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.
Table 1.1: Dawn’s data on Bob’s consumption of cat treats. ‘C’ [‘T’] is for chicken [tuna] flavored.

<table>
<thead>
<tr>
<th>Day</th>
<th>Flavor</th>
<th>Number Consumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>T</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>T</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>T</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>T</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>T</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>T</td>
<td>2</td>
</tr>
</tbody>
</table>

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
Table 1.2: Dawn’s data on Bob’s consumption of cat treats, by treatment.

<table>
<thead>
<tr>
<th>Observation Number:</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>Chicken:</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Tuna:</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1.3: Dawn’s data on Bob’s consumption of cat treats, sorted within each treatment.

<table>
<thead>
<tr>
<th>Position:</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>Chicken:</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Tuna:</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

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 consumption of 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
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:

$$\bar{x} = 51/10 = 5.1 \text{ and } \bar{y} = 29/10 = 2.9.$$ 

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 on page 7 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.
  
  $$\tilde{w} = (w_{(k)} + w_{(k+1)})/2 \quad (1.3)$$

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.

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 on page 8. 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

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

Below is a dot plot of our three numbers.

![Dot plot of three numbers](image)

Below is the same dot plot with the deviations identified.

<table>
<thead>
<tr>
<th>Deviations:</th>
<th>Observations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>( \bullet ) ( \bullet ) ( \bullet )</td>
</tr>
<tr>
<td>-1</td>
<td>( \bullet ) ( \bullet ) ( \bullet )</td>
</tr>
<tr>
<td>+3</td>
<td>( \bullet ) ( \bullet ) ( \bullet )</td>
</tr>
</tbody>
</table>

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, both respectively. The sign of a deviation tells us whether its observation is smaller than (negative) or larger than (positive) the mean. The magnitude of a deviation tells us how far its 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.

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 on page 8. 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 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, a deviation is positive [negative] if its 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

<table>
<thead>
<tr>
<th>Observation</th>
<th>Chicken $x$</th>
<th>$x - \bar{x}$</th>
<th>$(x - \bar{x})^2$</th>
<th>Tuna $y$</th>
<th>$y - \bar{y}$</th>
<th>$(y - \bar{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>-1.1</td>
<td>1.21</td>
<td>3</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-0.1</td>
<td>0.01</td>
<td>5</td>
<td>2.1</td>
<td>4.41</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>-0.1</td>
<td>0.01</td>
<td>0</td>
<td>-2.9</td>
<td>8.41</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.9</td>
<td>0.81</td>
<td>4</td>
<td>1.1</td>
<td>1.21</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-4.1</td>
<td>16.81</td>
<td>7</td>
<td>4.1</td>
<td>16.81</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.9</td>
<td>0.81</td>
<td>3</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1.9</td>
<td>3.61</td>
<td>1</td>
<td>-1.9</td>
<td>3.61</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>-2.1</td>
<td>4.41</td>
<td>3</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0.9</td>
<td>0.81</td>
<td>1</td>
<td>-1.9</td>
<td>3.61</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>2.9</td>
<td>8.41</td>
<td>2</td>
<td>-0.9</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>51</strong></td>
<td><strong>0.0</strong></td>
<td><strong>36.90</strong></td>
<td><strong>29</strong></td>
<td><strong>0.0</strong></td>
<td><strong>38.90</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>$\bar{x} = 51/10 = 5.1$</th>
<th>$\bar{y} = 29/10 = 2.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>$s_1^2 = 36.9/9 = 4.100$</td>
<td>$s_2^2 = 38.9/9 = 4.322$</td>
</tr>
<tr>
<td>Stand. Dev.</td>
<td>$s_1 = \sqrt{4.1} = 2.025$</td>
<td>$s_2 = \sqrt{4.322} = 2.079$</td>
</tr>
</tbody>
</table>
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 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 \(n = 10\) to obtain 3.690. Statisticians divide by \((n - 1) = 9\) to obtain 36.90/9 = 4.100.

**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 **variance of the data** and is denoted by \(s^2\) for our \(x\)'s and \(s^2\) for our \(y\)'s. Let me summarize the above with the following formula.

**Definition 1.1** Suppose that we have \(m\) numbers, denoted by 
\[
w_1, w_2, \ldots, w_m
\]
with mean denoted by \(\bar{w}\). The variance of these numbers is denoted by \(s^2\) and is computed as follows:
\[
s^2 = \frac{\sum_{i=1}^{m}(w_i - \bar{w})^2}{m - 1} = \frac{\sum (w - \bar{w})^2}{m - 1} \tag{1.4}
\]

In particular, for the data from treatment 1, the variance is denoted by \(s^2\) and is computed as follows:
\[
s^2_1 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 1} = \frac{\sum(x - \bar{x})^2}{n_1 - 1}. \tag{1.5}
\]

For the data from treatment 2, the variance is denoted by \(s^2\) and is computed as follows:
\[
s^2_2 = \frac{\sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_2 - 1} = \frac{\sum(y - \bar{y})^2}{n_2 - 1}. \tag{1.6}
\]

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 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.

Let’s return to the question:

Why do statisticians divide by the sample size minus one?

The answer 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 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 cellphone 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 $y = b_0 + b_1 x$ 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 open-source software 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. Earlier, I found a serious error in the Binomial Probabilities website, but later it had been fixed. 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. (By exams I mean the midterm(s) and final exams.)

**Dot plots.** There is a website that will draw a dotplot of a set of data. You won’t need to use it in this course, but I include it for completeness. In general, if I want you to see a dotplot, I will provide it for you. The website is:


If you are interested in this website, I suggest you try it with the chicken data from Dawn’s study.

**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 I.2 or Equation I.3 depending on whether your sample size is odd or even, respectively. (Both equations are on page 10.) If you don’t know how to use a spreadsheet program, no worries; on exams, except possibly for very small data sets, 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 again at Table 1.4 on page 13, 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; i.e., you may not enter more than one observation per line.

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 includes more than one observation, it won’t work.
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 on page 6.

2. The data can be presented as above, but separated into groups by treatment. An example of this is Table 1.2 on page 7.

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 on page 7.

   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 on page 8.

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.
Table 1.5: Sorted speeds, in MPH, by time, of 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 29 30 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 31 32 32 32 32 32 32 32 33 33 33 33 33</td>
</tr>
<tr>
<td>33 33 33 33 34 34 34 34 35 35 35 35 36 36 36 37 37 37</td>
</tr>
<tr>
<td>37 37 37 37 38 38 39 39 40 40 40 40 40</td>
</tr>
</tbody>
</table>

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
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 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 on page 10, 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 \( \tilde{x} \) 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, \( \bar{x} = 29.408 \) and \( s_1 = 2.071 \). For the median, the sample size is now 49, an odd number. From Equation 1.2 on page 10 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 the driver responsible for the large outlier at 43 MPH decided to drive down the street being studied when Kenny was collecting data. But it’s certainly possible that he/she could have chosen a different route or a different time. 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 projects. 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.