# Contents

## 1 Inference Based on Randomization

1. The Completely Randomized Design with a Numerical Response
   1.1 Comparative Studies
   1.2 Dawn’s Study: Various Tools
       1.2.1 A Connection Between Pictures and Numbers
   1.3 The Standard Deviation
   1.4 Computing
   1.5 Summary
   1.6 Practice Problems
   1.7 Solutions to Practice Problems
   1.8 Homework Problems

2. The CRD with a Numerical Response: Continued
   2.1 Kvyn the Rower
   2.2 Sara’s Golf Study: Histograms
       2.2.1 Kernel Densities
   2.3 Interpreting the Standard Deviation
   2.4 Cathy’s Running Study
   2.5 Computing
   2.6 Summary
   2.7 Practice Problems
   2.8 Solutions to Practice Problems
   2.9 Homework Problems
Part I

Inference Based on Randomization
Chapter 1

The Completely Randomized Design with a Numerical Response

A Completely Randomized Design (CRD) is a particular type of comparative study. The word design means that the researcher has a very specific protocol to follow in conducting the study. The word randomized refers to the fact that the process of randomization is part of the design. The word completely tells us that complete randomization is required, in contrast to some form of incomplete randomization, such as the randomized pairs design we will study later in these notes. What is a numerical response? See the following section.

1.1 Comparative Studies

So, what is a comparative study? Let’s look at its two words, beginning with the word study. According to dictionary.com (http://dictionary.reference.com) the fifth definition of study is:

Research or a detailed examination and analysis of a subject, phenomenon, etc.

This reasonably well fits what I mean by a study. Next, again according to dictionary.com, the first definition of compare (the root word of comparative) is:

To examine (two or more objects, ideas, people, etc.) in order to note similarities and differences.

Because of time limitations, for the most part in these notes we will restrict attention to exactly two, as opposed to two or more, things being compared.

In the examples of the first two chapters, Dawn wants to compare two flavors of cat treats; Kymn wants to compare two settings on an exercise machine; Sara wants to compare two golf clubs; and Cathy wants to compare two routes for jogging. In the practice and homework problems of these first two chapters you will be introduced to several other comparative studies. Indeed, a large majority of the chapters in this book are devoted to comparative studies. Why? Two reasons:
1. Comparative studies are extremely important in science.

2. The discipline of Statistics includes several good ideas and methods that help scientists perform and analyze comparative studies.

Next, some terminology: the two things being compared are called the **two levels** of the **study factor**. For our examples we have the following study factors and levels.

- **Dawn’s study factor is the flavor of the cat treat**, with levels equal to *chicken-flavored* and *tuna-flavored*.

- **For Kymn’s study**, her exercise apparatus is called an ergometer which requires two choices by its operator. Kymn’s study factor is the machine setting with first level defined as *small gear with the vent closed*; her second level is *large gear with the vent open*.

- **Sara’s study factor is the golf club she used** with levels *3-Wood* and *3-Iron*.

- **Cathy’s study factor is the route for her one mile run** with levels *at her local high school* and *through the park*.

The remaining components of a comparative study are:

- The **units** that provide the researcher with information.

- The **response** which is the particular information from the unit of interest to the researcher.

- One of the following **methods**:
  
  - The researcher identifies each unit with its level of the study factor, or
  
  - The researcher assigns each unit to a level of the study factor.

I choose to introduce you to the units and the response for each of our studies in the various sections below. I do want to say a bit about the **method** in the last bullet of the above list.

Examples of identifying, sometimes called classifying, are: comparing men and women; comparing old people with young people; comparing residents of Wyoming with residents of Wisconsin. Our development of randomization-based inference—beginning with Chapter 3—in Part I of these notes, **will not consider any studies that involve identifying** units with levels.

As the last sentence implies, randomization-based inference is restricted to studies in which the researcher has the **option** of assigning units to levels. In fact, as the name suggests, we attend only to those studies in which the researcher exercised the option of assignment by using a method called **randomization**. You will learn about randomization in Chapter 3.
1.2 Dawn’s Study; Various Tools

Dawn completed my class several years ago. In this section you will be introduced to Dawn’s project.

The choice of a project topic, or, indeed, any research, should begin with a curiosity about how the world operates. Here is Dawn’s motivation for her study, as she wrote in her report.

I decided to study my cat’s preference to two different flavors of the same brand-name cat treats. I was interested in this study because I figured that Bob, my cat, would prefer the tuna-flavored treats over the chicken-flavored because Bob absolutely loves canned tuna with a passion. My interest came when I looked at the ingredients on the two labels. I noticed that the percentage of real chicken in the chicken-flavored treats was larger than the percentage of real tuna in the tuna-flavored treats.

Thus, Dawn had a pretty good idea of what she wanted to study. Her next step was to operationalize the above notions into the standard language of a comparative study. We know her study factor and its levels from the previous section. Now, we need to specify: the definition of the units and the response.

A unit consisted of presenting Bob with a pile of ten treats, all of the same flavor. The flavor of the treats in the pile determined the level of the study factor for that unit: either chicken (level 1) or tuna (level 2). The response is the number of treats that Bob consumed in five minutes.

The technical term unit is not very descriptive. In this course there will be two types of units: trials and subjects. Dawn’s units are trials. Essentially, we have trials if data collection involves doing something repeatedly. In Dawn’s case this something is setting out a pile of ten treats. And then doing it again the next day. By contrast, in many studies the different units are different people. When the units are people, or other distinct objects, we call them subjects. As we will see later in these notes, sometimes the distinction between trials and subjects is blurry; fortunately, this doesn’t matter.

Dawn decided to collect data for 20 days, with one trial per day. Dawn further decided to have 10 days assigned to each level of her study factor. (Sometimes, as here, I will speak of assigning units to levels. Other times I will speak of assigning levels to units. These two different perspectives are equivalent.) Dawn had to decide, of course, which days would be assigned to chicken-flavored and which days would be assigned to tuna-flavored. She made this decision by using a method called randomization. Randomization is very important. We will learn what it is in Chapter 3. In fact, the word randomized in CRD emphasizes that trials are assigned to levels by randomization. Without randomization, we have some other kind of comparative study; not a CRD.

For example, suppose a researcher wants to compare the heights of adult men and women. The study factor would be sex with levels male and female. Note that the researcher most definitely cannot assign subjects (individual adults) to level (female or male) by randomization or any other method! Sex as a study factor is an example of a classification factor, also called observational factor because each unit is classified according to the level it possessed before entry into the study. Thus, for example, Sally is a female before the study begins; she is not assigned by the researcher to be a female.
Whenever units are assigned by randomization to levels, we call the levels treatments. Thus, if you hear a researcher talking about the treatments in a study, you may conclude that randomization was utilized.

When Dawn randomized, she arrived at the following assignment:

Trials 1, 5, 7, 8, 9, 11, 13, 15, 16 and 18 are assigned to chicken-flavored treats and the remaining ten trials are assigned to tuna-flavored treats.

