with a peak at about 29 MPH, in agreement with our earlier pictures. The 11:00 PM kernel density has a single peak at about 33 MPH. Its two tails have approximately the same length, but the right tail is heavier.

   Here is a feature that I do not like about the 11:00 PM kernel density: To me, one of the most interesting features in the data set is the fact that while five cars were traveling 40 MPH, none was going faster. This feature is obliterated in the kernel density.

4. The first interval for the 6:00 PM data is:

   \[ \bar{x} \pm s_1 = 29.68 \pm 2.81 = [26.87, 32.49]. \]

   From the table, we see that two observations are smaller than 26.87 and seven observation are larger than 32.49. Thus, this interval encompasses \( 50 - (2 + 7) = 41 \) observations, which is 82% of the total of 50 observations. The Empirical Rule approximation of 68% is not good.

   The second interval for the 6:00 PM data is:

   \[ \bar{x} \pm 2s_1 = 29.68 \pm 5.62 = [24.06, 35.30]. \]

   This interval encompasses \( 50 - 1 = 49 \) observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 95% is a bit small.

   Finally, the third interval for the 6:00 PM data is:

   \[ \bar{x} \pm 3s_1 = 29.68 \pm 8.43 = [21.25, 38.11]. \]

   This interval encompasses \( 50 - 1 = 49 \) observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 99.7% is a bit large.

   The first interval for the 11:00 PM data is:

   \[ \bar{y} \pm s_2 = 34.42 \pm 3.25 = [31.17, 37.67]. \]

   From the table, we see that eight observations are smaller than 31.17 and nine observation are larger than 37.67. Thus, this interval encompasses \( 50 - (8 + 9) = 33 \) observations, which is 66% of the total of 50 observations. The Empirical Rule approximation of 68% is quite good.

   The second interval for the 11:00 PM data is:

   \[ \bar{y} \pm 2s_2 = 34.42 \pm 6.50 = [27.92, 40.92]. \]
This interval encompasses $50 - 1 = 49$ observations, which is 98% of the total of 50 observations. The Empirical Rule approximation of 95% is a bit small.

Finally, the third interval for the 11:00 PM data is:

$$
\bar{y} \pm 3s_2 = 34.42 \pm 9.75 = [24.67, 44.17].
$$

This interval encompasses all 50 observations, The Empirical Rule approximation of 99.7% is quite good.

5. (a) Given that the height is 0.03:
   
   i. This is impossible. For a frequency histogram the height of each rectangle must be an integer.
   
   ii. Three percent of the 1000 observations are in this interval:
   
   $$
   0.03(1000) = 30 \text{ observations}.
   $$
   
   iii. The area of the rectangle is $3(0.03) = 0.09$. Thus, 9% of the 1000 observations are in this interval:
   
   $$
   0.09(1000) = 90 \text{ observations}.
   $$

(b) Given that the height is 3:

   i. 3. For a frequency histogram the height of a rectangle tells us how many observations are in its class interval.
   
   ii. This is impossible. For a relative frequency histogram the height of a rectangle cannot exceed one.
   
   iii. The area of the rectangle is $0.05(3) = 0.15$. Thus, 15% of the 500 observations are in this interval:
   
   $$
   0.15(500) = 75 \text{ observations}.
   $$

(c) We are given that the relative frequency of the interval is 0.10, which makes its frequency $0.10(600) = 60$.

   i. 60. For a frequency histogram the height of a rectangle equals the number of observations in the class interval.
   
   ii. 0.10. For a relative frequency histogram the height of a rectangle equals the relative frequency of observations in it.
   
   iii. The width of the class interval is 2. Thus, the density of the interval is $0.10/2 = 0.05$, which is also the height of its rectangle.
Table 2.8: Frequency tables for Kenny’s data.

<table>
<thead>
<tr>
<th>Class Interval Width</th>
<th>6:00 PM</th>
<th>11:00 PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>26–29 3</td>
<td>19</td>
<td>0.38</td>
</tr>
<tr>
<td>29–32 3</td>
<td>23</td>
<td>0.46</td>
</tr>
<tr>
<td>32–35 3</td>
<td>6</td>
<td>0.12</td>
</tr>
<tr>
<td>35–38 3</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>38–41 3</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>41–44 3</td>
<td>1</td>
<td>0.02</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Figure 2.10: Frequency histograms of car speeds, by time.

6:00 PM:

11:00 PM:
2.9 Homework Problems

In the Chapter 1 Homework Problems we learned about Brian’s study of his running times. (See page [24]) Use Brian’s data, reproduced below, to solve problems 1–5. Wearing combat boots, Brian’s sorted times were:

321 323 329 330 331 332 337 337 343 347

Wearing jungle boots, Brian’s sorted times were:

301 315 316 317 321 321 323 327 327 327


2. Create the frequency table and draw the frequency histogram for Brian’s combat boot data, using the class intervals: 321–333, 333–345 and 345–357. Briefly describe the shape of the histogram.

3. Compare your answers to problems 1 and 2.

4. Create the frequency table and draw the frequency histogram for Brian’s jungle boot data, using the class intervals: 300–310, 310–320 and 320–330. Briefly describe the shape of the histogram.

5. Figure 2.11 presents kernel densities for Brian’s two data sets. Briefly describe what these pictures reveal about the data. Compare these kernel densities to the three frequency histograms from problems 1, 2 and 4.
6. Refer to Sara’s golfing data in Table 2.2. For this problem only I am going to combine Sara’s two sets of data to obtain one set of data with 80 observations. The mean of these 80 numbers is 102.52 and the standard deviation is 29.25.

How do the Empirical Rule (Result 2.2) approximations perform for these data?

7. Below is a dot plot of \( m = 19 \) observations.

(a) How would you label the shape of this dot plot? Approximately symmetric? Skewed? Other?

(b) Calculate the mean and median of these data.

(c) Given your answers to (a) and (b), comment on Result 2.1