**Very imprecisely, the purpose of a CRD is to learn whether the treatments influence the responses given by the units.** For Dawn, this becomes: Does the flavor of the treat influence the number of treats that Bob consumes?

Here is an aside, especially for readers with a strong background in experimentation: A key word in the above is influence. In many types of studies, hoping to find influence is too optimistic; in these cases we can seek only an association between the two levels being compared and the response. As we will see later, randomization plays a key role here. Roughly speaking, with randomization we might find influence; without randomization, the most we can hope for is association.

There are several other important features of how Dawn performed her study, but I will defer them for a time and introduce the numbers she obtained. Dawn’s data are presented in Table 1.1. Let me make a few brief comments about this display.

I use the terms specific to Dawn’s study as my labels in this table; namely, day, flavor and number of treats consumed as opposed to trial (or unit), treatment and response. A major goal of mine is to develop a unified approach to CRDs and for this goal, general language is preferred. When we are considering a particular study, however, I prefer language that is as descriptive of the study’s components as possible.

Take a few moments to make sure you understand the presentation in Table 1.1. For example, note that on day 6, Bob was presented tuna-flavored treats and he consumed four of them.

My next step in presenting Dawn’s data is to separate, by treatment, the list of 20 numbers into two groups of 10. These appear in Table 1.2. Note that in this table I preserve the order in which the data, within treatment, were collected; e.g., the first time Bob was presented with chicken (day #1), he consumed 4 treats; the second time (day #5) he consumed 5 treats; and so on. I have done
this because sometimes a researcher wants to explore whether there is a *time-trend* in the data. We won’t look for a time-trend in this study, in part, to keep this presentation simple.

Usually it is useful to sort—always from smallest to largest in Statistics—the data within each treatment. I present these new lists in Table 1.3. Of the three tables I have created with Dawn’s data, I find the sorted data easiest to interpret quickly. For example, I can see easily the smallest and largest response values for each treatment and, more importantly, that, as a group, the chicken responses are substantially larger than the tuna responses.

In Statistics we talk a great deal about *within group* and *between group* variation. *Within* the sorted list of chicken data, I see variation. Indeed, I would say that there is a great deal of day-to-day variation in Bob’s interest in chicken-flavored treats. Similarly, I see a great deal of variation within the sorted tuna data. Finally, I see a substantial amount of variation between the flavors: the responses to chicken are clearly larger, as a group, than the responses to tuna. In fact, you can easily verify that overall Bob ate 51 chicken treats compared to only 29 tuna treats.

Statisticians and scientists find it useful to draw a picture of data. We will learn a variety of pictures, starting with the *dot plot* also called the *dot diagram*. The dot plots for Dawn’s data are in Figure 1.1. Some of you are already familiar with dot plots. Others may find them so obvious that no explanation is needed, but I will give one anyways. I will *explain* a dot plot by telling you how to construct one. Look at the dot plot for chicken. First, I draw a segment of the number line that is sufficient to represent all data, using the method described below. The plot contains 10 dots, one for each of the 10 chicken responses. Dots are placed above each response value. When a particular response value occurs more than once, the dots are stacked so that we can see how many are there. For example, there are three dots above the number ‘6’ in the dot plot of the chicken data because on three chicken-days Bob consumed six treats.

Statisticians enjoy looking at a dot plot and labeling its shape. I don’t see a shape in either of these dot plots. Indeed, I would argue that it is extremely unusual to see a shape in a small amount of data. One thing that I do see is that neither of these dot plots is symmetric. Left-to-right symmetry is a big deal in Statistics-pictures (as we will see). With real data perfect left-to-right symmetry is extremely rare and we are usually happy to find approximate symmetry. In fact, I
Figure 1.1: The dot plots for the cat treat study.

Chicken:

![Chicken Dot Plot]

Tuna:

![Tuna Dot Plot]

would say that the tuna dot plot is approximately symmetric. You may reasonably disagree and it turns out not to matter much in this current example.

Do you recall my earlier remarks about within and between variation in these data? I comment now that these features are easier to see with the dot plots than with the listings of sorted data. This is a big reason why I like dot plots: Sometimes they make it easier to discover features of the data.

It is, of course, a challenge to look at 10 (for either treatment) or 20 (for comparing treatments) response values and make sense of them. Thus, statisticians have spent a great deal of time studying various ways to summarize a lot of numbers with a few numbers. This can be a fascinating topic (well, at times, for statisticians, if not others) because of the following issues:

1. Are there any justifications for selecting a particular summary?
2. For a given summary, what are its strengths and desirable properties?
3. What are the weaknesses of a given summary?

Statisticians classify summaries into three broad categories. (There is some overlap between these categories, as you will learn later.)

1. Measures of center. Examples: mean; median; mode.
2. Measures of position. Examples: percentiles, which include quartiles, quintiles and median. Also, percentiles are equivalent to quantiles.
3. Measures of variation (also called measures of spread). Examples: range; interquartile range; variance and standard deviation.

Don’t worry about all of these names; we will learn a little bit about some of them now and will learn more later. I suspect that many of you already know a little or a lot about several of these summaries. (For example, my granddaughter learned about medians at age nine in fourth grade.)
For Dawn’s data we will learn about the three measures of center listed above. To this end, it will be convenient to adopt some mathematical notation and symbols. We denote the data from the first [second] treatment by the letter $x$ [$y$]. Thus, Dawn’s chicken data are denoted by $x$’s and her tuna data are denoted by $y$’s. We distinguish between different observations by using subscripts. Thus, in symbols, Dawn’s chicken data are:

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}.$$  

Obviously, it was very tedious for me to type this list and, thus, in the future I will type simply $x_1, x_2, \ldots x_{10}$. Similarly, Dawn’s tuna data are denoted by $y_1, y_2, \ldots y_{10}$.

Our next bit of generalization is needed because we won’t always have 10 observations per treatment. Let $n_1$ denote the number of observations on treatment 1 and $n_2$ denote the number of observations on treatment 2. For Dawn’s data, of course, $n_1 = n_2 = 10$. Whenever $n_1 = n_2$ we say that the study is balanced.

The subscripts on the $x$’s and $y$’s denote the order in which the data were collected. Thus, for example, $x_1 = 4$ was the response on the first chicken-day; $x_2 = 5$ was the response on the second chicken-day; and so on.

We will need notation for the sorted data too. We try to avoid making notation unnecessarily confusing. Thus, similar things have similar notation. For the sorted data we still use $x$’s and $y$’s as above and we still use subscripts, but we denote sorting by placing the subscripts inside parentheses. Thus, Dawn’s sorted chicken data are:

$$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}, x_{(8)}, x_{(9)}, x_{(10)}.$$

More tediously, for Dawn’s chicken data

$$x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 4, x_{(4)} = 5, x_{(5)} = 5, x_{(6)} = 6, x_{(7)} = 6, x_{(8)} = 6, x_{(9)} = 7, x_{(10)} = 8.$$

Note that the collection of sorted values

$$x_{(1)}, x_{(2)}, \ldots x_{(n_1)}$$

is called the order statistics of the data.

The mean of a set of numbers is its arithmetic average. For example, the mean of 5, 1, 4 and 10 is:

$$(5 + 1 + 4 + 10)/4 = 20/4 = 5.$$  

We don’t really need a mathematical formula for the mean, but I will give you one anyways. Why? Well, later you will need to be comfortable with some formulas of this type, so we might as well introduce an easy one now.

Suppose we have $m$ numbers denoted by

$$w_1, w_2, w_3 \ldots w_m.$$  

The mean of these $m$ numbers is

$$\bar{w} = \frac{w_1 + w_2 + w_3 + \ldots + w_m}{m} = \frac{\sum_{i=1}^{m} w_i}{m} = \frac{\sum w}{m}. \quad (1.1)$$
As you can see, we denote the mean by \( \bar{w} \), read w-bar. For our notation for a CRD, we will have means denoted by \( \bar{x} \) and \( \bar{y} \). Often in these notes I will be informal in my use of summation notation; for example, \( \sum w \) in these notes will be an instruction to sum all the \( w \)'s in the problem at hand.

Note, of course, that for any set of data, summing sorted numbers gives the same total as summing unsorted numbers. You may easily verify that for Dawn’s data:

\[
\bar{x} = \frac{51}{10} = 5.1 \quad \text{and} \quad \bar{y} = \frac{29}{10} = 2.9.
\]

In words, in Dawn’s study, the mean number of chicken treats eaten by Bob is larger than the mean number of tuna treats eaten by Bob. Usually (but not always; exceptions will be clearly noted), we compare two numbers by subtracting. Thus, because \( 5.1 - 2.9 = 2.2 \) we will say that the mean number of chicken treats eaten by Bob is 2.2 larger than the mean number of tuna treats eaten by Bob.

The idea of the **median** of a set of numbers is to find the number in the center position of the sorted list. This requires some care because the method we use depends on whether the sample size is an odd number or an even number. For example, suppose we have five sorted numbers: 1, 3, 6, 8 and 8. There is a unique center position, position 3, and the number in this position, 6, is the median. If, however, the sample size is even, we need to be more careful. For example, consider four sorted numbers: 1, 4, 5 and 10. With four positions total, positions 2 and 3 have equal claim to being a center position, so the median is taken to be the arithmetic average of the numbers in positions 2 and 3; in this case the median is the arithmetic average of 4 and 5, giving us 4.5.

For both sets of Dawn’s data (chicken and tuna) there are 10 numbers; hence, there are two center positions, namely positions 5 and 6. If you look at Table 1.3 again, you will see that the median for Dawn’s chicken data is \( (5+6)/2 = 5.5 \) and the median for her tuna data \( (3+3)/2 = 3 \).

If we have \( m \) numbers denoted by \( w \)'s then the median is denoted by the symbol \( \tilde{w} \), which is read as w-tilde. There are two formulas for calculating the median.

- If the sample size \( m \) is an odd integer, define \( k = (m + 1)/2 \), which will be an integer.

\[
\tilde{w} = w_{(k)} \quad (1.2)
\]

- If the sample size \( m \) is an even integer, define \( k = m/2 \), which will be an integer.

\[
\tilde{w} = \frac{w_{(k)} + w_{(k+1)}}{2} \quad (1.3)
\]

You are not required to use these formulas. Usually, I find it easier to visually locate the center position(s) of a list of sorted data. Later in these notes, however, you will need to learn to use similar formulas for calculating percentiles.

I have one additional comment on Equation 1.2, which applies to many of the equations and formulas in these notes. When you are reading this equation, do not have your inner-voice say, “w-tilde equals w-sub-parentheses-k.” This sounds like gibberish and won’t help you learn the material. Instead, read what the equation signifies. In particular, I recommend reading the equation as “We obtain the median by finding the number in position \( k \) of the sorted list.” Similarly, I read Equation 1.3 as, “We obtain the median by taking the arithmetic average of two numbers. The first
of these numbers is the number in position \( k \) of the sorted list. The second of these numbers is immediately to the right of the first number.”

We won’t use the mode much in these notes, but I will mention it now for completeness. It is easiest to explain and determine the mode by looking at a dot plot of the data. Refer to Figure [1.1]. In the chicken dot plot the tallest stack of dots occurs above the number 6. Thus, 6 is the mode of the chicken data. Similarly, the mode of the tuna data is 3. If two (or more) response values are tied for being most common, they are both (all) called modes. As an extreme case, with a set of \( m \) distinct numbers, every response value is a mode. It seems bizarre to claim that reporting \( m \) values of the mode is a summary of the \( m \) observations!

### 1.2.1 A Connection Between Pictures and Numbers

For any set of numbers, there is a very strong connection between their dot plot and their mean.

Suppose that we have \( m \) numbers denoted by \( w_1, w_2, \ldots, w_m \). As usual, let \( \bar{w} \) denote the mean of these numbers. From each \( w_i \) we can create a new number, called its deviation, which is short for deviation from the mean. We create a deviation by taking the number and subtracting the mean from it. Symbolically, this gives us the following \( m \) deviations:

\[
w_1 - \bar{w}, w_2 - \bar{w}, \ldots, w_m - \bar{w}.
\]

Let’s have a quick example. Suppose that \( m = 3 \) and the three observations are 4, 5 and 9, which gives a mean of \( \bar{w} = 6 \). Thus, the deviations are

\[
4 - 6, 5 - 6 \text{ and } 9 - 6; \text{ or } -2, -1 \text{ and } 3.
\]

Below is a dot plot of our three numbers.

```
4 5 6 7 8 9
• • •
```

Below is the same dot plot with the deviations identified.

```
Deviations: -2 -1 +3
Observations: • • •
```

4 5 6 7 8 9

The following points are obvious, but I want to mention them anyways:

- The observation 4 has deviation \(-2\) because it is two units smaller than the mean.
- The observation 5 has deviation \(-1\) because it is one unit smaller than the mean.
- The observation 9 has deviation \(+3\) because it is three units larger than the mean.
In general, a non-zero deviation has a sign (positive or negative) and a magnitude (its absolute value). Thus, the deviations for the three observations 4, 5, and 9, have signs: negative, negative and positive, and magnitudes: 2, 1, and 3, respectively. The sign of an observation tells us whether it is smaller than (negative) or larger than (positive) the mean. The magnitude tells us how far the observation is from the mean, regardless of direction. Note, of course, that an observation has deviation equal to zero (which has no sign, being neither positive nor negative, and magnitude equal to 0) if, and only if, it is equal to the mean of all the numbers in the data set.

You have probably noticed that the three deviations for my data set sum to zero; this is not an accident. For any set of data:

\[ \sum_{i=1}^{m} (w_i - \bar{w}) = \sum_{i=1}^{m} w_i - m\bar{w} = m\bar{w} - m\bar{w} = 0. \]

In words, for any set of data, the sum of the deviations equals zero. Statisticians (and others) refer to this property by saying that the mean is equal to the center of gravity of the numbers. If this terminology seems a bit mysterious or arbitrary, perhaps the following will help.

Below I have once again drawn the dot plot of my data set of \( m = 3 \) numbers: 4, 5, and 9, with one addition to the picture.

\[ \bullet \bullet \bullet \]

I have placed a fulcrum at the value of the mean, 6. Imagine the number line as the board on a seesaw and imagine that this board has no mass. Next, imagine each dot as having the same mass. Next, view the three dots as three equal weight (equal mass) children sitting on the board. We can see that with this scenario that with the fulcrum placed at the mean, 6, the seesaw will balance.

Look at the dot plots for Dawn’s data in Figure 1.1. First, consider the chicken data. We can see quickly that if we placed a fulcrum at 5, the picture would almost balance; in fact, recall, that \( \bar{x} = 5.1 \). Similarly, looking at the tuna data, we can see quickly that if we placed a fulcrum at 3, the picture would almost balance; in fact, recall, that \( \bar{y} = 2.9 \). Thus, we can look at a dot plot and get a quick and accurate idea of the value of the mean.

### 1.3 The Standard Deviation

I have mentioned, for example, the within-chicken variation in Dawn’s data. We need a number that summarizes this variation. The summary we choose is a function of the deviations defined above. Actually, there are two measures of variation, also called spread, that we will investigate: the variance and the standard deviation. These two measures are very closely related; the standard deviation is the square root of the variance. Or, if you prefer, the variance is the square of the standard deviation. As a rough guide, statisticians prefer to measure spread with the standard deviation while mathematicians prefer to use the variance. As the course unfolds, you can decide which measure you prefer. It’s not really a big deal; as best I can tell, mathematicians prefer the
Table 1.4: The computation of the variances and standard deviations for Dawn’s data on Bob’s consumption of cat treats.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Chicken</th>
<th>Tuna</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>-1.1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-0.1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>-0.1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.9</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-4.1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.9</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1.9</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>-2.1</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0.9</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>2.9</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Variance

- $s_1^2 = \frac{36.9}{9} = 4.100$
- $s_2^2 = \frac{38.9}{9} = 4.322$

Stand. Dev.

- $s_1 = \sqrt{4.1} = 2.025$
- $s_2 = \sqrt{4.322} = 2.079$

variance because lots of theorems are easier to remember when stated in terms of the variance. (For example, under certain conditions, the variance of the sum is the sum of the variances it certainly easier to remember than the same statement in terms of standard deviations. If you doubt what I say, try it!) On the other hand, for the types of work we do in Statistics, the standard deviation makes more sense. Our approach will be to calculate the variance. Once the variance is obtained, it is just one more step—taking a square root—to obtain the standard deviation. I will introduce you to the computational steps in Table 1.4. Let’s begin by looking at the treatment 1 (chicken) data. In the $x$ column I have listed the ten response values. I placed these numbers in the order in which they were obtained, but that is not necessary. If you want to sort them, that is fine. I sum the $x$’s to find their total, 51, and then divide the total by $n_1 = 10$ to obtain their mean, $\bar{x} = 5.1$. Next, I subtract this mean, 5.1, from each observation, giving me the column of deviations, $x - \bar{x}$. As discussed earlier the deviation is positive [negative] if the observation is larger [smaller] than the mean. For the chicken data, five deviations are negative and five are positive. In terms of magnitude, two deviations are very close to 0 (their magnitudes are both 0.1, for observations 2 and 3); the deviation for observation 5 has the distinction of having the largest magnitude, 4.1. The idea is that we want to summarize these ten deviations to obtain an overall measure of spread within the chicken treatment. In my experience many people consider it natural to compute the mean or median of the magnitudes of the deviations. Neither of these operations—calculating the mean or median magnitude—is shown in the table because neither turns out not to be particularly
useful in our subsequent work. What turns out to be very useful, as we shall see throughout these
notes, is to \textit{square each deviation}. The squared deviations appear in the \((x - \bar{x})^2\) column of the
table.

We find the total of the squared deviations, which appears in the table as 36.90 for the chicken
data.

Now, another strange thing happens. (Squaring the deviations was the first strange thing.)
Mathematicians and statisticians disagree on what to do with the total of the squared deviations,
again 36.90 for the chicken data. Mathematicians argue in favor of calculating the mean of the
squared deviations; i.e., to divide 36.90 by \(n_1 = 10\) to obtain 3.690. Statisticians divide by
\((n_1 - 1) = 9\) to obtain \(36.90/9 = 4.100\).

\textbf{In these notes we will follow the lead of statisticians and divide the sum of squared deviations
by the sample size minus one.} The resultant number is called the \textit{variance of the data} and is
denoted by \(s_1^2\) for our \(x\)’s and \(s_2^2\) for our \(y\)’s. Let me summarize the above with the following
formula.

\textbf{Definition 1.1} \textit{Suppose that we have} \(m\) \textit{numbers, denoted by}

\[ w_1, w_2, \ldots, w_m \]

\textit{with mean denoted by} \(\bar{w}\). \textit{The variance of these numbers is denoted by} \(s^2\) \textit{and is computed as
follows:}

\[ s^2 = \frac{\sum_{i=1}^{m}(w_i - \bar{w})^2}{m - 1} = \frac{\sum(w - \bar{w})^2}{m - 1} \]  

\textit{In particular, for the data from treatment 1, the variance is denoted by} \(s_1^2\) \textit{and is computed as
follows:}

\[ s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1} = \frac{\sum(x - \bar{x})^2}{n_1 - 1} \]  

\textit{For the data from treatment 2, the variance is denoted by} \(s_2^2\) \textit{and is computed as follows:}

\[ s_2^2 = \frac{\sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_2 - 1} = \frac{\sum(y - \bar{y})^2}{n_2 - 1} \]

Why do statisticians divide by the sample size minus one? As discussed earlier—when talking
about the center of gravity interpretation of the mean—I noted that for any data set the sum of the
deviations equals 0. This fact is illustrated for both the chicken and tuna data sets in Table 1.4.

Let’s focus on the chicken data. Each deviation gives us information about the spread in the data
set. Thus, initially, one might think that there are 10 items of information in the 10 deviations. But,
in fact, there are only nine items of information because once we know any nine of the deviations
the value of the remaining deviation is determined; it equals whatever is needed to make the 10
deviations sum to 0.

Here is a simpler example: suppose that I have \(m = 3\) numbers. Two of the deviations are: +4 and
\(-7\). Given these two deviations, we \textbf{know} that the third deviation must be +3. A picturesque
way of saying this is that for \( m = 3 \) observations, “Two of the deviations are free to be whatever number they want to be, but the third deviation has no freedom.” In other words, the three deviations have two degrees of freedom. Thus, in our chicken or tuna data, the ten deviations have nine degrees of freedom. In general, for \( m \) observations, the \( m \) deviations have \((m - 1)\) degrees of freedom. Thus, the answer to the question:

Why do statisticians divide by the sample size minus one?

Is because statisticians divide by the degrees of freedom. There are many reasons why statisticians divide by the degrees of freedom and we will learn some of them in these notes. I won’t, however, introduce new concepts in this chapter simply to explain why,

The variance of the chicken data is 4.100. You may follow the presentation in Table [1.4] and find that the variance of the tuna data is 4.322. This measure of spread is nearly identical for the two data sets; in words, the two data sets have almost exactly the same amount of spread; well, at least as measured by the variance.

As stated earlier, statisticians prefer to take the (positive) square root of the variance and call it the standard deviation. There are three main reasons statisticians prefer using the the standard deviation to measure spread rather than the variance.

1. In many of the formulas we will see in these notes, especially those for population-based inference, the standard deviation appears, not the variance.

2. In Chapter 2, I will give you a guide, called the empirical rule, which allows us to interpret the value of the standard deviation. There is no such useful guide for interpreting the variance. I am saving this for Chapter 2 in order to keep the size of the current chapter—your first after all—less daunting.

3. The standard deviation gets the units of the variable correct; the variance does not. I will explain this below for the data from Dawn’s study.

Regarding the third reason above, consider the unit of the variable for the study of Bob the cat. (Yes, this is unfortunate language. The unit of the variable is not the units of the study. This is one reason I prefer to call the units of the study either trials or subjects.) Each observation counts the number of cat treats consumed. For example, on the first chicken day, four cat treats were consumed by Bob. The mean for the chicken data is 5.1 cat treats. Each deviation is measured in cat treats: the chicken day when Bob consumed 7 treats has a deviation of \( 7 - 5.1 = 1.9 \) cat treats. This day gives a squared deviation of \((1.9)^2 = 3.61 \) cat treats squared, whatever they are. Thus, the variance for the chicken data is 4.100 cat treats squared. When we take the square root of 4.100 to obtain the standard of 2.025, we also take the square root of cat treats squared, giving us a standard deviation of 2.025 cat treats.

1.4 Computing

WARNING: In this course I will direct you to several websites for computing. In my experience, some of these websites do not work for all web servers. My recommendation is to use Firefox, Safari or Chrome. If you have difficulties, contact your instructor for this course.
In this chapter we have seen several tools for presenting and summarizing data: dot plots, means, medians, variances and standard deviations. I have presented these tools as if we perform all the necessary operations by hand. Obviously, we need to reduce the tedium involved in using these tools. Before I discuss the specifics of computing for this chapter, I want to give you a brief overview of computing in this course.

For simple computations, I recommend that you use an inexpensive handheld calculator. For example, I use the calculator on my cellphone; it performs the four basic arithmetic operations and takes square roots. Thus, if you tell me that the variance of a set of data equals 73.82, I pull out my cell phone and find that the standard deviation is \( \sqrt{73.82} = 8.592 \). Similarly, if you tell me that \( x = 5, b_0 = 20 \) and \( b_1 = -1 \), I can determine that the value of

\[
y = b_0 + b_1 x \text{ is } y = 20 - 1(5) = 20 - 5 = 15.
\]

There are literally dozens of approaches you could use to perform more involved computations required in this course; five approaches that come to mind are:

1. Using a variety of websites.
2. Using the statistical software package Minitab.
3. Using some other statistical software package. Of special interest is the freeware package R. (Yes, its name is a single letter.)
4. Using a sophisticated hand-held calculator.
5. Using a spreadsheet program, for example Excel.

In these notes I will provide guidance on the first of these approaches. I use Minitab extensively to produce output that is not available from any website. If you are interested in learning about Minitab, let me know. No promises as to what we will do, but I would like to know. Neither the TA nor I will provide any guidance on the other three options above or, indeed, any other option you might know. Thus, please do not ask us to do so.

The websites are great because:

- They are free.
- They have some quirks, but, for the most part, require little or no training before they are used.

The websites, however, have two potential problems.

- I cannot guarantee that they will remain available because I am not the tsar of the internet.
- Whereas I personally have a great deal of faith in the validity of answers provided by Minitab and R, I don’t really know about these sites. I have found a serious error in the Binomial Probabilities website and will warn you about it when the time comes. I have found some other errors that we can work around and will mention them when the time is appropriate. Are there other errors? Who knows? If you find or suspect an error, please let me know.
I have used Minitab in my teaching and research since 1974. Perhaps obviously, I am very satisfied with its performance. Advantages of Minitab include:

- If you do additional course work in Statistics, eventually you will need to learn a statistical software package.

- Knowledge of Minitab might be a useful addition to any application for employment. Might; no guarantee.

- If you enjoy programming, Minitab will give you a good understanding of the steps involved in a statistical analysis.

The two main drawbacks to learning Minitab are:

- It is not free. At the time of my typing this chapter, I do not know the price of Minitab for my course. The number I have heard is $30 which would buy you a six-month rental of Minitab.

- Compared to the websites, Minitab requires more time before you can get started. In my experience, a great feature of Minitab is this amount of time is much smaller than for any other statistical software package.

Now I will discuss each of the tools mentioned above and how I expect you to make use of them. By the way, in these Course Notes I focus on what I expect you to know in order to do the homework and to be successful on the exams. (Recall that by exams I mean our various quizzes and the final.) If you choose to submit an extra credit project—details to be provided—then you might need to do some work by hand.

**Dot plots.** I have been unable to find a website that will take your data as input and create the dot plot as output. As always in these notes, let me know if you find such a site. Meanwhile, I see no benefit to making you draw a dot plot by hand. Thus, if you need a dot plot I will provide it.

**Median.** The difficulty lies in taking the set of data and sorting its values. This is no fun by hand, but is easy with a spreadsheet program. Once you have the sorted data, you may use Equation 1.2 or Equation 1.3 depending on whether your sample size is odd or even, respectively. If you don’t know how to use a spreadsheet program, no worries; on exams I will give you the sorted list of data.

**Mean, variance and standard deviation.** As with the median, you may perform the arithmetic by using a spreadsheet program. In particular, if you look at Table 1.4 again, you can visualize how to create these columns using a spreadsheet. Again as with the median, if you don’t know how to use a spreadsheet program, no worries; on exams, I will give you the value of the mean and the value of either the variance or standard deviation. For homework, the computation of the mean, variance and standard deviation can be achieved by using our first website. Go to the site:

http://vassarstats.net
On the left side of the page is a blue border, with links in white type. About 75% of the way down the list, click on:

**t-Tests & Procedures.**

Click on the third of the four paragraphs that appear, Single Sample t-Test. You will be taken to a page with a heading Procedure followed by a Data Entry box. This website is a bit nasty, meaning that you need to be very careful how you enter the data. I entered the chicken data in Table 1.2 into the box and clicked on the calculate box. The website produced quite a collection of statistics, including the following:

- The sample size, 10; the sum of the observations, 51; the sum of squared (SS) deviations, 36.9; the variance, 4.1; the standard deviation, 2.0248; the mean, 5.1; and the degrees of freedom (df), 9.

If you use this website, **you need to be very careful with data entry.**

1. **If you enter your observations by typing:** After typing each observation, hit the enter key, *except* that you don’t hit the enter key after the last observation. If you hit the enter key after the last observation, the site reads this as another observation that is equal to 0. Thus, check the output to make sure the sample size is correct.

2. **If you enter your observations by ‘cutting and pasting:’** You must cut and paste a column of numbers; one number per row, as described above for typing. If you paste a row that possesses more than one number, it won’t work. As above, be careful not to have a blank line at the end of your data input.

**Exception:** If you use dot plots in your project, you will need to draw them by hand. If you choose to calculate the median, you will need to sort your data.
1.5 Summary

Scientists use comparative studies to investigate certain questions of interest. A comparative study has the following components:

- **Units**: Units are either trials or subjects. The researcher obtains information—the value(s) of one (or more) feature(s)—from each unit in the study.

- **Response**: The feature of primary interest that is obtained from each unit.

- The scientist wants to investigate whether the level of a study factor influences (strong) or is associated with (weak) the values of the responses given by the units. Almost always in these notes, a comparative study will have two levels.

- Of very great importance to the scientist is the method by which units are identified with or assigned to levels of the study factor.

Regarding this last item, if the method is:

- Units are assigned to levels (or vice versa) by the process of randomization (as described later in Chapter 3)

then the comparative study is called a Completely Randomized Design (CRD). For a CRD the levels of the study factor are called the treatments.

In the first few chapters of these notes the response always will be a number; hence, it is called a numerical response. When a CRD is performed or conducted, the result is that the researcher has data. Our goal is to discover how to learn from these data.

The data can be displayed in tables, in three ways of interest to us.

1. The table can present the data exactly as collected. An example of this is Table 1.1.

2. The data can be presented as above, but separated into groups by treatment. An example of this is Table 1.2.

3. The data in the previous table can be sorted, from smallest to largest, within each treatment. An example of this is Table 1.3.

It is instructive to draw pictures of the data, one picture for each treatment. The picture we learned about in this chapter is the dot plot. An example of a dot plot is in Figure 1.1.

Finally, we learned about four numbers that can be computed to summarize the information in a set of data. They include two measures of center: the mean and the median; and they include two mathematically equivalent measures of variation (spread): the variance and the standard deviation.

There is an exact connection between the dot plot and the mean; namely, the mean is the center of gravity of the dot plot.
1.6 Practice Problems

The idea behind this section is to give you an additional example that highlights the many ideas and methods that you need to learn from this chapter.

First, let me describe the data set we will use in this section. On a spring evening, a Milwaukee police officer named Kenny measured the speeds of 100 automobiles. The data were collected on a street in a “warehouse district” with a speed limit of 25 MPH. Fifty cars were measured between roughly 5:45 and 6:15 pm, referred to below as 6:00 pm. The remaining 50 cars were measured between roughly 10:40 and 11:20 pm, referred to below as 11:00 pm.

Each car’s speed was measured to the nearest MPH. The sorted data, by time, are in Table 1.5. The dot plots of the speeds, by time, are given in Figure 1.2. These speed data will be used to answer questions 1–6 below.

1. This is a comparative study, but not a CRD. Identify the following components of this study.

   (a) What are the units? Are the units trials or subjects?

   (b) What is the response?

   (c) What is the study factor? What are its levels?

   (d) Explain why this is not a CRD.

2. Look at the two dot plots in Figure 1.2. Write a few sentences that describe what the pictures reveal. You should discuss each picture separately and you should compare them.

3. Calculate the mean, median and standard deviation of the 6:00 PM data.

4. Calculate the mean, median and standard deviation of the 11:00 PM data.

5. Briefly discuss your answers to questions 2–4.

6. We will see repeatedly in these notes that the presence of even one outlier might have a big impact on our analysis. Let’s explore this topic a bit. Delete the largest observation from the
6:00 data set and recalculate the mean, median and standard deviation of the remaining 49 observations. Discuss your answers.

1.7 Solutions to Practice Problems

1. (a) The units are the cars driving past the police officer. I think of each car driving past as a trial. If you knew that the 100 cars were driven by 100 different people, you could view the units as subjects. (To paraphrase a well-known national association—channeling Harry Potter, that whose name we do not mention—cars don’t speed, drivers speed.) This is an example where either designation—trials or subjects—has merit. It really is not a big deal whether we call call the units trials or subjects.

(b) The response is the speed of the car, measured to the nearest integer miles per hour.
(c) The study factor is the time of day, with levels 6:00 PM and 11:00 PM.

(d) In my experience, many students find this question to be difficult. Some have said, “Yes, it’s a CRD because the cars are driving past at random.” This is an example of a very important issue in this class. Randomization has a very specific technical meaning. We must follow the meaning exactly in order to have randomization. Admittedly, I have not told you what randomization is, so you might think I am being unfair; if this were an exam, I would be unfair, but this is a practice problem. The key point is that in order to have randomization the police officer first had to have control over when the cars drove past. He had to have a list of the 100 cars (drivers) and say, “You 50, drive past me at 6:00; the remaining 50, you drive past me at 11:00.” Clearly, he did not have this control; he observed when the cars drove past. As is rather obvious from the dot plots, cars at the later hour are driven at substantially higher speeds than cars driven at the earlier hour. But—as we will see later and you can perhaps see now—does this mean that a given person tends to drive faster at the later time or does this mean that fast drivers come out late at night? You might have a strong feeling as to which of these explanations is better (or you might have some other favorite), but here is my point: The data we have will not answer the question of why. In my earlier language, we cannot say that time-of-day influences speed; we only can say that time-of-day is associated with speed.

2. Obviously, there are many possible good answers to this question. My answer follows. Don’t view this as the ideal, but rather try to understand why my comments make sense and think about ways to improve my answer.

**6:00 PM data:** Everybody is driving faster than the speed limit, 25. A substantial majority (32 of 50, if one counts) of the cars are traveling at 28, 29 or 30 MPH. There is not a very much spread in these 50 observations, except for the isolated large value of 43. (A large [small] isolated value is called a large [small] outlier.)

**11:00 PM data:** Everybody is driving faster than the speed limit; in fact, all but two drivers exceed the limit by at least 5 MPH. There is a lot of spread in these response values. There are three clear peaks: from 32–34; at 37; and at 40. The peak at 40 is curious; lots of people (well, five) drive 40, but nobody drives faster. Previous students of mine (I do like this example!) have opined that the drivers are trying to avoid a big increase in penalties for being caught driving more than 15 MPH over the speed limit.

**Comparing dot plots:** The most striking feature is that the speeds are substantially larger at 11:00. Also, there is more spread at 11:00 than at 6:00.

3. I used the website [http://vassarstats.net](http://vassarstats.net) following the method described in Section 1.4. I entered the 6:00 PM data and obtained:

\[
\bar{x} = 29.68 \text{ and } s_1 = 2.810.
\]

To obtain the median we note that \( n_1 = 50 \) is an even number. Following Equation 1.3, we compute \( k = 50/2 = 25 \) and \( k + 1 = 26 \). From Table 1.5 the response 29 is in both positions 25 and 26. Thus, the median \( \tilde{x} \) equals 29.
4. I used the website [http://vassarstats.net](http://vassarstats.net) following the method described in Section 1.4. I entered the 11:00 PM data and obtained:

\[ \bar{y} = 34.42 \text{ and } s_2 = 3.252. \]

From Table 1.5, the response 34 is in both positions 25 and 26. Thus, the median \( \bar{x} \) equals 34.

5. The mean [median] speed at 11:00 is 4.74 [5.00] MPH larger than the mean [median] speed at 6:00. The differences in these measures of center agree with what we see in the dot plots. The ratio of the standard deviations is \( 3.252/2.810 = 1.157 \). Thus, as measured by the standard deviation, there is almost 16% more spread in the later data.

6. With the help of the website, \( \bar{x} = 29.408 \text{ and } s_1 = 2.071 \). For the median, the sample size is now 49, an odd number. From Equation 1.2, we find that \( k = (49 + 1)/2 = 25 \). The observation in position 25 is 29 and it is the median.

The deletion of the outlier has left the median unchanged. The mean decreased by \( 29.68 - 29.41 = 0.27 \) MPH; or, if you prefer, the mean decreased by 0.9%. The standard deviation decreased by 26.3%! As we will see repeatedly in these notes, even one outlier can have a huge impact on the standard deviation.

I do not advocate casually discarding data. If you decide to discard data, you should always report this fact along with your reason for doing so.

Beginning with Chapter 3 we will devote a great deal of effort into learning how to quantify uncertainty, using the language, techniques and results of probability theory. It is important to learn how to quantify uncertainty, but it is equally important to realize that there are many situations in which we cannot, in any reasonable way, quantify uncertainty. Often we just need to accept that our answers are uncertain. In the current case, there is uncertainty about who drives down a street on any given night. I don’t know why driver ‘43’ decided to drive down the street under study, but it’s certainly possible that he/she could have chosen a different route. Thus, I think it is interesting to see what would happen to our analysis if one of the subjects/trials had not been included in the study.
1.8 Homework Problems

Brian performed a balanced Completely Randomized Design with 20 trials. His response is the time, measured to the nearest second, he needed to run one mile. Wearing combat boots, his 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

Figure 1.3 presents the two dot plots for Brian’s study. Use Brian’s data to solve problems 1–4.

1. Calculate the mean, median and standard deviation for the combat boots data.

2. Calculate the mean, median and standard deviation for the jungle boots data.

3. Recalculate the mean, median and standard deviation for the jungle boots data after you delete the small outlier (leaving a set of nine observations).

4. Write a few sentences to explain what you have learned from your answers to problems 1–3 as well as an examination of these dot plots.

Note: There is no unique correct answer to this problem. I don’t put questions like this on my exams—grading such questions is very subjective and I try to avoid such grading issues. Also, I don’t want you to feel you are at a disadvantage compared to the other students in this class who, surprisingly, are all majoring in military footwear. Seriously, I try to avoid grading you based on your scientific knowledge in any particular field that I choose to present. Answering this question is, however, good practice for your project. In a project, you choose the topic; if you choose a topic for which you have no knowledge, no interest and no aptitude, then your grade will suffer!
Reggie performed a balanced CRD of 30 trials. Each trial consisted of a game of darts, where a game is defined as throwing 12 darts. Treatment 1 [2] was throwing darts from a distance of 10 [12] feet. Reggie’s response is the total of the points obtained on his 12 throws and is called his score, with larger numbers better. Below are Reggie’s sorted scores from 10 feet:

181 184 189 197 198 198 200 200 205 205 206 210 215 215 220

Below are Reggie’s sorted scores from 12 feet:

163 164 168 174 175 186 191 196 196 197 200 200 201 203 206

Reggie’s two dot plots are presented in Figure 1.4. Use Reggie’s data to answer problems 5–7.

5. Calculate the mean, median and standard deviation for the scores from 10 feet.

6. Calculate the mean, median and standard deviation for the scores from 12 feet.

7. Write a few sentences to explain what you have learned from your answers to problems 5 and 6 as well as an examination of Reggie’s dot plots.

Keep in mind my comments in problem 4 above.
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 [2] is small gear, vent closed [large gear, 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.

Small Gear, Vent Open:

Large Gear, Vent Closed:

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 plot, 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 I note that means, medians and standard deviations of these data are:

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

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>3-Wood</th>
<th>22</th>
<th>32</th>
<th>38</th>
<th>56</th>
<th>58</th>
<th>77</th>
<th>81</th>
<th>93</th>
<th>99</th>
<th>101</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>101</td>
<td>101</td>
<td>104</td>
<td>107</td>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td></td>
<td>113</td>
<td>114</td>
<td>115</td>
<td>116</td>
<td>118</td>
<td>122</td>
<td>122</td>
<td>127</td>
<td>127</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>128</td>
<td>129</td>
<td>131</td>
<td>131</td>
<td>137</td>
<td>139</td>
<td>139</td>
<td>140</td>
<td>147</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3-Iron</th>
<th>27</th>
<th>52</th>
<th>53</th>
<th>57</th>
<th>58</th>
<th>59</th>
<th>68</th>
<th>68</th>
<th>68</th>
<th>82</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>84</td>
<td>88</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>97</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>101</td>
<td>105</td>
<td>107</td>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>127</td>
<td>132</td>
<td>132</td>
<td>136</td>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>139</td>
</tr>
</tbody>
</table>

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

3-Wood:

3-Iron:

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>3-Wood</th>
<th>3-Iron</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq. (f)</td>
<td>Rel. Freq. (r_f = f/n_1)</td>
<td>Density (d = r_f/w)</td>
</tr>
<tr>
<td>0–25</td>
<td>25</td>
<td>1</td>
<td>0.025</td>
</tr>
<tr>
<td>25–50</td>
<td>25</td>
<td>2</td>
<td>0.050</td>
</tr>
<tr>
<td>50–75</td>
<td>25</td>
<td>2</td>
<td>0.050</td>
</tr>
<tr>
<td>75–100</td>
<td>25</td>
<td>4</td>
<td>0.100</td>
</tr>
<tr>
<td>100–125</td>
<td>25</td>
<td>18</td>
<td>0.450</td>
</tr>
<tr>
<td>125–150</td>
<td>25</td>
<td>13</td>
<td>0.325</td>
</tr>
<tr>
<td>Total</td>
<td>—</td>
<td>40</td>
<td>1.000</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 Kynn’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 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-
frequency 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 falls in the interval 100–125, not in 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. Begin by drawing a portion of the number line and locating on it the various class interval boundaries.

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. 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. 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. 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.

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 with bandwidth $h = 15$ for both pictures. This choice of bandwidth is close to the values (which
Figure 2.4: Kernel densities for Sara’s 3-Wood and 3-Iron data.

are similar, but slightly different for the two data sets) given under the heading *Practical estimation of the bandwidth*.

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

\[ w_1, w_2, \ldots, w_m, \] with mean \( \bar{w} \) and standard deviation \( s \).

I hope that this won’t be confusing.

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.

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

(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

\[
\bar{y} - s = 98.175 - 28.33 = 69.845 \quad \text{to} \quad \bar{y} + s = 126.505.
\]
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 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 homework problem 6. 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.

### 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 $\bar{x} = 106.175$ $[\bar{y} = 98.175]$ and $s_1$ $[s_2]$ for the 3-Wood [3-Iron] data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Mean ± SD</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>
</tr>
<tr>
<td>3-Iron</td>
<td>22 (55.0)</td>
<td>39 (97.5)</td>
<td>40 (100)</td>
</tr>
</tbody>
</table>

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>

In 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 of Cathy’s data.

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

### 2.5 Computing

There are two new tools in this chapter: histograms and kernel estimates. I could find no websites that generate either of these.
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; hence, any one of them is sufficient. For a frequency table with variable-width class intervals, one should use the density histogram; the other two 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 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 \( s \ [2s; 3s] \).
Table 2.6: Sorted speeds, in MPH, by time-of-day, of Kenny’s 100 cars.

<table>
<thead>
<tr>
<th>Speeds at 6:00 pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 29 29 29 29 29 29 29 29 29 30 30 30 30 30 30 30 30</td>
</tr>
<tr>
<td>30 30 31 31 31 31 32 33 33 33 33 34 34 35 43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speeds at 11:00 pm</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 28 30 30 30 31 31 32 32 32 32 32 33 33 33 33 33 33</td>
</tr>
<tr>
<td>33 33 33 34 34 34 34 35 35 35 35 36 36 36 37</td>
</tr>
<tr>
<td>37 37 37 37 38 38 39 39 40 40 40 40 40</td>
</tr>
</tbody>
</table>

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. Comment on these pictures.

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

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

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

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.

5. 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 \textit{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 \textbf{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 \( m \).

\[
\bar{w} = 354.1, \quad \tilde{w} = 83.5 \quad \text{and} \quad s = 560.4.
\]

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 \ref{fig:histogram5} presents a constant-width frequency histogram of these data. The class intervals are: 0–100, 100–200, 200–300, \ldots, 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 \textit{puts emphasis on} the place with lots of streets and people and \textit{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
• One interval with width = 1500: 2000–3500.

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.

6. 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 2.6: Kernel estimates for Kenny’s 6:00 PM and 11:00 PM data.

Figure 2.7: 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 6 and the real baseball career data in Practice Problem 5.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Shape</th>
<th>Min.</th>
<th>Max.</th>
<th>( \bar{x} \pm s )</th>
<th>( \bar{x} \pm 2s )</th>
<th>( \bar{x} \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>
</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>
</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>
</tr>
</tbody>
</table>

Empirical Rule Approximation: 680 950 997

Baseball Careers: 0 3,308
Empirical Rule Approximation: 136 190 199
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.
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 *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 observations 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 observations 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:

\[ \bar{y} \pm 2s_2 = 34.42 \pm 6.50 = [27.92, 40.92]. \]
Table 2.8: Frequency tables for Kenny’s data.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Width</th>
<th>6:00 PM Rel. Density</th>
<th>6:00 PM Freq.</th>
<th>11:00 PM Rel. Density</th>
<th>11:00 PM Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>26–29</td>
<td>3</td>
<td>0.38 0.127</td>
<td>19</td>
<td>0.04 0.013</td>
<td>2</td>
</tr>
<tr>
<td>29–32</td>
<td>3</td>
<td>0.46 0.153</td>
<td>23</td>
<td>0.12 0.040</td>
<td>6</td>
</tr>
<tr>
<td>32–35</td>
<td>3</td>
<td>0.12 0.040</td>
<td>6</td>
<td>0.40 0.133</td>
<td>20</td>
</tr>
<tr>
<td>35–38</td>
<td>3</td>
<td>0.02 0.007</td>
<td>1</td>
<td>0.26 0.087</td>
<td>13</td>
</tr>
<tr>
<td>38–41</td>
<td>3</td>
<td>0.00 0.000</td>
<td>0</td>
<td>0.18 0.060</td>
<td>9</td>
</tr>
<tr>
<td>41–44</td>
<td>3</td>
<td>0.02 0.007</td>
<td>1</td>
<td>0.00 0.000</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>1.00 1.00</td>
<td>50</td>
<td>1.00 1.00</td>
<td>50</td>
</tr>
</tbody>
</table>

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

6:00 PM:

11:00 PM:

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.
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 present 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.

8. The previous problem provided some insight into why Result 2.1 is not a theorem. For a dot plot—an accurate representation of the data—that is clearly skewed to the right, the median is larger than the mean.

This and the following problems investigate a similar issue for histograms. I have seen books that state, “If the histogram is symmetric, then the mean equals the median.” As we will see below, this statement is false. It must be false because histograms throw away detail by putting observations into class intervals. Arithmetic pays attention to details!

I have an artificial data set consisting of \( m = 19 \) values. Below is a partial frequency table for these data.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10–20</td>
<td>5</td>
</tr>
<tr>
<td>20–30</td>
<td>9</td>
</tr>
<tr>
<td>30–40</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
</tr>
</tbody>
</table>

The frequency histogram for these data is symmetric with the point of symmetry at 25. As you will demonstrate below, both the mean and median can be very different from 25.

For the questions below, I will define the following four data sets. To save space I use the following notation. When I type 20-9, for example, I mean that the response value 20 occurs nine times in the data set.

- **Data set A**: 10-5; 20-9; and 30-5. (Thus, this data set consists of:
  
  10, 10, 10, 10, 10, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 30, 30, 30, 30, 30.)
• **Data set B:** 19-5; 20-5; 29-4; and 40-5.
• **Data set C:** 10-5; 20-4; 29-5; and 30-5.
• **Data set D:** Any data set with the following features:
  – Any five values that are at least 10 and smaller than 20;
  – Five values equal to 20;
  – Any four values that are at least 20 and smaller than 30; and
  – Any five values that are at least 30 and at most 40.

It is easy to verify that Data sets A–D all agree with the partial frequency table given earlier in this question. You should verify this fact, but please do **not** include your verification in your submitted homework.

(a) I assert that the smallest possible value of the mean is 20. You don’t need to prove this fact, but you must verify that the mean of data set A equals 20.

(b) I assert that the smallest possible value of the median is 20. You don’t need to prove this fact, but you must verify that the median of data sets A, B and D all equal 20.

9. Refer to the frequency table and data sets (A–D) given in the previous problem. Finding the largest possible values of the mean and median is a bit tricky because of our endpoint convention. For this problem assume that the 19 observations must be integers. This implies that: 19 is the largest possible observation in the class interval 10–20; and 29 is the largest possible observation in the class interval 20–30. Remember that the last class interval includes both of its endpoints. Thus, 40 is the largest possible observation in the class interval 30–40.

(a) I assert that the largest possible value of the median is 29. You don’t need to prove this fact, but you must verify that the median of data set C equals 29.

(b) For data set C, calculate the value of the mean.

(c) Recall that for data set D the median equals 20. Find the largest possible value of the mean for data set D.

(d) In view of your answers to (a)–(c), comment on the statement, “For a symmetric histogram, the mean equals the median.”